교육학박사학위논문

한국어 말하기 평가의 채점 변인 연구

-채점 과정 분석을 중심으로-

2019년 8월

서울대학교 대학원

국어교육과 한국어교육전공

이 성 준

한국어 말하기 평가의 채점 변인 연구

- 채점 과정 분석을 중심으로 -

지도교수 민 병 곤

이 논문을 교육학 박사 학위논문으로 제출함

2019 년 4 월

서울대학교 대학원

국어교육과 한국어교육전공

이 성 준

이성준의 박사 학위논문을 인준함

2019 년 6 월

위 원 장 김 호 정 (인)

부 위 원 장 소 영 순 (인)

위 원 지 현 숙 (인)

위 원 김 평 원 (인)

위 원 민 병 곤 (인)

국문 초록

본 연구의 목적은 채점자가 수험자의 말하기 평가 과제 응답에 대하여 점수를 부여하는 채점의 과정을 실증적으로 살펴보고, 이를 바탕으로 채점 과정 기반의 채점자 교육 방안을 제시하는 것이다.

말하기 평가에서 채점 과정에 개입할 수 있는 다양한 요인들은 채점의 진행에 영향을 미치며, 이는 채점자 사이의 상이한 평가 결과가 나타나게 하는 원인이 된다. 기존의 채점 과정에 관한 연구에서 주목하였던 변인인 채점자의 배경적 특징이 평가 결과에 영향을 주었을 것이라는 추정은 채점의 과정, 즉 채점자가 어떤 상황에서 어떤 채점 방식을 적용하였으며, 이를 바탕으로 도출한 점수를 어떻게 결정한 것인지에 관한 설명이 가능할 때 해석의 타당성을 확보할 수 있다. 이와 관련하여 본 연구에서는 말하기 평가의 채점에 관여하는 직접적인 변인으로서 채점 과정을 실증적으로 탐구하기 위한 통합방법연구를 설계하고, 채점자 영향과 채점 과정에 대한 분석을 통해 말하기 평가의 채점 과정을 규명하고자 하였다.

먼저 말하기 평가의 채점 과정을 설명할 수 있는 이론적인 모형을 구축하고자 하였다. 이는 채점 과정에 관여하는 여러 가지 변인들이 무엇이며, 어떤 경로를 통해 영향을 나타내는지에 관한 가설적인 모형으로서 채점 과정 분석을 위한 틀을 마련하고자 한 것이다. 이와 관련하여 본고에서는 채점 과정에 관한 언어 평가학과 인지 심리학의에서 제시한 추론과 추단의 관점을 바탕으로 채점자의 인지적인 접근을 설명하고자 하였다. 나아가 채점자의 외적·내적인 영향 요인으로 인하여 채점 과정의 변화가 나타날 수 있다는 점을 논의하고, 이를 종합하여 청취한 정보에 대한 지각과 판단, 점수 결정으로 이루어지는 말하기 평가 채점 과정 모형과 평가 맥락에 따른 채점 과정의 유형을 제시하였다.

다음으로 말하기 평가 채점 과정의 구체적인 특징과 양상을 알아보기 위하여 중·고급 학습자 대상의 컴퓨터 기반 한국어 말하기 평가인 ‘한국어 말하기 능력 시험’을 구안하였다. 평가의 문항은 ‘국제통용한국어교육과정’과 한국어 교재에 대한 검토를 바탕으로 구성하였다. 그리고 수험자 응답에 대한 채점을 위하여 선행 연구를 바탕으로 4개 평가 준거에 대한 혼합형 채점 척도를 구성

하였다. 시험에는 총 18명의 수험자가 응시하였으며, 이 중에서 채점 연습 대상을 제외하고 12명의 응답을 본 채점 자료로 사용하였다. 연구에 참여한 채점자는 한국어교육 경력 5년 이상의 교사 13인이었으며, 이들로부터 평가 결과인 점수의 기록과 채점 과정 보고를 녹음한 자료를 수집하였다.

다음으로 평가 결과에 나타난 채점자 영향을 파악하기 위하여 통계 분석을 실시하였다. 먼저 고전검사이론에 따른 분석 결과, 평가 결과에서 채점자들 간에 높은 내적 일관성 신뢰도가 나타났으나, 상대적으로 낮은 상관을 나타내는 채점자가 있음을 확인하였다. 채점자 배경 변인에 관한 분산 분석에서는 채점자의 교육 경력, 평가 경험, 평가 관련 교육 경험, 구사 가능한 외국어 수에 대하여 집단 간에 유의한 차이가 나타나지 않았다. 다음으로 다국면라쉬모형에 따른 문항반응이론 분석 결과에서는 전체 채점자는 대체로 관대한 채점을 한 것으로 나타난 가운데, 특정 문항이나 평가 준거, 척도에 따라서 엄격하거나 관대한 채점을 한 채점자가 있음을 확인하였다. 집중 경향성 분석에서는 중앙값인 3점에 집중하는 채점자와 최댓값인 5점에 집중하는 채점자가 나타났으며, 무작위성과 후광성 분석에서는 적합도 분석과 점수열 비교를 통하여 해당하는 사례가 있었음을 확인할 수 있었다. 평가 요소와 채점자의 상호작용에 따른 채점자 편향 분석에서는 문항 및 준거별로 다국면라쉬모형의 예상을 벗어난 채점자가 나타났다.

다음으로 말하기 평가 채점 과정의 구성과 특징을 파악하기 위하여 채점 과정 분석을 실시하였다. 분석을 위해 먼저 채점 과정 보고를 전사한 다음 분절하고, 각 발화를 논증 요소로 코딩하였다. 분석 결과, 대부분의 문항에서 채점자들이 순차형 채점을 한 것으로 나타난 가운데 ‘경험 말하기’ 문항에서 중·하위 수준 수험자에 대해 종합형 채점을 한 비율이 상대적으로 많이 나타나고 있었다. 순차형 채점에서는 채점 과정에서 응답의 전반적인 인상을 기준으로 전체적인 점수를 결정하는 경향이 나타났으며, 종합형 채점에서는 구체적인 응답 정보나 특징을 고려하여 점수를 결정하려는 경향이 나타났다. 또한 채점자별로 채점 과정에서 점수 결정을 위해 형성한 근거나 고려하는 가정의 양과 종류의 차이가 나타났는데, 이는 수험자 응답에 대한 지각 수준과 기억 체계 작동의 영향 때문인 것으로 해석하였다. 특히 중위 수준의 응답에 대한 채점 과정에서는 지각 정보를 바탕으로 계량적 접근을 하는 경우가 많이 나타났는

데, 이러한 특징은 중위 수준의 응답이 최상위나 최하위 수준의 응답에 비하여 수준 판단과 점수 결정이 까다롭기 때문에 이를 극복하기 위한 방책으로 지각한 정보를 적극적으로 활용한 것으로 해석하였다. 자료를 활용하는 통합형 문항의 채점 과정에서는 응답의 담화 구성 수준에 대한 고려를 바탕으로 전체 평가 준거에 대한 점수를 결정하는 경향이 나타났다.

다음으로 채점자 영향이 나타난 사례에 대한 채점 과정을 분석하였다. 먼저 엄격한 채점 경향을 나타낸 사례의 채점 과정 분석에서는 '전반적 수행'을 먼저 채점하였을 때 나머지 준거들의 채점에도 같은 경향을 적용하면서 결과적으로 엄격한 경향이 가중된 것으로 판단된다. 채점 과정에서 응답에 대한 증거 수집이 잘 이루어지지 않은 경우에는 상대적으로 관대한 채점을 하거나 무작위적인 채점을 한 것으로 나타났다. 평가 척도를 제한적으로 사용하는 경향이 나타난 경우에는 채점 과정에서 인상 기반 접근과 평가적 가정을 고려하는 양상이 나타났으며, 이는 수험자의 응답 내용을 기존의 기억 체계에 통합하여 처리하려고 하면서 그 밖의 정보는 채점 과정에서 고려하지 않기 때문에 일어난 현상으로 판단된다. 무작위적인 채점 경향이 나타난 채점자의 채점 과정은 다른 채점자에 비해 간결한 구조로 나타난 점이 특징이었는데, 응답 내용에 관한 판단과 점수 결정을 위한 이론적·경험적 가정을 고려하지 않고 채점 척도에 대한 주관적인 해석에 의존하여 채점 과정이 진행되면서 무작위적인 결과가 나타난 것으로 해석하였다.

끝으로 본 연구에서는 연구 결과로부터 도출한 시사점을 바탕으로 채점 과정 기반의 한국어 말하기 평가 채점자 교육의 원리와 방안을 제시하였다. 채점 과정 기반 채점자 교육의 원리는 채점 척도의 내재화, 증거 타당성 확보, 내적 일관성 유지이며, 이와 관련하여 제시한 교육 방안은 채점자가 채점 과정에 대한 분석을 통해 자신의 채점 경향을 성찰하는 활동을 중심으로 구성하였다.

주요어: 한국어, 말하기 평가, 채점 변인, 채점 과정, 채점자 영향, 채점 과정, 채점 과정 보고, 채점 과정 기반 채점자 교육

학번: 2013-30424

목 차

<표> 목차

[그림] 목차

I. 서론

1. 연구의 필요성 및 목적

본 연구의 목적은 한국어 말하기 평가 채점의 변인으로서 채점 과정을 실증적으로 살펴보고, 이를 바탕으로 채점 과정 기반의 말하기 평가 채점자 교육 방안을 제시하는 것이다.

언어 평가 가운데 수행 평가로 이루어지는 말하기 평가는 수험자가 주어진 평가 과제를 수행하면서 발화한 구어 담화에 대해 채점자가 평가 준거를 바탕으로 점수를 부여하고, 이를 활용하는 일을 가리킨다. 이러한 수행 중심의 말하기 평가는 1960년대 이후 의사소통 중심 접근의 확대와 함께 언어 사용 능력 평가도 실제적이며 직접적인 방법으로 이루어져야 한다는 논의를 바탕으로 나타났다. 말하기 평가를 수행 평가[1]로 실시한다는 것은 평가에 관한 수험자의 지식과 수행 과정을 포함하여 평가한다는 의미를 갖고 있다. 이는 이해의 측면을 중심으로 하는 선택형이나 단답형 평가와 달리 언어 생산에 초점을 두는 평가 방법이며, 수험자의 응답이 복잡한 형태의 언어적 구성물로 나타나기 때문에 이를 평가할 수 있는 체제를 요구한다. 이와 관련하여 가장 일반적인 접근 방법은 수준 판정을 위해 전문적인 채점자를 선정 또는 양성하여 평가의 역할을 맡기는 것이다. 그런데 채점자가 채점에 관한 전문성을 갖추기 위하여 훈련을 받는다고 해도, 말하기 평가에서 수험자들이 응답에 나타난 모든 양상을 절대적인 기준을 바탕으로 완벽하게 일치시켜 평가하는 일은 현실적으로 불가능하다. 따라서 수행 평가로서의 말하기 평가에서는 절대적인 답이 존재하여 채점의 일관성을 확보할 수 있는 선택형 평가에 비하여 채점자의 판정 전

1) 수행 평가의 일반적인 특징에 관해 백순근(2002: 49-54)은 첫째, 교사의 전문적 판단에 근거하며, 둘째, 학생이 스스로 답을 구성하며, 셋째, 실제 상황 맥락을 바탕으로 교육목표 달성 여부를 파악하고, 넷째, 교수·학습의 내용뿐만 아니라 과정도 중시하며, 다섯째, 학습 과정 진단 및 촉진을 위한 방법이며, 여섯째, 개인별 평가뿐만 아니라 집단적인 평가에도 적용 가능하며, 일곱째, 종합적인 평가를 목적으로 지속적인 평가를 강조하며, 여덟째, 학생의 인지적 영역뿐만 아니라 정의적 영역과 심동적 영역에 대한 평가를 중시한다고 하였다.

문성이나 점수 산출의 신뢰성 확보에 영향을 미치는 변인들의 간섭을 최소화하여 채점의 타당도를 확보할 수 있도록 해야 한다. 이러한 말하기 평가의 채점 타당도 확보와 관련하여 위어(Weir, 2005: 46)가 제시한 평가 준거/채점 척도(criteria/rating scale), 채점 과정(rating procedure), 채점자(raters), 성적 부여(grading and awarding)의 채점 수행 관련 변인 중에 '채점 과정'은 채점자가 채점 척도를 바탕으로 성적을 부여하고, 이를 분석 및 해석하여 활용 가능한 결과로 도출하는 일로 이루어지기 때문에 채점 관련 변인의 영향이 집약적으로 나타나게 된다. 그러므로 말하기 평가 채점의 타당성을 확보하기 위해서는 채점 과정이 타당하게 이루어져야 하며, 이를 입증하기 위하여 증거의 체계적인 수집과 분석이 이루어져야 한다.

언어 평가에서 채점(scoring/rating)은 평가에 참여한 학습자 또는 수험자가 주어진 평가 과제에 대해 응답한 바에 대하여 특정한 점수나 등급을 부여하여 평가 결과를 도출하는 행위이며, 말하기 평가에서는 평가자가 수험자의 과제 수행을 관찰 및 청취하는 과정을 포함하는 것이 특징적이다. 청취 과정을 통해 수집한 수험자 응답에 관한 정보는 채점자의 인지 체계를 중심으로 즉각적으로 판단이 내려지거나 또는 반성적인 검토를 반복해서 거치면서 판단이 내려질 수도 있다. 이러한 채점의 과정에서 채점자는 정보를 인지적으로 처리하고 점수를 결정하기 위하여 여러 평가 관련 요소들을 고려해야 한다. 이는 채점자가 채점을 수행하면서 이러한 요소들을 충분하고 적절하게 고려하여 수험자에게 타당한 평가 결과를 제공해야 하는 책임을 갖고 있기 때문이다. 따라서 채점자는 채점 과정을 통해 타당한 평가 결과를 도출하기 위하여 자신의 판단과 결정을 보장하는 근거를 확보할 수 있도록 노력해야 한다.

지금까지의 말하기 평가 연구에서 채점의 타당도는 채점자가 채점 수행으로 부여한 점수가 수험자, 과제, 채점자 등의 측면에 따라 나타내는 일관성의 정도인 신뢰도(reliability)로 파악하여 왔다. 신뢰도는 평가 결과에 대한 통계 분석을 통해 파악하는데, 같은 평가를 반복해서 시행하였을 때 안정적으로 일관된 평가 결과를 얻을 수 있을 것이라는 점을 보증하는 지표이다. 그런데 총점을 기준으로 산출하는 신뢰도를 바탕으로 평가 결과를 타당화하는 접근은 시험 시행의 조건이 서로 다를 때 수치의 비교가 불가능하다는 한계가 있다. 이는 신뢰도 산출에 활용하는 자료가 채점자를 통해 수집한 관찰 점수라는 점과

관찰 점수를 구성하는 진점수 외의 오차에 대한 해석이 어렵다는 점에서 기인한다.[2] 이러한 신뢰도 산출과 해석의 문제 외에도 말하기 평가와 같은 수행평가에서 채점 타당도의 확보와 관련하여 고려해야 할 중요한 변인은 채점자 영향(rater effect)이다. 채점자 영향은 평가 결과를 수험자의 과제 수행에 대하여 채점자가 평가 결과를 도출하는 과정에 관여하면서 발생하는 수준 판정 및 점수 결정에 변화를 가져오는 것을 가리킨다. 평가 결과를 채점자가 결정하는 말하기 평가에서 합리적인 평가 결과의 해석을 위해서는 사전에 채점자 영향을 통제하거나, 그 영향의 정도를 파악하여 해석에 반영할 수 있다.

지금까지의 채점자 영향에 관한 연구에서는 채점자의 경력이나 전공, 어학 능력 등의 비모수적 정보에 주목하거나, 문항반응이론 분석을 통하여 채점자 영향의 정도와 특징을 객관화하여 확인하는 접근이 이루어져왔다. 이러한 통계적 접근은 채점자 영향을 수치화하며, 비교 가능한 결과를 얻을 수 있어 채점자 영향의 실제적인 의미 해석에 기여한다는 장점이 있다. 그러나 채점자 개인의 영향과 집단적인 영향, 그리고 채점에 관여하는 특정 변인들과의 상호작용적 측면을 고려하지 않기 때문에 제한적인 접근에 그친다는 한계가 있다. 또한 채점자 영향의 연구를 통해 확인한 채점의 정확성 또는 일관성의 문제는 그 원인이 불명확한데다 처음부터 문제가 나타날 수 있는 조건에서 조작적으로 연구를 실시하여 문제를 확인하기 때문에 순환논증의 오류에 빠질 수 있다는 점을 주의해야 한다. 따라서 채점이 타당하게 이루어졌는지를 알아보기 위해서는 채점자가 평가 결과를 도출하는 과정에서 수험자의 수행에 나타난 어떤 점에 대해 어떻게 반응하며, 무엇을 고려하여 점수를 부여하는지에 대한 직접적인 접근이 이루어져야 할 것이다.

본 연구에서는 채점자 영향을 채점 과정에서 채점자가 여러 인지적 요소들을 처리하면서 나타나는 점수 결정에 대한 체계적인 개입으로 본다. 이러한 관점은 채점자 영향을 채점자의 본성적인 문제나 무능으로 해석하는 것을 경계하고, 특정 평가 맥락 속에서 인지적인 상호작용하는 과정을 통해 나타난 특성

---

2) 심리측정학적인 관점에서 관찰 점수 산출과 관련하여 주목한 오차 요인 중에 평가 문항에 대한 모수로는 곤란도, 변별도, 추측도 등이 있으며, 수행 평가에서는 답안을 추측하여 구성하는 일이 불가능하기 때문에 문항의 곤란도와 변별도를 중심으로 오차를 해석할 수 있다.

으로 보는 것이다. 이러한 채점자 영향의 재개념화는 언어 수행 평가에서 이루어지고 있는 채점자 훈련의 방향이 채점 특성에 대한 통제와 조정에서 채점 과정에 대한 이해와 탐구로 변해야 함을 함의한다.[3] 이와 같은 접근이 이루어지기 위해서는 채점자가 수험자의 과제 응답 청취에서부터 점수를 결정하기까지의 말하기 평가 채점 과정에서 채점자의 머릿속에서 일어나는 일에 대한 파악이 이루어져야 한다.

한국어 말하기 평가 연구에서 채점에 관한 연구는 자료 수집과 분석을 위한 접근법의 한계로 인하여 매우 제한적으로 이루어져 왔다. 지금까지의 연구에서는 주로 채점자의 채점 신뢰도에 영향을 끼치는 배경 변인에 대한 탐구가 중심이었다. 그런데 말하기 평가 결과 도출에 개입하는 여러 평가 요소들 중에 채점자 배경 변인의 문제를 체계적으로 다루기 위해서는 과제나 채점 방법 등의 평가 특성과의 상호작용을 통제한 가운데 배경 변인이 채점 과정에 개입하였음을 검증해야 한다.

한편, 채점자의 말하기 평가 채점 과정에 대한 파악을 위하여 본 연구에서는 채점자의 한국어 말하기 성취도 평가를 중심으로 채점 과정에 나타난 양상을 파악하고자 하였다. 한국어 말하기 성취도 평가는 한국어 교육 현장에서 이루어지는 여러 평가 유형 중에 중간·기말 시험과 같이 형식성이 높은 시험들을 가리키며, 일정한 단위의 학습을 종료한 후에 치르는 종합적인 성격의 시험으로서 가장 보편화된 말하기 평가 양식이다. 성취도 평가에서는 학습자의 수준에 따라 평가 내용의 수준을 달리하여 구성하는 것이 일반적인데 기본적인 언어 지식의 습득 여부를 확인하는 문항에서부터 실제적인 언어 사용 능력을 평가하는 자유 발화 및 과제 기반 발화 문항 등을 통해 평가가 이루어진다. 평가 문항 유형 가운데 발음 구사나 담화 구성과 같은 평가 준거를 바탕으로 채점 척도를 따라 점수를 부여하는 문항에서는 채점자의 적극적인 개입이 이루어진다. 이러한 평가 상황을 통해 제시하는 평가 결과는 학습자의 평가 과제 수행 결과이기 이전에 채점자의 채점 과정 산출물이라고 볼 수 있다.

---

3) 채점자 주관성이나 채점자 배경 변인 영향에 대한 일시적인 통제와 조정으로 이루어지는 채점자 훈련은 단기간의 채점 수행에서는 채점자 간의 일치도를 높일 수 있는 방법으로 알려져 있지만, 채점자의 주관적인 채점 경향은 제거할 수 없으며, 훈련 효과가 일시적이라는 한계가 있다.

이러한 맥락에서 볼 때 말하기 평가의 채점 과정을 확인하는 것은 채점자가 채점을 잘 하고 있는가를 평가하기 위한 접근으로 볼 수 있다. 그러나 기존의 채점자에 대한 신뢰도 중심의 평가적 접근에서 채점의 원인을 다루지 못하고 있는 것과 달리, 채점 과정을 통한 접근에서는 특정 과제에서 특정 학습자의 과제 수행에 대한 특정 채점자의 점수 결정 경향을 확인할 수 있어 보다 객관적인 접근을 할 수 있을 것으로 보인다. 이러한 채점 과정 중심 접근은 말하기 평가의 채점자에게 기대하는 채점 수행 능력의 실체를 명시적으로 제시하여 이를 바탕으로 채점자 교육 및 훈련의 새로운 토대를 구축하는 데 기여할 수 있을 것이다.

본 연구에서는 이상의 문제의식에서 한국어 말하기 평가의 채점 과정에 대한 이론적·실증적 접근을 통해 채점 과정에 대한 규명이 이루어져야 한다고 보았다. 이를 위하여 선행 연구를 바탕으로 채점자의 채점 수행이 이루어지는 채점 과정에 대한 이론적 모형을 수립하고, 실제 말하기 평가 채점 과정에 대한 양적·질적 자료를 수집한 후에 분석하여 채점 과정의 실제적인 양상을 확인하고자 한다. 끝으로 연구 수행을 통하여 살펴본 말하기 평가 채점 과정의 특징을 바탕으로 채점 과정 기반의 말하기 평가 채점자 교육을 위한 원리와 방안을 제시하고자 한다. 연구 목적을 달성하기 위한 연구 문제는 다음과 같다.

첫째, 말하기 평가의 채점 과정이란 무엇이며, 채점자의 채점 과정에서는 어떠한 인지적 처리가 이루어지는가? 채점자의 머릿속에서 이루어지는 말하기 평가의 채점 과정을 파악하기 위해서는 인지적으로 이루어지는 채점에 관한 정보의 처리와 그에 영향을 미치는 여러 영향 요인에 대한 이해가 뒷받침되어야 한다. 이와 관련하여 Ⅱ장에서는 채점을 바라보는 관점으로 언어 평가학과 인지심리학에서의 접근을 살펴보고, 말하기 평가에서의 채점에 관한 국내외 선행 연구 검토를 바탕으로 채점자의 내·외재적 영향 요인을 파악한다. 이러한 이론적 검토를 바탕으로 채점자의 머릿속에서 이루어지는 말하기 평가 채점 과정의 체계와 유형을 제시하고자 한다.

둘째, 말하기 평가 결과를 통해 확인할 수 있는 채점자 영향은 무엇이며, 채점자 영향의 특성이 채점 과정에 대해 시사하는 바는 무엇인가? 말하기 평가에서 채점자는 여러 평가 요소들에 대한 복합적인 고려 가운데 채점 척도를

중심으로 수험자의 응답에 대하여 점수를 부여하는데, 평가 결과인 점수는 채점자의 채점 경향의 차이를 단적으로 확인할 수 있는 증거이며, 채점 과정을 통해 결정된다는 특징을 갖고 있다. 이와 관련하여 Ⅲ장에서는 먼저 말하기 평가 결과에 나타난 채점자 영향을 실증적으로 파악하기 위하여 개발한 한국어 말하기 평가 도구를 제시하였다. 그리고 이를 활용하여 수집한 수험자의 응답을 바탕으로 채점을 실시하고, 수집한 평가 결과에 나타난 채점자 영향을 파악하기 위하여 통계 분석을 실시하였다. 통계 분석은 고전검사이론과 문항반응이론을 통한 접근으로 이루어지는데, 고전검사이론 분석은 채점 과정을 통해 발생한 채점자 영향의 전반적 특징을 파악하기 위하여 실시하며, 문항반응이론 분석을 통해서는 채점 과정을 통해 나타난 채점자 영향이 구체적으로 무엇이며, 채점자가 채점 과정에서 고려하는 과제, 준거 등의 평가 국면과 관련하여 어떤 관계에 있는지를 확인하고자 하였다.

셋째, 말하기 평가의 채점 과정은 어떻게 이루어지며, 이는 과제 유형 및 채점자 영향 유형에 따라 어떠한 차이가 있는가? 그리고 채점 과정에 대한 분석 결과가 말하기 평가의 채점자 특성에 관해 시사하는 바는 무엇인가? 말하기 평가 채점 과정의 양상을 체계적으로 살펴보는 한 가지 방법은 같은 채점 대상에 대한 채점자별 채점 과정의 차이를 체계적으로 비교하는 것이다. 본 연구의 Ⅳ장에서는 이와 관련하여 채점자들로부터 수집한 채점 과정 보고를 바탕으로 말하기 평가의 채점 과정을 체계적으로 분석하기 위하여 전사한 후에 키엔포인트너(1992)의 논증 도식을 바탕으로 주장, 근거, 전제로 코딩하여 채점 과정 구성의 차이를 알아보고자 하였다. 이를 바탕으로 이루어진 전체 문항 유형별 채점 과정 분석에서는 평균 점수를 기준으로 중위·최상위·최하위 수험자의 응답에 대한 전체 채점자의 채점 과정을 분석하여 문항 유형에 따른 채점 과정의 차이를 확인하고자 하였다. 또한 다국면라쉬모형 분석에서 확인한 채점자 영향이 나타난 채점 사례에 대한 채점 과정 분석에서는 같은 대상에 대한 채점 사례에서 상대적인 결과를 제시한 채점자의 채점 과정을 비교하여 평가 결과의 변인으로서 채점 과정의 영향이 무엇인지를 실증하고자 하였다.

넷째, 채점 과정 기반의 채점자 교육이란 무엇이며, 어떻게 이루어져야 하는가? 말하기 평가 채점 과정을 파악하기 위하여 실시한 양적·질적 자료 수집 및 분석과 그 결과는 채점자의 말하기 평가 과정을 탐색하기 위한 체계적인 접근

방법이 될 수 있다. 또한 연구를 통하여 파악한 평가 과제 유형 및 채점자 영향에 따라 달라질 수 있는 채점 과정과 평가 결과에 관한 정보는 말하기 평가 채점자가 자신의 채점 경향을 성찰하기 위한 교육에서 다루어야 하는 내용이 될 수 있을 것이다. 이에 본 연구에서는 말하기 평가의 채점자들이 자신의 채점 특성을 성찰하고, 채점 타당도를 확보할 수 있도록 하는 채점자 교육이 필요하다고 보고, 이와 관련하여 Ⅴ장에서 채점 과정 기반 채점자 교육의 원리와 방법을 제시하고자 하였다.

2. 연구사

본 연구에서는 말하기 평가의 채점 변인에 대한 탐구를 위하여 먼저 한국어 교육에서의 말하기 평가에 관한 연구사적 흐름을 살펴본 후에 채점 과정과 관련하여 채점자 영향과 채점 과정의 양상, 그리고 채점자 훈련에 대한 국내외 연구를 살펴보았다. 그리고 선행 연구 검토를 바탕으로 한국어 말하기 평가 채점 과정 연구를 위한 설계의 토대를 마련하고자 하였다.

1) 한국어 말하기 평가 연구

한국어교육 연구에서 말하기 평가에 관한 논의는 재외동포 자녀를 위한 한국어 평가의 목표와 내용을 제시한 노대규(1983)의 연구에서 발화 능력 시험에 관한 언급을 통해 처음 나타난 것으로 보인다. 연구사 초기에는 체계적인 한국어 말하기 평가 도구 개발 및 시행의 필요성을 바탕으로 국외 말하기 평가 도구 사례를 참조하여 한국어 말하기 평가에 적용하는 접근이 이루어졌다(김정숙·원진숙, 1993; 정광 외, 1994; 전은주, 1997; 정화영, 2000).

김정숙·원진숙(1993)은 한국어 말하기 평가의 구인 설정에 관한 논의와 이를 교수·학습에 적용하기 위한 방안을 제시하였는데, 수준에 따라 구인별 비중의 차이를 두는 관점을 제시하였다는 의의가 있다. 정광·고창수·김정숙·원진숙(1994)은 언어 능력 및 의사소통 능력에 관한 이론과 숙달도 평가에 관한 ACTFL OPI의 평가 체제를 검토하여 통합적인 한국어 능력 시험에 대한 방안을 제안하였는데, 수준별 평가 체계로 발화 단위 및 유형, 내용 및 범주, 기능,

정확성의 수준을 제시하였으며 또한 말하기 평가의 실행을 위한 시험 절차와 등급별 문제 유형을 제안했다. 전은주(1997)는 이전의 연구들이 이론적 바탕으로 삼았던 ACTFL의 평가 체계에 대한 문제를 지적하고, 한국어 말하기 평가에 맞는 평가 범주 설정을 제안하였다. 정화영(2000)도 영어 말하기 평가 시험(FSI)을 기초로 하여 한국어 말하기 숙달도 평가에 관한 연구를 하였는데, 직접 FSI 말하기 시험을 실시하고, 시험 중 이뤄진 대화에 대한 담화 분석을 실시하였으며, 사후 설문지 조사와 면담을 통해 그 효용성을 밝혔다는 것이 특징적이다.

2000년대에 들어서는 국외 연구 사례에 대한 검토와 함께 한국어교육 기관에서 이루어지고 있는 말하기 평가에 관한 비판적 검토와 현대적인 평가 방법을 제안하는 연구가 나타났다(지현숙, 2005; 최은규, 2006; 윤은경, 2006; 전나영 외, 2007). 지현숙(2005)은 국내 한국어교육 기관에서 수행한 성취도 평가인 학기말 말하기 시험 중 인터뷰 담화를 분석하였는데, 기존의 평가들이 의사소통 능력에 초점을 두었던 것과 달리 상호작용능력에 초점을 두고, 시험 담화의 구조, 말순서 취하기(turn-taking), 인접쌍(adjancey pairs), 주제 주도성을 분석하였다. 최은규(2006)는 5개 한국어 교육 기관에서 시행하고 있는 배치 시험과 성취도 평가의 유형과 평가 영역(채점 기준)에 대한 조사를 실시하여 한국어 능력 평가의 이론적 토대를 다지는 연구를 수행하였다. 전나영 외(2007)는 소속 기관의 말하기 능력 평가 도구 개발에 관한 연구 수행의 결과를 논문으로 발표하였는데, 영어, 일본어, 유럽권 언어 시험 및 한국어 시험들과 영어 말하기 시험들에 대한 검토와 기존의 한국의 말하기 평가 관련 선행연구를 바탕으로 말하기 능력 평가 도구의 평가 범주와 평가 방법, 평가 문항 예시, 채점 기준표를 등급별로 제시하였다. 윤은경(2008)은 ACTFL OPI에 대한 비판적 연구들을 검토하고, 이를 한국어 말하기 능력 평가 연구에 적용할 방안을 도출하였다.

2010년 이후의 한국어 말하기 평가 연구에서는 말하기 평가에 대한 상황적·이론적 접근을 넘어서 실제적인 평가 도구 개발과 그와 관련된 실증 연구가 나타나기 시작하였으며(김경선·이규민·강승혜, 2010; 이향, 2013; 김상경, 2015; 민병곤 외, 2017), 성취도 평가와 관련된 연구(이준호, 2009; 강현주, 2013; 나카가와 마사오미, 2014), 그리고 말하기 평가의 타당화에 관한 연구(이성준,

2018; 박현정 외, 2018; 오승영, 2019)가 이루어졌다.

김경선·이규민·강승혜(2010)는 한국어 말하기 성취도 평가의 실제 자료를 바탕으로 통계적 검증 과정을 통해 오차 요인을 분석하고, 일반화가능도 이론에 근거하여 효율적인 평가의 조건을 모색하였다. 이향(2013)은 말하기 수행 평가의 발음 범주 채점에 대한 타당성을 검증하는 것을 목적으로 하며, 이를 위해 발음 채점에 대한 이론 기반 타당도와 사전·사후 타당도를 검증하였다. 김상경(2015)은 학문 목적 말하기 평가 도구를 직접 개발하는 연구를 수행하였는데, 실제로 도구를 개발하고 실험과 검증과정을 포함하고 있다.

강현주(2013)는 말하기 성취도 평가에서 상호작용능력에 대한 평가의 필요성을 바탕으로 평가 범주와 기준, 원리와 과제 등을 제시하였다. 이 연구에서는 상호작용능력에 대한 평가자들의 인식에 관해 설문 조사를 실시하였으며, 상호작용 능력 평가를 위한 대화 과제 수행 결과를 분석하고 채점하는 평가의 과정을 제시하고 있다. 나카가와 마사오미(2014)는 일본에서의 한국어 성취도 평가연구를 수행하였는데, 설문조사와 담화 분석, 사후 면담을 활용하여 연구를 진행하였다. 이 연구에서는 한국어 말하기 수행 평가의 필요성과 방향을 제시하고, 실제 교수·학습에 적용할 수 있는 평가 과제의 평가 기준 및 방법을 구안하였으며, 학습자와 교수자의 평가 후 피드백을 함께 조사하여 제시한 것이 특징이다.

한국어 말하기 평가에 관한 타당도 문제를 다룬 연구 중에서 이성준(2018)은 한국어 말하기 평가의 체계적인 개발과 타당한 결과 활용을 연계하는 방법으로 논거 기반 접근법을 적용하여 한국어 교실 평가와 대규모 평가에 적용하는 방안을 제시하였다. 박현정 외(2018)의 연구에서는 학문 목적 한국어 말하기 평가의 대규모 검사에 대하여 문항반응이론을 적용한 분석 결과를 바탕으로 검사 내적 타당도 증거로 채점자 영향을 분석하고, 공인 타당도 증거로서 수험자 배경 및 한국어 능력에 대한 상관을 제시하고 있다. 오승영(2019)에서는 이주민 대상 한국어 말하기 평가 도구에 대한 맥락 타당도를 검증하는 연구를 수행하였다.

이상의 내용을 종합하여 보았을 때, 지금까지의 한국어 말하기 평가 연구는 대규모 평가를 위한 시험 개발을 목적으로 국외 평가도구들을 벤치마킹하거나 구체적인 평가 구인과 과제의 특성에 관한 이론적 연구, 그리고 평가 결과에

나타난 통계적 분석 및 담화 분석을 통한 실증적인 접근으로 이루어져왔다. 이러한 접근 속에서 나타난 평가 목적과 대상 설정의 부재와 평가 구성 요소에 관한 오개념 사용, 교육적 환류를 고려하지 않는 한계는 한국어 말하기 평가 연구에서 지속적으로 다루어야 할 과제라고 할 수 있다(이성준, 2015).

2) 말하기 평가의 채점자 영향 연구

채점자 영향(rater effect)은 말하기 평가에서 채점자가 채점을 수행하면서 나타나는 주관적 채점자 영향으로서, 심리학적으로는 “수행에 대한 채점에서 채점자로 인한 체계적 변화에 따른 광범위한 영향”(Scullen, Mount, & Goff, 2000)이며, 평가하는 특성에 대한 측정의 관점에서는 “측정 오류를 내포하는 채점자의 채점 경향”(Wolfe & McVay, 2012:32)이라고 볼 수 있다.[4] 채점자 영향에 관한 연구는 수험자(McNamara & Lumley, 1997; Meiron, 1998; Pollitt & Murray, 1996; ; Seong & Bottcher, 2011), 채점 척도(김평원, 2010; 김현아, 2016), 평가 방법(주미진, 2014)같은 평가 맥락적 요인과 채점자의 경력(원미진·김지영, 2017; Isaacs & Thompson, 2013; Winki, Gass, & Myford, 2012), 전공(강석한·안현기, 2014; 이향, 2013; Isaacs & Trofimovich, 2011), 언어적 배경(김현주, 2007; 이향, 2013; Gass & Varonis, 1984; Kim, 2009a, 2009b; Matsugu, 2013; Zhang & Elder, 2011; Joe, Harmes, & Hickerson, 2011)과 같은 배경적 특성에 따른 연구, 그리고 채점자의 채점 경향에 대한 연구(Myford & Wolfe, 2004)로 이루어져왔다.

수험자의 응답 특성의 영향을 주목한 Pollitt and Murray(1996)는 수험자의 응답에서 원숙함과 발화 의지가 나타났을 때 채점에서 긍정적인 효과가 있다고 하였으며, Meiron(1998)은 창의적이고 유머러스한 발화가 비의사소통적 과제에서 효과적이었다는 점을 발견하였다. 응답의 음향적인 측면에 관한 문제를 다룬 McNamara & Lumley(1997)는 채점자들이 음성 자료의 녹음 수준의 문제로

4) 채점자 영향을 오류로 보는 관점에서는 진점수가 포함하는 고정적 오차(constant error)로 보기도 하는데, 일관된 영향을 나타내는 경우에 해당하기 때문에 채점 척도의 이상으로 인한 문제 또는 채점자의 평가 목적이나 측정 구인에 대한 근본적인 이해 부족과 같은 문제에서 비롯된다. 채점자별로 다양한 채점 과정을 통해 이루어지는 가변적인 채점 행위는 채점 과정에 비체계적으로 영향을 미칠 수 있다는 점에서 무선적 오차(random error)도 나타날 수 있다.

잘 들리지 않는 응답을 불완전한 응답으로 인식하여 상대적으로 엄격하게 채점하는 경향을 나타냈음을 확인하였다. Seong & Bottcher(2011)에서는 채점자들이 중급에게 관대하고 고급에게는 엄격한 경향이 나타났으나, 가르쳐보지 않은 대상을 채점하는 경우에는 중급에서 엄격하고 고급에서는 관대한 경향이 나타났음을 보고하였다.

Gass & Varonis(1984)는 채점자의 비원어민 화자 발화에 대한 친숙성에 관한 연구를 수행하였는데, 평가 결과에서 화제에 대한 친숙성의 영향이 가장 컸으나, 비원어민 학습자 발화에 대한 친숙성이 이해를 촉진시키고 보다 관대한 평가를 내리도록 했음이 나타났다. 김현주(2007)는 영어 말하기 평가에서 채점자의 언어 배경의 영향을 알아보기 위하여 ENL, ESL, EFL 각각의 환경에서 영어를 가르치는 총 65명의 교사들에게 6개의 수험자 발화 자료를 제공하고, 이를 총체적 채점과 분석적 채점으로 각각 채점한 후에 구조방정식 모형을 바탕으로 EFL > ESL > ENL의 순으로 수험자의 비표준 영어에 대한 관대성의 차이가 있으며, 평가 준거로는 문법과 발음, 원어민성이 채점자 언어 배경에 따른 영향을 받을 수 있는 것을 확인하였다. Kim(2009a)에서는 원어민 채점자와 비원어민 채점자의 말하기 평가 결과에서 영어 원어민 교사들은 비원어민 교사들에 비해 발음, 문법, 정확성과 관련하여 세밀하고 정교하게 채점하려는 경향이 있음을 확인하였다. 또한 후속 연구에서는 과제에 따라서 원어민 채점자와 비원어민 채점자가 중요하게 고려하는 평가 구인과 척도에 차이가 있음을 확인하였다(Kim, 2009c). Zhang & Elder(2011)는 중국에서의 영어 말하기 평가에서 평가 결과에 나타난 채점자들의 특성을 다국면 라쉬 모형을 통하여 분석하고, 사후 면담을 통해 이유를 확인하였는데, 원어민 채점자들이 의사소통적 문제를 주목하여 채점을 하는 반면에 비원어민 채점자들은 형태에 중심을 두어 상대적으로 비의사소통적인 측면으로 채점하는 경향이 있음을 확인하고, 채점자들의 교육 경험의 차이를 원인으로 지목하였다. Matsugu(2013)에서는 채점자들이 학습자를 알고 있는지의 여부에 따라서 발화 채점의 차이를 알아보았는데, 분산분석 결과 유의한 차이가 나타났으며, 학습자 관련 지식이 보상적인 채점을 유도하였을 가능성을 제기하였는데, 이 연구에서는 교사와의 사후 면담을 실시하여 관대한 채점의 원인으로 학습자 문화권에 대한 이해와 자신의 지위에 대한 지속성 확보 등의 영향에 주목하였다. 윈크, 개스, 미포드(Winke,

Gass, & Myford, 2012)는 말하기 평가 채점에서 청취자인 채점자에게 친숙한 수험자의 억양이 미치는 영향을 연구하였는데, 특정 채점자 집단에서 친숙한 특정 언어권 수험자에게 눈에 띄게 관대한 채점을 하였음을 확인하였다. 이 연구에서는 채점자 영향의 원인으로 채점 과정에서 수험자 응답 청취와 점수 결정에 수험자 L1 친숙성이 관여할 수 있다는 점에서 이를 상쇄할 수 있는 방안으로 채점자 훈련에서 채점자별로 친숙한 언어권의 영향 정도를 확인하고, 그 영향이 두드러진 경우에는 억양 친숙성에 관한 민감도를 올리는 훈련이 필요하다고 하였다.

채점자의 경력이 말하기 시험 채점에 미치는 영향에 관한 연구에서는 채점자 경력이 많을수록 일관된 채점의 경향이 나타나며, 평가에 관한 전문 지식을 활용하여 응답에 대한 표현을 적절하게 할 수 있는 능력을 갖고 있는 것으로 나타났다. Isaacs & Thompson(2013)은 영어 발음 평가 연구에서 채점 경험이 많은 채점자들에게서 초인지 전략을 보다 적극적으로 사용하는 양상이 나타났으며, 이들은 세부적인 발화의 문제들을 구체적으로 표현하고 지적할 수 있는 능력이 있다고 하였다. 원미진·김지영(2017)은 한국어 말하기 평가의 훈련 효과 및 신뢰도 향상을 위한 연구에서 채점자의 한국어교육 경력에 따른 영향을 확인하였는데, 경력자 집단에서는 초급은 관대하고, 중·고급으로 올라갈수록 엄격한 경향이 나타난 반면에 무경험자 집단에서는 초급은 엄격하게 채점한다는 점을 확인하였다. 측정 구인의 측면에서는 경력자 집단은 내용과 조직을 엄격하게 채점한 반면에 무경험자 집단은 내용적 측면에는 관대하고 조직에 관해서는 엄격한 경향이 나타났다. 이 연구에서는 이와 같은 특징으로부터 채점자 교육을 통한 지식 제공의 필요성을 주장하였는데, 엄격성이나 관대성의 차이가 채점자의 경력에서 비롯되었다고 볼 수 있기 때문에 지식 제공뿐만 아니라 지식에 대한 이해를 높일 수 있는 구체적인 방안이 필요할 것으로 보인다. 조, 함즈, 힉커슨(Joe, Harmes, & Hickerson, 2011)은 말하기 시험에서의 채점 경험에 따라 채점 과정의 인지적 특성에 차이가 있는지를 알아보기 위하여 채점자들에게 녹화된 수험자의 과제 수행 장면을 시청하면서부터 채점을 마칠 때까지의 과정을 보고하도록 하였다. 연구 결과에서 채점자들은 채점 과정에서 채점 척도를 참조하지 않는 경향을 확인하였는데, 이는 채점자들이 복잡한 채점 척도를 단순하게 적용하려고 하며, 채점자 자신의 내적 채점 규칙을 적용하

고 있다는 것으로 볼 수 있다. 이 연구에서는 숙련된 채점자와 숙련되지 않은 채점자의 차이가 부여하는 점수의 다양성에 관한 측면과 주목한 응답 특성의 다양성에 있음을 확인하였다.

채점자들의 전공에 따른 전문성이 채점에 미치는 영향과 관련하여 Isaacs & Trofimovich(2011)은 L2 발음 채점에서 채점자들의 인지적 차원에 대한 접근으로 음성학적인 부분의 인식과 관련하여 기억력, 주의 통제력, 음악적 능력에 따라 발음 평가에서 차이가 난다는 점을 확인하였는데, 억양(accentedness)에서 민감도가 높은 음악을 전공한 채점자들이 음악을 전공하지 않은 채점자에 비하여 더 엄격한 채점을 한 것을 확인하였다. 이 연구에서는 음악적 능력에 비해 기억력과 주의 통제력에 대한 측정이 발음 판정과는 관련성이 적어 측정 조건 간의 균형이 맞지 않다는 한계가 있었다. 이향(2013)은 한국어 발음 평가에서 채점자들의 전공에 따른 채점 경향의 차이를 Facets 분석을 통하여 알아보았는데, 경력 5년 이상의 채점자들 중 음운론 전공자들의 발음 정확성에 대한 채점에서 엄격성이 가장 높았음을 확인하였다. 강석한·안현기(2014)의 연구에서는 한국어 말하기 평가의 전문가 집단과 비전문가 집단의 채점 경향의 차이를 Facets을 통해 알아보았는데, 문법과 발음 채점에서 채점자 집단 간의 채점 엄격성의 차이가 나타났으며, 과제에 따라서도 집단 간의 차이가 있음을 확인하였다. 이 연구에서는 전문가 집단이 개방형 과제에서 더 엄격하고, 그림 설명 과제와 같은 폐쇄형 과제에는 관대한 경향이 나타난 반면에 비전문가 집단은 반대의 경향이 나타났다고 하였다. 이 연구는 전문성에 따른 채점자 집단 구분이 다소 모호하며 집단 간 차이의 원인을 알 수 없다는 아쉬움이 있지만, 채점자의 경력에 따라 전문성을 구분하여 평가 과제에 따른 채점자 영향의 차이를 실증적으로 확인하였다는 의의가 있다.

채점자의 척도 사용에 따른 영향에 관하여 김평원(2010)은 국어 말하기 능력 평가에서 전문가와 비전문가의 평가 척도 적용 양상의 차이에 대한 퍼지 분석을 바탕으로 모호한 채점 척도를 명료화하는 방안을 제시하였다. 김현아(2016)는 Facets를 활용하여 한국어 말하기 시험에서의 원어민 채점자와 비원어민 채점자의 채점 경향에 대한 훈련의 효과를 알아보는 연구를 하였다. 비원어민 채점자들에게서는 원어민 채점자에 비하여 총체적 채점에서는 관대함이, 분석적 채점에서는 정확성과 풍부성 범주에서의 엄격성이 나타나던 것이 채점자 훈련

을 거친 후에 상쇄되는 경향이 나타났음을 확인하였는데, 이 연구에서는 사전 채점에서는 구체적인 채점 기준표 없이 평가 구인과 척도만을 제시하고, 사후 채점은 채점 척도 교육과 훈련을 적용하는 방식을 택하고 있어 사전·사후 검사를 비교하는 조건이 일관되지 못하다는 한계를 갖고 있다.

주미진(2014)에서는 영어 말하기 평가에서 면대면 방법과 컴퓨터 기반 방법의 채점자 영향을 알아보는 연구를 수행하였는데 연구 결과에서 기존의 연구 결과들과 달리 컴퓨터 기반 평가가 면대면 평가에 비해 채점 일관성이 부족한 것으로 나타났으며, 이를 채점자 피로도의 영향이라고 해석하였다. 또한 관대한 채점자의 특징으로 원어민이며, 나이가 적고, 학사 학위자이며, 비전공자라고 하였는데, 이 또한 기존의 연구 결과들과 상반된 결과였다. 이러한 결과가 나타난 원인을 연구 설계를 바탕으로 추정해 보면, 먼저 25명의 학생에 대한 직접식 평가 후에 준직접식 평가를 실시하면서 채점 수행에 신체적·인지적 부담이 증가하였을 것으로 보인다. 또한 채점자 배경 변인 분석 결과에서 관대한 채점을 한 집단이 원어민이라는 점을 중심으로 이해할 때 수용 가능한 결과이며, 나이나 전공은 무선적인 영향 요인이다.

그밖에 채점자 영향이 다른 평가 맥락 요인들과의 관계에서 갖는 설명력을 알아보는 연구도 있었는데, Lee(2006)는 TOEFL 말하기 시험에서 다양한 변수들의 영향에 관한 양적 연구를 수행하였는데, 일변량 분산 분석 결과에서 수험자(51.3%), 과제(27.9%)의 설명력에 비해 채점자(1.8%) 변량의 설명력은 매우 낮은 수준으로 나타났다. 이 연구에서는 채점자 영향이 적게 나타난 것과 관련하여 사용한 과제의 수준들이 서로 비슷하였고, 채점자 엄격성의 차이로 인한 설명력 상쇄, 그리고 높은 채점 일관성의 영향 때문이라고 보았으며, 이는 측정의 표준 오차(SEM)를 따라 신뢰도를 산출하는 접근을 활용하는 연구에서 생각해야 할 지점을 알려준 것이다.

3) 말하기 평가의 채점 과정 연구

말하기 평가의 채점 과정에 관한 연구는 채점자가 수험자의 과제 수행에 대해 최종 점수를 결정하기까지의 과정을 심층면담이나 사고 구술법, 또는 언어 프로토콜을 활용하여 접근이 이루어져 왔다(김지영, 2018; 김평원, 2010; 송민

영 · 이용상, 2015; Brown, 2000, 2006; Brown, Iwashita & McNamara, 2005; Kim, 2015; Wei & Llosa, 2015). 채점자들은 채점을 할 때 수험자의 과제 수행에 대한 인식과 해석, 판단에 관여하는 채점자의 이전 교육 및 평가 경험의 영향을 받으며, 이러한 경험을 통해 구성한 내재적인 평가 참조틀(frame of reference)은 채점자의 채점 과정을 변화시킨다.

브라운(Brown, 2000)은 자극 회상법(stimulated verbal recall)을 통해 IELTS의 말하기 시험에서 채점자의 채점 수행 과정에 대한 질적 연구를 수행하였다. 이 연구에서는 종합적인 채점 척도 사용의 문제로 언어 수행 수준에 따른 응답 특성에 대한 인식에 차이가 있음을 확인하였다. 예를 들어 낮은 수준의 화자에게서는 이해가능도나 언어 산출에 관심을 기울이며, 문제가 있을 때에만 인지적인 주목을 한다는 것이다. 또한 채점자에 따라 우선시 여기는 특성의 차이가 있으며, 이는 과제 수행만을 기반으로 하는 채점과 이를 맥락화하여 추론 기반으로 접근하는 양상이 있음을 확인하였다. 흥미로운 점은 최근에 채점 훈련을 받은 채점자의 경우 수험자에 대한 추론이 적고, 채점 척도에 나타난 응답 특성을 구별하여 개별적으로 초점을 두는 모습을 확인하였다. 브라운(2006)에서도 자극 회상법을 사용하여 IELTS 말하기 시험의 분석적 채점 척도 사용의 타당도를 조사하는 연구를 수행하였다. 말하기 시험 채점 과정에서 분석적 채점 척도를 사용한 것과 관련하여 채점자들은 종합적 척도에 비해 채점 기준표에 근접한 채점을 하려는 경향이 나타났으며, 더 쉽게 해석하고 적용했다고 하였다. 또한 분석적 척도 사용에 따른 각 채점 기준별 수준 구별의 어려움과 기준별 특성 사이의 효과로 인한 어려움이 있으며, 특히 유창성과 응집성 척도에 대한 판정이 가장 어려운 것으로 나타났다고 보고하였다.

브라운, 이와시타, 맥나마라(Brown, Iwashita & McNamara, 2005)는 TOEFL의 말하기 시험의 과제가 수험자 응답과 채점자에게 미치는 영향을 알아보는 연구를 수행하였다. 이 연구에서는 채점 과정에서 채점자가 과제를 어떻게 고려하는지를 알아보기 위하여 언어 보고 프로토콜(verbal report protocol)을 적용하여 채점 과정을 보고한 자료를 분석한 결과, 채점자들은 과제의 유형에 따라 채점의 내용과 활용 자원의 범주가 다른 것을 확인하였다.

김평원(2010)은 전문가와 일반인의 말하기 평가 과정에 대한 프로토콜 분석을 바탕으로 채점자의 평가 과정 모형을 제시하고 있는데, 전문가의 경우에는

총체적 판단과 분석적 판단을 조화롭게 사용하는 경향이 나타난 반면에 일반인의 경우에는 총체적 판단과 분석적 판단이 순차적으로 이루어졌다. 또한 전문가들은 감성적 판단과 이성적 판단을 복합적으로 사용한다면, 일반인은 감성적 판단만을 내리고 있었고, 객관적 판단과 주관적 판단을 동시에 고려하면서 상호주관성을 나타내는 전문가와 달리 일반인은 주관적 판단만을 고려한다는 점을 확인하였다. 이 연구는 전문가 집단과 비전문가 집단의 채점 과정에서 수준 판정을 위한 인지적 활동의 구조를 모형으로 제시하였으며, 또한 채점 전문성에 따라 채점 경향의 차이가 있음을 보여주었다는 의의가 있다.

킴(Kim, 2015)은 ESL 학습자들의 말하기 평가에서 채점자들을 채점 관련 경력에 따라 초보, 개발 중, 전문으로 나누고, 집단별 채점 척도 사용 경향을 조사하였다. 그 결과 각각의 집단에서 채점자들의 채점 능력이 구별되며, 서로 다른 채점 수행의 진보를 확인하였는데, 그 원인으로 채점에 대한 경험의 양과 그에 따른 루브릭 내재화의 차이에 주목하였다.

웨이와 로사(Wei & Llosa, 2015)는 TOEFL 말하기 시험에서 인도인 수험자에 대한 미국인 채점자와 인도인 채점자의 차이를 조사하기 위하여 다국면 라쉬 분석(MFRM)과 언어 보고 분석(VPA)을 실시하였다. 연구 결과 채점자들은 신뢰도 측면에서 통계적으로 유의한 차이가 없었으나, 채점 과정에서 인도인 채점자들은 인도인 응시자들의 반응을 여러 평가 준거와 관련하여 잘 파악하는 경향이 나타났다. 또한 채점자의 목표 언어에 관한 정체성이 채점에 영향 미친다는 것을 확인하였다.

송민영·이용상(2015)은 한국 고등학생을 대상으로 실시한 대규모 영어 말하기 평가에서 채점자의 채점 행동 특성에 관한 연구를 수행하였는데, 채점 신뢰도를 기준으로 상·하위 집단을 구분하고, 각 채점자들에게 채점 수행 후에 점수 결정의 이유에 대해 사후 보고식 사고 구술을 요청하였다. 연구 결과에서 상위 수준 채점자의 특징으로 노트하기, 과제 완성도 기준의 점수 조정, 명확한 판정 기준 적용을 확인하였으며, 하위 수준의 특징으로는 과제 완성도를 기준으로 나머지 점수를 조정하는 전략이 없었다는 점과 평가 준거별 점수가 상호 간섭하는 현상, 그리고 모호한 채점 기준 및 척도 적용이 나타나는 것을 확인하였다. 이 연구는 점수 결정에 대한 근거로 채점 척도 사용 외에도 채점 과정 중 행동 및 인지적 정보 처리 양상에 대한 접근을 시도하였다는 의의가 있

으나, 신뢰도를 기준으로 채점자 집단을 구분하는 근거가 부족하고, 하위 집단으로 분류한 채점자들에게만 채점 교육을 실시하였기 때문에 자료 수집 조건의 차이가 있었으며, 사고 구술 과정이 연구자를 마주 보고 있는 상황에서 이루어져 그로 인한 영향을 고려하지 않았다는 한계가 있다.

김지영(2018)은 한국어 말하기 평가에서 채점자의 채점 경향을 양적·질적 방법을 병렬적으로 연계하는 통합방법연구를 수행하였는데, 질적 연구 수행 결과에서 채점자별 심층 면접을 통해 채점자들이 양적 연구에서 나타낸 채점 경향의 원인을 확인하였다. 이 연구는 말하기 평가 채점 수행에 관한 통합방법연구 접근을 통하여 풍부한 자료 수집과 분석 결과를 제시하였다는 점에서 의의를 갖고 있으나, 수집한 질적 자료가 회상적 사고 구술 및 문제 원인에 대한 반성적 면담을 토대로 하고 있어 자료의 타당성 확보가 필요하고, 연구 결과를 논증하지 않고 있다는 한계가 있다.

4) 말하기 평가의 채점자 훈련 연구

말하기 평가에서 채점자 훈련의 영향을 다룬 연구들(Brown, 1995; Lumley & McNamara, 1995; Myford & Wolfe, 2000; Orr, 2002; Papajohn, 2002; Davis, 2012; Davis, 2016)에서는 채점자 훈련이 일관성을 높이는 데는 기여하지만, 엄격성을 완벽하게 제거할 수는 없음이 밝혀져 왔다. 채점자의 엄격성이 높을수록 보수적으로 채점하기 때문에 상대적으로 낮은 수준으로 점수를 부여하는 경향이 나타난다.

브라운(Brown, 1995)에서는 일본어 사용 관광 가이드의 말하기 능력 평가에서 채점자의 훈련이 미치는 영향을 확인하였는데, 채점자가 이전 경험을 바탕으로 형성한 평가 준거에 관한 내재적 인식은 채점 훈련에서 제거할 수 없음을 확인하였다. 럼리와 맥나마라(Lumley & McNamara, 1995)는 호주의 직업목적 영어 시험(OET)의 말하기 시험에서 채점자의 훈련 횟수의 영향과 상호작용 효과를 알아보았는데, 채점자 훈련의 2회와 3회 사이에 큰 차이가 나타났다. 이는 채점자 훈련이 매번의 평가 시행에서 반복적으로 이루어져야 하는 당위성의 근거를 제시한 것이라고 볼 수 있다.

미포드와 울프(Myford & Wolfe, 2000)는 TSE에서 훈련받은 채점자의 영향이

어떻게 나타나는지를 MFRM 분석을 통해 알아보았는데, 엄격성의 영향이 점수 척도 전체적으로 발생한다는 점을 발견하였다. 이는 채점자가 엄격하다는 의미를 평가 전반에 대한 의미로 해석할 수 있는 근거를 제시한 것으로 볼 수 있으나, '훈련받은 채점자'라는 점에서 훈련을 통해 일시적으로 특정 수행 특성에 인지가 강화되어 나타난 현상일 수 있기 때문에, 훈련의 시기에 따라 채점 과정에서 과제와 구인의 상호작용으로 인한 영향의 정도를 확인할 필요가 있다.

오어(Orr, 2002)는 UCLES의 FCE 말하기 시험에서 채점자들이 채점 과정에서 수행한 언어 보고를 분석하였는데, 훈련을 받은 채점자들은 채점 과정에서 채점 기준을 주목하는 부분에 차이가 있었으며, 평가 준거와 무관한 정보에 주목하는 모습도 나타났다. 또한 같은 점수를 부여하면서도 전혀 다른 인식을 바탕으로 이루어진 채점 사례가 나타났음을 확인하였다. 이러한 채점 과정과 평가 결과의 불일치 현상은 평가 결과 해석의 타당성을 저해하며, 말하기 평가의 목적을 기계적으로 실현한다는 비판을 받는 원인이 된다.

파파존(Papajohn, 2002)은 학문 목적 영어 말하기 평가(SPEAK) 채점자들의 '개념 지도 산출'에 관한 연구에서 채점자들마다 중요하게 생각하는 원칙과 수준의 차이가 있음을 채점 과정에 나타난 개념 지도 분석을 통하여 확인하였다. 이 연구에서는 채점자들의 채점 과정에 관한 언어 보고의 관계를 구조화하여 제시하는 접근을 하였다는 점과 그것의 교육적 효과를 알아보았다는 점에서 의미가 있다.

데이비스(Davis, 2012)는 영어 숙련 교사들이 채점자 훈련 전후에 TOEFL 말하기 시험 응답을 채점하고, 자극 회상법(stimulated recall)을 활용하여 채점 과정에 대한 회상적 보고를 수집하여 채점 훈련의 영향에 관한 연구를 진행하였다. 먼저 통계 분석 결과에서는 채점자 훈련 후에 채점자 간 신뢰도와 일치도가 상승하는 효과를 확인하였다. 이 연구에서는 능숙한 채점자의 특징으로 기준 응답(benchmark response) 사례를 더 자주 검토하며, 점수 결정의 시간이 더 길다는 것을 확인하였으며, 이러한 채점 행위가 채점 정확성과 신뢰도에 영향을 준 것으로 추론하였다. 또한 자극 회상 보고 내용에 대한 분석에서는 채점 패턴과 채점자가 언급한 언어적 요소들과는 특별한 관계가 나타나지 않았다.

데이비스(Davis, 2016)는 채점 능력의 차이로 일관성과 내용적인 변화에 주

목하여 연구를 하였다. 채점 훈련이 채점자 간 일관성을 향상시킬 수 있다는 점을 확인하였는데, 가장 정확성이 뛰어난 채점자의 사례에서 나타난 특징으로 응답에 나타난 내용에 대한 언급이 더 자주 나타나며, 점수 결정 시간도 최소 수준의 정확성을 나타낸 채점자에 비하여 더 길었다는 점을 확인하였다.

박현정 외(2017)는 문항반응이론을 바탕으로 학문 목적 한국어 말하기 평가(KoSTAP)의 채점자 특성을 분석하였는데, 가교 수험자(anchor-examinee) 설계로 이루어진 채점에 대한 분석 결과에서 전체 채점자의 채점자 모수가 ±2 표준편차 안으로 나타나 채점자 협의 중심의 워크숍형 채점자 교육을 실시한 것이 효과적이었다고 하였다. 이 연구에서는 채점자 영향으로 인한 부정적인 영향을 통제하는 방안을 제시하였는데, 채점자의 배경 정보를 고려하여 채점자 집단을 구성하면 채점 편향성이 상쇄되어 결과적으로 안정된 채점 결과를 가져올 수 있다고 보았다.

말하기 평가 연구사에서 채점에 관한 연구 검토를 통해 얻을 수 있는 시사점은 다음과 같다. 첫째, 말하기 평가에서의 채점자 영향에 관한 접근에서 상관을 통한 접근은 채점에 영향을 미치는 여러 변인이 평가 결과에 미치는 영향에 주목한 것이었다. 특히 채점자의 전공이나 경력과 같은 변인은 말하기 평가와 관련성이 높아 보이지만, 채점 수행에 대한 간접적이고 잠재적인 영향 요인이며, 따라서 개별 평가 사례와 맥락에 따라 다양한 결과가 나타날 수 있다는 점에서 기본적으로 무선적인 요인으로 보아야 한다. 둘째, 채점자 영향 중 채점자의 채점 행위에 나타난 엄격성이나 관대성, 적합도와 같은 특성에 주목한 MFRM 분석 연구에서 나타난 결과를 바탕으로 해당하는 채점자 영향의 의미를 파악할 수는 있으나, 이를 간접 변인인 채점자 배경 정보 등을 바탕으로 해석할 경우 일반화가 어렵고, 반대로 채점 수행 전에 이를 고려하여 채점자를 선발하거나 채점자 집단 구성을 하면 조절할 수 있는 변수로 간주하였을 때는 평가 윤리에 어긋나는 연구 수행이라고 보아야 할 것이다. 셋째, 발음과 같이 말하기 수행 전반에 영향을 미치는 준거에 한정하여 연구가 이루어질 경우, 채점 과정에서 그 영향이 지속된다는 점을 고려하여 결과를 해석해야 한다. 넷째, 말하기 평가의 채점 과정에 관한 연구는 언어 보고 분석을 중심으로 이루어져 왔으며, 각 연구에서 적용한 담화 분석 및 코딩 기준이 내용 중심적으로 상이하여 결과 해석에 한계가 있다. 다섯째, 말하기 평가의 채점자 훈련은 매

평가 시행을 통하여 자격 여부를 확인 받을 필요가 있으며, 스스로 채점자 영향을 성찰하고, 목표를 재설정하는 과정을 중심으로 접근이 이루어져야 한다.

3. 연구 대상 및 연구 방법

1) 연구 대상

한국어 말하기 평가의 채점 과정 연구를 위하여 5년 이상의 교육 경험을 갖고 있는 한국어 경력 교사를 연구 참여자로 선정하였다. 이들은 한국어교육을 전문으로 하는 교사 및 연구자로서 한국어교육 기관 소속이며, 대학원에서 한국어 및 한국어교육 관련 전공을 하였으므로 전문적인 한국어 교육자라고 볼 수 있다. 연구 참여자들은 소속 기관에서 실시하는 말하기 평가에 최소 10회 이상 참여하여 채점 및 출제 경험을 가지고 있었으며, 본 연구에서 실시한 말하기 평가가 중·고급 수준 학습자를 대상으로 한다는 점을 고려하여 해당 수준에 대한 교육 및 평가 경력이 있는 자로 한정하여 목적적 표집을 실시하였다. 연구에 참여한 이들은 6개 대학 기관에 소속된 13명의 한국어 교사였으며, 상세한 정보는 아래 <표 Ⅰ-1>과 같다.

<표 Ⅰ-1> 연구 참여자 정보: 채점자

ID	성별	경력	학력		외국어
		한국어 교육 (년)	전공	최종 학위	
R01	여	10	한국학	박사 수료	프랑스어, 영어
R02	여	6.8	한국어교육학	석사	영어
R03	남	15.5	현대문학	석사	영어
R04	여	7	한국어교육학	박사 수료	중국어, 영어
R05	여	18.7	한국어교육학	박사	러시아어
R06	여	16	한국어교육학	석사	영어
R07	남	11	한국어교육학	박사 수료	일본어, 영어, 러시아어
R08	여	10	한국어교육학	박사 수료	중국어
R09	여	9	한국어교육학	석사	일본어, 영어
R10	여	5	한국어교육학	석사	영어, 스페인어
R11	여	6	한국어교육학	석사	영어, 일본어
R12	여	10	한국어교육학	박사 수료	영어
R13	여	5.5	한국어교육학	석사	영어, 중국어

본 연구 진행을 위하여 설계한 말하기 시험의 응시자는 중·고급 수준의 한국어 학습자 18명이 참여하였다. 응시자의 한국어 수준을 정밀하게 통제하기 위하여 중급 상 수준에서 고급 상 수준에 해당하는 4 ~ 7급 학습자로서 1개월 이내에 해당 급을 수료 예정 또는 수료한 이들로 한정하여 모집하였다. 상세한 수험자 정보는 아래 <표 Ⅰ- 2>와 같다.

<표 Ⅰ-2> 연구 참여자 정보: 수험자

	ID	성별	국적	한국어수준
1	E01	여	카자흐스탄	고급(중)
2	E02	여	일본	고급(하)
3	E03	여	일본	고급(상)
4	E04	남	말레이시아	고급(상)
5	E05	여	일본	중급(상)
6	E06	여	이탈리아	고급(하)
7	E07	여	일본	고급(하)
8	E08	여	일본	고급(중)
9	E09	여	태국	중급(상)
10	E10	여	중국	고급(상)
11	E11	여	일본	중급(상)
12	E12	여	중국	고급(중)
13	E13	여	대만	고급(하)
14	E14	여	일본	고급(상)
15	E15	여	베트남	중급(상)
16	E16	여	중국	고급(상)
17	E17	여	대만	고급(하)
18	E18	여	일본	중급(상)

2) 연구 방법

본 연구는 한국어 말하기 평가에서 채점자들의 채점자 영향의 파악과 원인을 규명하기 위하여 양적 및 질적 자료 수집 및 분석을 병행하는 통합방법 연구를 적용하였다. 통합방법 연구(mixed method research)는 혼합 연구 방법으로도 부르는데, 연구 문제와 자료의 특성을 고려하여 더 나은 이해를 창출할 수 있는 제 3의 연구 방법론(Tashakkori & Teddlie, 2003; Creswell & Plano Clark, 2007)으로서 현상에 대한 메타 추론을 가능하게 한다는 장점을 갖고 있다(Tashakkori & Teddlie, 2008). 통합방법연구는 접근 방법에 따라 연구의 설계에서부터 자료 수집과 결과 분석 및 도출에 통합적 접근을 하는 것(Teddlie & Tashakkori, 2003)과 통합적으로 자료를 수집하고 분석하는 방법(Creswell & Plano Clark, 2007)으로 나누어 볼 수 있으며, 양적·질적 방법론의 통합을 통해 보다 정확하고 광범위한 답변을 추구하는 것을 목표로 한다(김영천, 2013: 547).

본 연구에서는 말하기 평가 결과의 원인을 제공하는 채점 과정을 설명하기

위하여 타샤코리와 테들리(Tashakkori & Teddlie, 2009)에서 제시한 통합방법 다국면 설계 모형 중 상호보완적인 접근을 취하는 병렬적 통합 설계를 따른다. 이러한 접근은 그린(Greene, 2007)에서 제시한 구성 요소 설계 유형[5] 중 양적 자료와 질적 자료의 상보성(complementarity)을 목적으로 하는 설계에 해당한다. 병렬통합모형을 선택한 이유는 질적 자료에 해당하는 채점 과정 보고 속에 양적 자료에 해당하는 평가 결과인 점수가 나타나는 자료 수집 설계를 갖고 있으며, 평가 결과에서 나타난 채점자 영향을 채점 과정을 통해 설명하는 상보성을 바탕으로 연구를 진행하기 때문이다.

(1) 채점자 영향 분석을 위한 양적 연구

말하기 평가 채점 과정의 영향을 확인하기 위하여 채점자의 평가 결과에 대한 통계 분석을 실시한다. 기술 통계 분석은 수집한 자료의 분포에 나타난 표면적인 특징에 관한 정보를 제공한다. 검사 이론(test theory)은 심리 측정 연구의 과거와 현재를 연결하는 핵심적인 이론으로서 확률 기반 추론을 바탕으로 검사 타당화 및 문항 양호도와 피험자 능력 추정 등에 관한 정보를 제공한다. 검사 이론을 구성하는 대표적인 두 이론은 고전검사이론과 문항반응이론이다.

고전검사이론(Classical Test Theory; CTT)은 실시한 검사(시험)의 결과가 얼마나 정확한지(신뢰도)에 관한 정보를 확인하려는 목적으로 발전되어 왔으며, 개인별 차이를 기준으로 전체 집단과의 상관을 통해 신뢰도를 확인하는 방법을 사용한다. CTT를 통한 접근은 분석 과정이 단순하기 때문에 명쾌한 결론을 제공한다는 장점을 갖고 있으나, 모형에서 가정하는 진점수(true score)의 개념이 모호하고, 시간에 따른 변화 가능성을 수용하지 않는다는 한계를 갖고 있다(Haertel, 2006). CTT는 관찰자가 부여한 점수인 관찰 점수가 능력 모수인 진점수와 검사 특성 모수인 오차의 합으로 이루어졌다고 가정하는데, 부여한 점수에 대하여 일회적인 해석 정보만을 제공하고, 분석 결과에 대한 비교가 불가능하며, 검사 총점을 기준으로 삼기 때문에 검사 문항의 특성을 고려하지 않는다는 한계가 있어 사용과 결과 해석을 유의해야 한다. 다시 말해서 CTT는 고

5) Greene(2007)에서는 통합 형태에 따라 구성요소 설계와 통합 설계로 나누고, 통합 설계는 다시 구성요소 범주와 통합 범주로 나누어 제시하였다.

정된 장면에 대한 측정 결과를 제공할 뿐, 자료에서 나타나고 있는 경향(pattern)은 모형화하지 않고 있어서 무엇으로 인하여 이런 결과가 나타났는지에 관한 정보를 제공하지 않는다(Mislevy, 2006)는 한계가 있다.

CTT를 통한 채점자 영향에 대한 분석이 평가 결과에 대한 종합적인 특성에 관한 정보를 제공한다면, 문항반응이론(Item Response Theory; IRT)은 통한 채점자 영향의 차이를 객관화하고 구체적으로 비교할 수 있게 한다. IRT는 개별 문항에서 나타난 수행의 확률적 특징을 고려하기 때문에 점수의 의미를 CTT보다 정확하게 추정한다는 장점이 있다. 본 연구에서는 채점자들로부터 수집한 채점 자료를 바탕으로 채점 과정에 대한 채점자 영향을 파악하기 위하여 곤란도 기반의 1모수 IRT 모형인 다국면라쉬모형(Multi-Facets Rasch Model; 이하 MFRM)을 통해 분석을 실시한다. MFRM은 곤란도 모수를 바탕으로 잠재 변수인 능력에 대한 확률적 추정을 가능하게 하는 1모수 IRT 모형 가운데 문항 변별도(item discrimination)를 1로 고정하는 라쉬 모형(Rasch model)(Rasch, 1960)을 바탕으로 한다.

MFRM 분석은 평가 맥락 요소 간의 상호작용에 대한 세밀한 정보를 제공할 수 있는데, 집단 수준 분석에 효과적인 일반화가능도(G-theory) 분석과 달리 측정을 통해 개별 대상의 잠재적인 문제들을 확인할 수 있다는 장점이 있다(Lynch & McNamara, 1998). MFRM 분석은 평가에서 어떤 국면을 고려할 것인지에 따라 그에 따른 세부 모형을 설정하여 이루어진다. 본 연구에서 수집한 양적 자료가 구성형 문항 말하기 평가의 분석적 채점을 통해 도출한 점수라는 점에서 기본적인 수험자와 채점자 국면 외에 평가 과제와 평가 준거를 분석 국면으로 설정하였으며, 결정한 점수의 척도가 0 ~ 5점으로 이루어졌으므로 다분 반응 모형을 바탕으로 한다.

$$\ln\left[\frac{P_{niljk}}{P_{niljk-1}}\right]=\theta_n-\beta_i-\delta_l-\alpha_j-\tau_k \qquad (1.1)$$

$P_{nilj k}$ = 수험자 n이 채점자 j로부터 과제 l과 채점 준거 i에 대하여 k를 받을 확률
$P_{nilj k-1}$ = 수험자 n이 채점자 j로부터 과제 l과 채점 준거 i에 대하여 $k-1$을 받을 확률
θ_n = 수험자 n의 말하기 능력
β_i = 채점 준거 i의 곤란도
δ_l = 과제 l의 곤란도
α_j = 채점자 j의 채점 경향(엄격 – 관대)
τ_k = 점수 $k-1$보다 k를 받는 것의 상대적 곤란도

수식(1.1)은 MFRM에서 다분 반응을 고려하는 모형으로서, 수험자 능력 추정에 영향을 미치는 평가 준거와 과제(문항), 채점 경향성(엄격-관대), 점수 획득 확률의 국면을 포함한다. MFRM 분석은 전용 프로그램인 Facets(Ver. 3.81.2)(Linacre, 2019)으로 이루어진다.

IRT 분석의 장점은 오차 변수의 통제를 바탕으로 측정 결과의 객관성과 비교 가능성을 확보한다는 점이다. 채점자가 부여한 점수인 총점을 중심으로 수험자와 검사 특성을 분석하는 CTT와 달리 개별 문항에서의 개별 피험자의 능력 수준을 모형화하여 다양한 검사 관련 예측 정보를 제공한다. IRT는 1900년대 초에 과학적 측정의 필요성에 관한 관련 논의가 등장한 이래로 1970-80년대에 계산 모형의 개선과 소프트웨어의 개발이 이루어지면서 선다형 문항의 분석에 본격적으로 적용되었으며, 1990년대 이후에는 심리측정에 관한 관심이 증가하면서 구성형 문항이나 혼합 유형 검사 등에 대한 다양한 적용이 가능하도록 발전해 오고 있다(Yen & Fitzpatrick, 2006).

본 연구에서는 CTT를 통한 말하기 평가 결과 분석이 채점자의 채점 경향에 대한 전반적인 특징을 나타내는 정보라면, IRT를 통한 분석은 평가 결과에 나타난 채점자별 특성을 비교 가능한 수치로 변환하여 나타낸다는 점을 주목하였다. 또한 문항반응이론 분석을 통해 채점에 개입한 채점자 영향을 상세하게 확인할 수 있으며, 이를 통해 얻을 수 있는 채점자 영향의 유형들은 이후의 채점 과정에 관한 질적 분석의 틀로서 연계한다.

(2) 채점 과정 분석을 위한 질적 연구

본 연구에서 채점자들이 말하기 평가의 채점 과정에서 어떤 사고 과정을 거

치면서 점수를 결정해 나가는지 알아보기 위하여 언어 보고 채점(verbal report rating)을 실시하고, 언어 경로 분석법(Ericsson & Simon, 1984; 1993)을 바탕으로 채점 과정 보고 담화를 정리하고, 점수 결정에 이르는 논증 구조를 분석하여 채점 과정에서 어떻게 평가 결과 도출에 영향을 주는가를 확인한다.[6] 질적 접근에서는 채점자가 보고한 내용을 채점 과정과 관련하여 인지적 차원으로 생성 및 처리하는 모든 정보들이라고 가정하며, 이는 채점 과정에 관한 실증적인 자료를 얻기 위한 최선의 접근법으로서 채택하였다.

언어적 보고법은 보고 시점에 따라서 자극이 주어지는 동안 떠오르는 생각을 보고하는 동시적 보고법(concurrent report)과 자극이 끝난 후에 기억을 떠올리며 보고하는 회상적 보고법(retrospective report)으로 나눌 수 있다. 학습자 응답 청취를 기준으로 청취 중에 떠오르는 생각을 보고하는 동시적 보고와 청취 후에 점수를 결정하는 과정에서 청취한 내용을 기억하며 떠오르는 생각을 보고하는 회상적 보고는 서로 연결되어 있지만, 듣자마자 직관적으로 반응하는 동시적 보고와 달리 회상적 보고는 들은 후에 특정한 사고 체계를 거쳐서 나타나기 때문에 이러한 인지적 처리 방식의 변화가 있다는 점을 주의해야 한다.

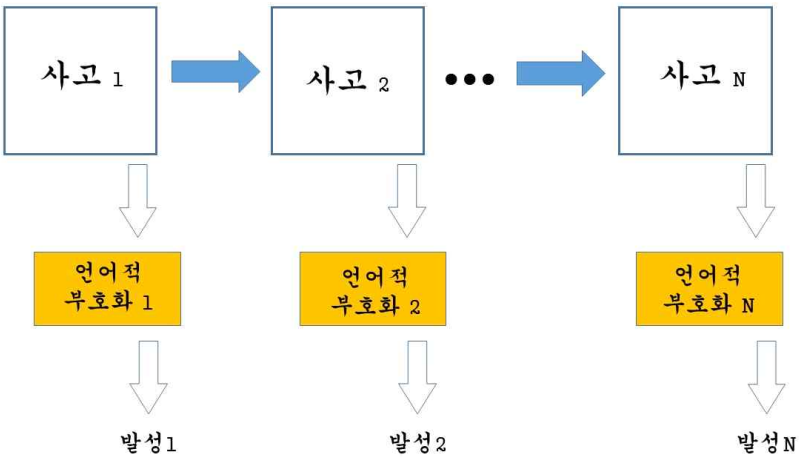
채점 과정의 언어적 보고와 사고 구술 보고의 공통점은 연구 참여자가 모든 사고의 내용을 보고한다는 것이며, 차이점은 언어적 보고 채점 과제와 채점 기준이 보고의 체계 안에서 보고가 이루어지기 때문에 기타 범주의 응답이 나타날 가능성이 매우 적지만, 일반적인 사고 구술에서는 불특정한 범주까지 보고에 나타날 수 있음을 고려해야 한다는 점이다.

개인적인 생각에 대한 보고를 통해 타당한 연구 자료를 수집하기 위해서는 자료 수집과 분석에 걸쳐 체계화된 접근 방법을 요한다. 내재된 사고에 대한 접근의 시초는 아리스토텔레스라고 볼 수 있다. 그는 자신의 정보 회상 과정을

---

6) 본 연구의 질적 자료 수집 방법인 언어 경로법(verbal protocol)은 교육 평가와 인지 심리학적 연구를 통합하는 결정적인 역할을 한 것으로서(Leighton, 2017), 보고에 나타난 문제 해결의 고등 사고 처리 과정을 주목한다. 사고의 언어화를 통해 인지적 과정을 탐구하는 대표적인 학문 분야는 인지 심리학이며, 대표적으로는 전문가와 비전문가의 인지적 차이를 알아보는 전문성(expertise) 연구가 있다. 체스 선수와 같이 전문가들의 사고 과정이 일반인 또는 보통 수준의 사람과 어떤 차이가 있는지를 다룬 전문성 연구에서는 구두 보고법을 통해 문제(체스 게임)를 해결하는 과정에서 구조에 대한 인식과 계획에 기반한 잠재적 경로 확보, 필수적인 정보의 부호화 등에서 능숙한 모습을 확인하였다(De Groot, 1978; Charness, 1981; Simpson & Gilhooly, 1997; Crowley, Naus, Stewart & Frieman, 2003).

통해 사고의 분석을 시도하였는데, 사고 자체를 가리켜 생각의 연속체이며, 이어진 생각들 사이의 전환기(transition periods)는 보고할 정보가 없다고 여겼다. 이러한 주장은 주목되지 않은 정보는 없는 것으로, 그리고 주목하여 보고된 정보만이 실체가 있는 것으로 보려는 합리주의(rationalism)적인 철학이 바탕에 있다고 할 수 있다. 그러나 이러한 입장은 17～19세기 철학자들이 제기한 고차원적인 사고 능력의 실체로 인해 반박되었다(Ericsson & Crutcher, 1991). 합리주의를 따르게 된다면 MRI나 뇌파 촬영을 통해 머릿속이나 머릿속에서 일어나는 일을 촬영하는 것이 유효하다고 할 수 있지만, 실제 심리학에서의 전문성 연구에서 나타난 바와 같이 두뇌의 신경활동 만으로는 설명할 수 없는 차원이 다른 사고가 존재하며, 이를 포착하기 위해서는 언어화를 통해 명시적으로 표현하도록 하는 것이 내적 언어에 접근할 수 있는 한 방법이라고 볼 수 있다(Watson, 1920; Ericsson & Simon, 1993). 한편 비고츠키(Vygotsky, 1987: 86)는 사적 언어(private speech)가 인지적으로 복잡하거나 요구가 많은 상황 속에서 잘 나타나며, 언어화(verbalization)를 통해 잠재된 지식에 대한 이해와 계획, 처리, 기억을 통제하는 일이 이루어진다고 하였다.



[그림 Ⅰ-1] 언어 경로 보고를 통한 사고의 언어화(Ericsson, 2006: 227)

연구 참여자의 언어화를 통해 머릿속을 들여다보는 인지적 접근이 신경 과학에서 fMRI 촬영을 통해 나타난 영상을 활용하는 것과의 차이는 영상 촬영을

통해 확인할 수 있는 것이 특정 뇌 부위에서 일어난 일정 기간의 활성화된 자극 부위라면 언어화는 구체적으로 인지한 내용과 과정을 살펴볼 수 있다는 점이다. 이러한 보고의 타당성을 확보하기 위해서는 언어적 보고를 유도하는 적절한 방법과 설명을 제공해야 하며(Ericsson & Simon, 1980), 이용한 정보를 즉각적으로 반영할 수 있도록 해야 하고, 또한 충분한 시간을 제공하여 표면적인 언어화(overt verbalization)가 가능하도록 해야 한다는 점이다(Ericsson, 2006) ([그림 Ⅰ-1] 참조). 다만 한 가지 유의해야 할 점은 사고를 보고하는 것을 어색하게 생각하거나, 개인적인 성격으로 인해 보고가 잘 이루어지지 않을 수 있다는 점이다. 이와 관련하여 에릭슨과 사이먼(Ericsson&Simon, 1993)은 간단한 교육과 적응 연습을 통해 극복이 가능한 부분이라고 보기도 하였으며, 에릭슨(Ericsson, 2006: 229)은 언어 보고의 타당도 증거는 언어화된 사고가 연역적으로 정답 생성과 결합하여 나타나는 것으로 확인할 수 있다고 하였다. 정리하면 채점자들의 말하기 평가의 인지적 과정을 살펴보기 위해서 최선의 방법으로 언어 보고법을 활용할 수 있으며(Suto & Greatorex, 2006), 보고가 이루어지는 맥락의 체계화를 통해 타당도를 높일 수 있고, 결과의 측면에서도 과정과 주장의 관계가 확보될 수 있도록 함으로써 타당성을 확보할 수 있다.

언어 보고에 대한 질적 분석에서 나타날 수 있는 또 다른 문제는 연구 결과의 비교 가능성이다. 주관적인 개인의 보고에 대한 비교는 질적으로는 거의 불가능하다고 여겨지지만, 보고의 핵심 사항에 대한 코드의 정확성(Simson & Gilhooly, 1997; Crowley, Naus, Stewart & Frieman, 2003)과 그를 바탕으로 한 논리적인 추론의 뒷받침되어야 한다(Ericsson & Kintsch, 1995; Ericsson et al., 2000; Patel, Arocha, & Kaufmann, 1994). 무엇보다 언어화된 보고의 표현 자체를 비교하는 것만으로는 구두 보고의 의미나 목적과 무관할 수 있기 때문에, 언어화 목표가 서로 일치되는 상황들을 가려내어 이를 기준으로 보고 내용의 비교가 이루어져야 할 것이다.

본 연구에서는 채점자별로 다양한 채점 과정의 비교 가능성을 확보하기 위하여 CTT 및 IRT 분석 결과를 바탕으로 문항 유형 및 채점자 영향에 따른 채점 과정의 양상을 비교한다. 이는 평가 결과가 임의적이기만 한 것이 아니라 모든 채점자가 공통적으로 채점 과정에서 고려해야 하는 기준인 채점 척도를 바탕으로 한다는 점을 고려한 것이다. 또한 말하기 평가 맥락에서 채점 과정의

역할이 점수 결정에 이르는 채점자의 추론이라는 점을 고려하여 채점 과정을 채점자가 자신이 부여한 최종적인 점수를 주장하기까지 근거와 가정을 고려하는 일종의 평가 논증으로 보고, 그에 따른 질적 분석을 실시하고자 한다. 이는 기존의 채점 과정 연구에서 이루어졌던 언어적인 측면의 분석에 비해 분석의 체계성을 확보할 수 있으며, 또한 채점 과정의 양상을 도식화하여 비교할 수 있다는 점에서 질적 자료 분석의 타당도를 높이는 방법이라고 볼 수 있다.

Ⅱ. 말하기 평가 채점 변인 연구를 위한 이론적 토대

본 장에서는 말하기 평가의 채점 변인 연구를 위한 이론적 모형 수립을 목표로 채점 과정에 관한 언어 평가학과 인지심리학적 관점을 검토한다. 그리고 합리적인 채점 결과 도출을 목적으로 이루어지는 채점 과정에서 인지적 정보 처리의 단계별 특징을 제시한다. 다음으로 말하기 평가의 채점에 관여하는 채점자의 외재적 · 내재적 영향 요인을 검토하고, 이를 바탕으로 말하기 평가 채점 의 모형과 모형 수립의 기반을 제공하는 가설을 제시한다.

1. 채점 과정에 대한 관점과 구성

1) 채점 과정에 대한 관점

채점 과정은 평가 결과를 얻기 위하여 평가자 또는 채점자가 평가 관련 요소들과 연속적으로 상호작용하는 것으로 이루어진다. 채점 과정을 파악하기 위해서는 먼저 채점이 무엇인지를 이해해야 한다. 채점의 사전적 정의는 '시험 답안의 맞고 틀림을 살피어 점수를 매김'이다. 이는 채점을 평가 맥락에서 정오나 수준에 관한 판정 결과를 추상화한 점수를 결정하는 행위라고 보는 것이다. 이러한 정의에 비추어 볼 때, 채점은 평가 결과 산출 행위라고 볼 수 있는데, 이 때 채점자가 어떻게 채점을 수행하며 인지적 처리 과정을 거치느냐에 따라 점수의 변이가 나타난다(Purpura, 2013).

한국어교육에서 이루어지는 말하기 평가 중에 대표적인 것은 교실에서 이루어지는 성취도 평가가 있다. 성취도 평가(achievement test)는 학습자가 교육을 통해 해당 영역에 대한 지식이나 기능의 습득 정도를 파악하여 교육 목표의 달성 여부를 확인하기 위하여 실시한다. 성취도 평가 중에 가장 대표적인 것은 일련의 교육 과정을 마친 후에 실시하는 중간·기말 고사와 같은 총괄 평가이다. 수업 중에 이루어지는 평가는 즉각적으로 평가 결과와 피드백을 제공할 수 있는 간단하면서 명료한 과제로 이루어지기 때문에 직관적이고 간명한 채점 수행이 가능하다. 하지만 총괄 평가는 교육과정에서 다루고 있는 주요 내용들

에 관한 과제를 구성하고, 학습자의 수행을 바탕으로 교육 목표 달성 여부를 추리해야 하기 때문에 평가 결과의 타당성을 보장할 수 있는 채점에 대한 체계적인 계획을 필요로 한다. 즉 말하기 평가의 타당성은 평가 결과 산출을 위한 평가자의 판단과 점수 결정의 타당성을 의미하며, 이 때 채점 과정을 통하여 평가자가 어떤 평가 요소와 어떻게 상호작용하여 어떤 평가 결과를 어떤 의미로 결정하는지를 파악할 수 있으므로 채점 타당성의 입증을 위해서는 이를 직접적으로 살펴볼 필요가 있다.

언어 평가에서 채점은 평가의 계획과 개발, 실행, 결과 해석 및 활용의 순환에서 실제적으로는 평가 실행 단계에서 수행이 이루어지지만 결과적으로는 평가 전반에 영향을 미친다는 점에서 평가의 영향력을 결정하는 핵심적인 활동이라고 볼 수 있다. 평가의 계획 단계에서 마련하는 측정 구인과 평가 준거는 채점 과정에서 평가 결과를 정하는 참조틀로서 활용되며, 평가 개발 단계에서 과제를 구성할 때는 채점 가능한 형태의 응답을 도출할 수 있도록 과제 수행 요건을 구성해야 한다. 평가 결과의 해석 및 활용 단계에서는 채점을 통하여 결정한 평가 결과를 평가 목적을 따라 해석하고 활용하는데, 이 때 채점의 타당성은 평가 결과의 타당도로서 결과 활용을 정당화한다. 평가 결과의 타당화 증거로서 기존의 연구에서 주목하였던 평가 결과는 채점에 영향을 미치는 여러 요인에 대한 고려가 필요하다는 유용한 정보를 제공해 왔다. 이는 산술적으로 계산된 결과인 평가 결과를 통해 특정 요소로 인한 영향을 확률적으로 예측할 수 있도록 한다는 점에서 의미가 있지만, 그 영향이 어떤 경로를 거쳐서 나타난 것인지 알 수 없다는 점에서 기계적인 해석과 형식적인 처방을 낳을 수 있다는 한계가 있다. 따라서 채점의 타당성을 본질적으로 규명하기 위해서는 평가 결과뿐만 아니라 연속적인 채점의 과정이 어떻게 이루어졌는가에 대한 접근이 함께 이루어져야 하며, 이는 채점 수행과 평가 결과의 타당성을 확보할 수 있는 방법이다.

채점 과정이 어떻게 이루어지는가에 대한 질문에 대해 언어 평가 연구에서는 수험자의 언어 능력에 대한 평가자의 연속적인 추론을 주목한다면, 인지심리학 연구에서는 불확실한 상황에서 판단을 내리기 위한 ‘추단’의 개념에 주목한다. 이러한 두 접근은 채점자의 머릿속에서 이루어지는 채점 과정에 관심을 갖고 있다는 공통점을 갖고 있지만, 채점이 합리성에 기초한 평가자의 행위인

가에 대해서는 입장의 차이가 있다. 채점 과정이 '추론'으로 이루어진다는 것은 채점자가 평가 구성 요소를 바탕으로 합리적인 의사 결정을 할 것이라는 점을 가정하고 있다면, 추단으로서의 채점 과정은 채점자의 합리성이란 이상적인 것이며, 오히려 채점 과정에서 나타나는 여러 불확실한 요인들로 인해 일정한 정도의 편향성을 띤 결정을 내린다는 것이다. 이와 관련하여 본 연구에서는 말하기 평가의 채점 과정에 대한 이론적 관점을 정립하기 위하여 추론과 추단으로서의 채점 과정이란 무엇이며, 각각 어떠한 특징을 나타내는 지에 관하여 언어평가학과 인지심리학 연구를 바탕으로 각각의 관점을 살펴보고자 한다.

(1) 추론으로서의 채점 과정

'추론(推論, inference)'의 사전적 정의는 "미루어 생각하여 논함"이며, 채점 과정과 관련하여 이를 적용하면, 채점자가 평가 결과를 도출하기 위하여 여러 평가 요소를 바탕으로 수험자의 과제 수행 및 능력 특성에 대해 특정한 해석에 도달하는 논리적인 사고 과정으로 볼 수 있다(Chapelle et al., 2008). 언어 평가에서 채점자는 평가해야 하는 항목이 무엇인지를 해석하고, 구체적으로 그것이 의미하는 바가 무엇인지를 인지적으로 추론하여 점수를 결정한다(Wolfe, 1997). 이와 같이 평가 결과의 타당성 확보를 목적으로 하는 채점 과정은 연속적인 추론 행위로 이루어진다. 평가에서 추론은 어떤 평가적 판단을 도출하는 일이며, 그 과정에서 타당한 근거를 바탕으로 확실한 답을 내리는 것을 의미한다. 채점 과정을 추론으로 보는 것은 교육 평가의 일반적인 관점으로 평가의 정확성과 타당도 요구를 가능하게 한다. 언어 평가의 타당성에 관한 구체화가 필요하다는 인식을 나타낸 연구(Cronbach, 1971; Messick, 1989; Kane, 1992)에서는 평가 타당성의 핵심으로 평가 결과의 해석과 사용의 타당성에 주목한다. 평가 결과의 해석과 사용의 타당성은 결과로 도출한 점수나 수준을 정당화할 수 있는 기제를 요구하는데, 이와 관련하여 케인(1992: 8)에서 제시한 평가 타당화에 대한 세 가지 접근은 추론으로서 이루어지는 채점 과정의 목표에 해당한다. 첫째, '시험 점수에 기반한 결론의 정확성'은 시험 점수가 산출하는 방법과 산출 과정으로부터 직접적으로 영향을 받는다는 점을 고려할 때 영향의 유무와 정도를 파악하기 위한 추론을 가리킨다. 둘째, '평가 결과 점수 사용의

적절성’은 평가에서의 추론이 시험 점수를 평가 목적에 맞게 사용할 수 있도록 적절한 해석을 유도하는 가와 관련이 있다. 셋째로 ‘평가 결과를 지지할 수 있는 자료 수집 절차 설계의 품질 확보’는 평가 결과를 산출하기까지의 과정이 타당하게 이루어졌는가에 대한 절차적 증거를 주목하는 것이다. 이러한 추론 중심의 평가 접근은 평가 타당도 증거의 속성으로 정확성과 적절성, 체계성을 주목한 것이며, 이는 평가 결과를 도출하는 채점 과정에서의 추론이 지향해야 하는 목표라고 볼 수 있다.

① 정확성 추론으로 이루어지는 채점 과정

언어 평가 연구에서 정확성을 중심으로 채점 과정을 보는 관점에서는 채점자의 신뢰도를 주목해 왔다. 이는 쓰기 평가 연구에서 채점 과정을 다루면서 평가 결과의 신뢰도에 대한 채점자의 영향을 주목하는 것으로 나타난 바 있다. 초기 연구들을 살펴보면, 쓰기 평가에서 채점자들의 일치도가 낮게 나타나는 현상과 채점자의 성격의 관계를 살펴보는 연구(Branthwaite et al., 1981)나 답안의 수준과 채점자 배경의 영향을 알아본 연구(Milaanovic et al., 1996), 과목별 쓰기 채점 신뢰도의 차이와 점수 조정(adjustment)에 관한 연구(Newton, 1996)는 채점 과정을 거치면서 채점자가 평가 맥락과 상호작용한 것이 평가 결과를 어떻게 변화시키는 가에 관한 증거로 채점자 신뢰도를 주목하였다.

채점자 신뢰도 중 채점자 간 신뢰도(inter-rater reliability)는 채점자 개인별 또는 채점자 집단별로 평가 결과인 점수의 일관성이 있는지에 관한 분석을 통해 확인할 수 있다. 채점자 간 신뢰도 분석을 통하여 특정 채점자나 특정 채점자 집단의 신뢰도가 낮은 수준으로 나타났을 경우에는 채점자 간의 일관성이 부족한 것으로 볼 수 있는데, 신뢰도 수준 저하에 기여하였을 것으로 예상한 채점자 배경 변수나 평가 맥락 변수가 평가 결과 산출에 체계적으로 관여하였음을 가정하고 있다는 점을 유념해야 한다. 일반적으로 신뢰도를 추정하는 내적 일관성 지수들은 진점수와 오차의 합으로 보는 CTT를 바탕으로 접근이 이루어지는데, 이는 수험자나 문항, 채점자 특성 등의 오차 요인을 고려하지 않는 점을 고려하여 해석이 이루어져야 한다. 본래 신뢰도(reliability)는 측정의 정확성에 관한 정보로서 평가 결과의 안정성을 가정하며, 평가 도구 품질을 나타낸

다(강승호·김양분, 2004). 신뢰도는 응답한 여러 문항의 답안에 대하여 채점자가 부여한 점수들의 분포에 나타나는 일관성 경향에 대한 분석을 통해 파악한다. 신뢰도 산출에 이용하는 평가 결과는 관찰 점수(observed score)이며, 관찰 점수를 구성하는 진점수는 같은 검사에 대한 반복 측정에서 변하지 않는다고 본다. 이는 관찰 점수의 평균값을 수험자의 능력치로 수용할 수 있는 근거가 되는데 이러한 가정은 반복 측정의 영향을 고려하지 못할 뿐만 아니라 오차 변량에 대한 고려가 획일적이라는 문제가 있다(Haertel, 2006).

② 적절성 추론으로 이루어지는 채점 과정

적절성을 중심으로 채점 과정을 바라보는 관점은 채점자가 채점 과정에서 고려한 요소들이 평가 목적에 부합하는 것들이라는 점을 가정하며, 그에 따른 점수 사용의 타당성을 나타낸다. 이와 관련하여 교육 평가 연구에서 주목하였던 구인 타당도(construct validity)는 평가를 통해 측정하고자 하는 대상인 심리적인 구인(construct)을 적절하게 측정하는 가를 확인하기 위하여 이론적·실증적 접근을 통해 수집한 증거를 통해 파악한다. 수험자의 평가 수행과 밀접한 관련이 있는 구인으로 이루어진 평가는 구인의 타당도가 높다고 할 수 있는데, 구인이 여러 개로 이루어진 평가에서는 구인들 간의 관계 설정의 어려움이 있으며, 평가 결과 해석과는 직접적인 연관을 설명할 수 없다는 한계가 있다. 채점 과정에서 채점자는 구인을 평가 과제를 통해 유도한 수험자 응답을 평가하는 핵심 요인인 동시에 측정의 대상이다. 구인은 채점 척도에서 평가 범주를 구성하기 때문에 평가 결과 산출에 대한 절대적인 영향을 주는데, 채점 방법 가운데 여러 구인을 바탕으로 이루어지는 분석적 채점에서는 채점자의 구인 이해 정도와 구인 구별 가능성이 채점 과정을 통해 체계적으로 평가 결과에 영향을 미칠 수 있다.

③ 체계성 추론으로 이루어지는 채점 과정

채점 과정의 타당성은 평가 결과를 산출하기까지의 절차가 일정한 체계를 갖추고 있음으로 인하여 판단과 점수 결정을 방해하는 무관 요소의 개입을 최

소화하고, 평가의 본질에 부합하는 평가가 이루어지도록 노력하는 가운데 확보할 수 있다. 평가 결과를 도출하는 채점 과정의 추론이 체계성을 갖추기 위해서는 채점 과정에 관여하는 평가 요소들을 구체화해야 한다. 추론을 구체화하기 위해서는 평가의 설계, 개발, 실행, 해석의 순환에서의 추론에 대한 체계적인 접근이 이루어져야 한다. 이와 관련하여 평가 타당화를 위한 논거 기반 접근의 체계를 제시한 케인(1992, 2006, 2013)이 평가 결과의 사용과 해석에 관한 추론을 지지하는 체계적인 논거(argument)를 제시한 것을 주목할 필요가 있다. 케인은 평가 설계와 개발 단계에서 개발하는 해석 논거와 평가 실행으로 확보한 평가 결과에 대한 타당도 논거로 제시하였다. 논거 기반 접근에서 논거는 어떤 주장을 뒷받침하는 논리적 근거인데, 논거를 구성하는 추론(inference)과 가정(assumption)의 타당성은 평가의 타당성을 결정한다. 논거 기반 접근은 평가 전반에 대한 타당화 접근을 한다는 점과 이론적·실증적으로 정당화한 해석 논거를 바탕으로 점수를 해석 한다는 점에서 차이가 있다(Kane, 2006). 논거 기반 접근에서 채점 과정은 평가 목적과 결과 사용에 따라서 다양한 추론 과정으로 나뉠 수 있는데, 그 첫 과정이자 기본은 관찰 점수를 얻는 채점 추론이다. 채점 추론은 수험자의 과제 수행 결과물로부터 관찰 점수를 결정하기까지의 논증 과정으로 이루어진다. 케인은 추론의 체계성을 확보하기 위하여 툴민의 논증 도식을 따라 자료, 주장, 근거, 보증, 반박, 한정 등으로 이루어진 추론 모형을 바탕으로 접근하였는데, 툴민의 논증 도식이 형식성이 높고 법적인 추론에 가까운 것들을 대상으로 한다는 점에서 모든 평가 맥락에 대한 고려를 반영하기 보다는 형식적인 측면에서의 고려에 가깝다고 볼 수 있다. 비록 이러한 접근의 한계를 해소하기 위하여 논박(rebuttal)과 같은 예외적인 것들을 평가 추론에서 고려하도록 하고 있으나, 채점 과정에서 나타날 수 있는 예외적인 상황으로 평가 맥락뿐만 아니라 수험자 맥락, 채점자 맥락과 각각의 상호작용적인 영향도 나타날 수 있다는 점에서 상호간의 관계를 규명하는 데 어려움이 따른다.

이상의 검토를 통해서 추론으로서의 채점 과정은 평가 결과 도출의 정확성과 적절성, 체계성을 추구하기 위한 목적으로 이루어진다고 정리할 수 있다. 채점 과정에서 평가 결과의 정확성을 보장하기 위해서는 관찰 점수 기반 접근의 진점수 고려로 인한 한계를 극복해야 할 것이다. 또한 평가 결과의 적절한

해석을 지향하는 채점 과정은 측정한 구인이 평가 결과에 체계적으로 영향을 줄 수 있음을 고려하여 분석 조건을 구체화하여 접근할 필요가 있다. 평가 결과 수집을 위한 체계적 접근은 채점 과정에서 이루어지는 추론을 구성하는 자료와 근거, 주장, 보증, 반박의 형식적인 측면뿐만 아니라 평가를 둘러싼 맥락적 고려에 따라 다양하게 나타나는 증거들을 수용할 수 있도록 실증적으로 접근이 이루어져야 한다.

(2) 추단으로서의 채점 과정

채점 과정에 관한 또 다른 관점은 실제적인 측면에서 채점 과정이 논리적인 추론으로 이루어지는 것이 아니라 채점 상황에 존재하는 여러 불확실한 요소에 대한 인식으로 인하여 연쇄적인 추단으로 이루어진다는 것이다. '추단(推斷, heuristics)'은 인지심리학 연구에서 휴리스틱, 어림셈법, 경험법 등으로 부르기도 하는데, 사전적 정의는 "미루어 판단한다"는 것이다. 채점 과정의 측면에서 이를 살펴보면 실제적인 채점 과정에 개입하는 채점자의 특정한 판단의 경향으로서, 판단의 근거가 비명시적인 상황에서 특정 상황 요소와의 상호작용으로 인하여 특정한 판단에 이르게 되는 것을 추단이라고 볼 수 있다. 인간 행동의 경향으로서 추단을 제시한 심리학자 트벌스키와 카너먼(1974)은 정보가 충분하지 않은 상황에서 합리적인 판단을 내리려고 하거나, 합리적으로 느껴지는 판단, 또는 합리적으로 보이기 위한 판단을 하는 사람들의 심리에 주목하였다. 여기서 판단(judgement)은 인지적 처리를 통해 이루어지는데, 트벌스키와 카너먼은 이를 자의적인 확률의 해석과 연관하여 설명한다. 대표적인 추단의 방법으로 가용성 추단과 대표성 추단, 고정과 조정 추단이 있으며, 카너먼과 프레러릭(2002)에서는 동적인 추단으로 정의적 추단과 위험 추단, 재인 추단(recognition heuristics)을 추가로 제시하였다.

추단은 사고 과정에서의 인지적인 처리를 통하여 이루어지는데, 카너먼과 프레더릭(2002)은 그에 관한 체계로서 인지적인 직관(intuitive)(체계 1)과 반성(reflective)(체계 2)으로 보는 틀(<표 Ⅱ-1> 참조)을 제시하였다.

<표 Ⅱ-1> 판단을 위한 인지 체계의 특성과 역할(Kahneman & Frederick, 2002: 51)

인지 체계	직관	반성
인지 과정 특성	자동성	통제성
	수월성	노력을 요함
	결합성	연역성
	신속 · 평행적	느림 · 순차적
	불투명성	자기 인식 기반
	숙련된 행동	규칙 적용
인지 과정 역할	정서적	중립적
	인과적 경향	통계적
	구체화, 특정적	추상적
	원형	집단

직관에 의한 판단은 판단자에게 익숙한 기존의 신념 체계를 바탕으로 입력된 정보를 자동적으로 처리하는 것이며, 이는 개별적이고 즉각적으로 일어나기 때문에 그 원인을 명시적으로 확인하기 어렵다는 특징이 있다. 반성에 의한 판단은 직관에 의한 판단이 이루어진 후에 작동하며, 특정한 규칙을 적용하여 순차적인 처리를 통해 이루어진다. 그런데 선행하는 직관에 의한 판단은 모든 가능성을 고려한다기보다는, 인식 가능한 수준 또는 인식 가능한 상황의 영향 속에 이루어지는데, 이 과정에서 가용할 수 있는 근거를 고려하지 않는 경우에는 직관에 의한 판단을 수정할 수 없게 되면서 편향(biases)이 나타날 수 있다 (Tversky & Kahneman, 1974; Kahneman & Frederick, 2002).

채점 과정에 대한 추단은 과제를 통해 얻을 수 있는 정보가 충분하지 않은 경우에도 발생할 수 있다. 이때는 매우 단순한 판단 기준을 적용하여 쉽게 인출할 수 있는 정보를 기반으로 접근이 이루어지는데, 카너만과 프레더릭(2000)은 이러한 판단 정보 부족으로 인한 편향적 결과는 피할 수 없는 결과이기 때문에 오류로 보지 않는다. 이와 반대로 정보가 부족해서가 아니라 평가하는 대상의 속성에 대한 접근이 어렵거나, 의미적이고 연합적으로 관련 짓기 쉽고, 반성적 판단의 관여가 없을 때는 대상 속성을 대체하는 일이 일어날 수도 있다.

채점 과정에서 채점자가 고려하는 여러 평가 요소들은 채점 수행 대상인 동시에 대상을 구성하는 속성으로서 존재하는데, 각 요소들에 대한 채점자의 접근성(accessibility)에 따라서 특정한 속성을 쉽게 판단하거나, 오히려 어렵게 판

단을 내릴 수 있다. 채점 대상의 속성의 영향에 관해 트벌스키와 카네만(1983)은 “일상적으로 평가한 것에 대한 지각과 이해가 추단의 역할을 한다.”고 보았는데, 이는 대상 속성에 대한 경험으로부터 구성한 채점자의 상대적 가용성(relative availability)을 바탕으로 추단이 이루어짐을 말한다.

채점 과정에서 반성적 인지 처리가 앞서 이루어진 직관적인 인지 처리의 결과를 항상 정교화하거나 의미있는 결과로 유도하는 것은 아니다. 어떤 해결하기 쉬운 문제에서는 반성적 인지 처리가 짧은 시간 안에 지나치게 단순하게 이루어지면 직관적 판정을 보증하여 강화시키는 역할만을 할 수 있다. 이는 채점 과정에서 반성적 인지가 어느 정도의 충분한 시간과 노력을 기반으로 이루어지기 때문인데, 이는 평가 상황에서 채점자가 채점 과정을 충분하게 운용할 수 있는 여건이 주어져야 한다는 점을 의미한다.

채점 과정에서 채점자가 경험을 통하여 인식한 두드러진 응답의 양상은 대표적인 응답의 특성으로 간주되며, 비슷해 보이는 사례들에 대해 확률적으로 같은 것으로 보고 직관적으로 판단하여 대표적인 응답 양상과 결합하여 처리할 가능성이 높다.

판정자가 판정을 내리는 과정에서 고려하는 여러 변수들의 크기나 속성은 추단에 영향을 미치면서 본질에 대해 내려져야 하는 판정을 왜곡하거나 변질시킬 가능성이 있다. 이와 관련하여 카너먼과 프레더릭(2002)에서는 정서적인 측면에 관한 질문지 구성 순서에 따른 추단과 판단의 본질을 왜곡시킬 수 있는 위험에 대한 추단을 사례로 제시하였는데, 판단을 내리는 과정의 구성과 고려하는 특정 요인으로 인해 판단을 심화시킬 수도 있지만, 반대로 중요한 사실을 간과하거나 배제하는 채점과 관련이 될 수 있으며, 또한 이러한 관점은 여러 외적 요소들을 고려하는 과정에서 판단의 목적에서 벗어나는 현상에 대한 이해를 제공한다.

채점 과정에서의 추단은 직관으로부터 형성한 대표성과 가용성, 접근성에 따른 추단 외에도 이전의 경험으로부터 발생하는 재인(recognition)에 의한 추단이 있다(Gigerenzer, 2007). 재인에 의한 추단은 장기 기억을 바탕으로 이루어지는데, 특정 요소에 대한 재인 가능 여부의 차이로 인하여 추단의 양상이 달라질 수 있다. 채점 과정에서 재인에 의한 추단은 채점자가 수험자의 과제 수행에 나타난 특정 요소와 관련지어 자신의 이전 경험과 신념 등을 고려하는

양상과 관련될 수 있다. 재인에 의한 추단에서 나타나는 차이는 해당 인지 요소의 인출이 잘 일어날 수 있는지의 인지적 거리에 기인하기 때문에 채점자의 해당 측면에 대한 자연스러운 평가를 가능하게 한다.

한편, 추단으로 이루어지는 채점 과정에 대한 관점은 채점 타당도를 보장하는 기존의 두 가지 핵심적인 가정에 대하여 의문을 제기한다. 첫 번째 가정은 채점을 수행하는 채점자가 충분한 자격을 갖추고 있다는 것이다. 기존의 언어 평가 연구에서 채점자의 자격과 관련하여 이루어진 접근은 채점의 일치도 향상을 목적으로 채점자의 교육 경력, 평가 경험 등을 주목하여 왔다. 그리고 연구를 통해서 나타난 사실은 경력이 많은 채점자가 그렇지 않은 집단에 비하여 더 일관된 채점을 하였다는 것이다. 이러한 결과는 충분한 경력이 채점의 일관성을 확보하는 데 긍정적으로 기여할 수 있으며, 따라서 이를 '좋은 채점자'의 자격 요건으로 여기도록 하는 근거로서 받아들여져 왔다. 그런데 추단의 관점에서 경력이 많은 채점자는 특정한 조건에서 반복적으로 채점을 경험하면서 자신만의 정보 처리 방식을 형성하였을 가능성이 높고, 이는 어떤 새로운 자극을 인식하였을 때 기존의 체계로 포섭하여 처리하려는 경향이 나타나거나 또는 중요하게 고려하지 않으려는 경향이 나타날 수 있음을 의미한다. 반대로 경력이 적은 채점자의 경우에는 채점 과정에서 인식한 정보를 처리할 수 있는 확고한 인지 체계가 마련이 되어 있지 않았기 때문에 각각의 정보를 독립적으로 처리해야 하며, 따라서 세밀한 평가가 가능하다는 장점이 있으나 채점 근거에 대한 검토가 충분하게 이루어지지 않았을 때에는 지나치게 관대하거나 엄격한 채점을 할 가능성이 높을 것이다.[7] 두 번째 채점 타당도에 대한 가정은 채점에 적용하는 준거(rating criterion)와 척도(scale)가 타당하다는 것이다. 성취도 평가의 경우, 일반적으로 교육과정을 근거로 하여 평가 준거와 척도를 구성한다. 평가 준거는 평가 구인의 수준 또는 점수에 따라 기술한 내용으로 이루어지는데, 평가를 담당하는 교사는 평가 준거와 척도의 타당성을 확보하기 위

7) 채점자의 오랜 교육 경력이나 많은 평가의 경험이 채점을 더 정확하게 한다는 결과를 채점 과정을 통한 채점자의 체계적인 영향으로 볼 때, 이는 특정한 경험이 많은 채점자가 가진 편의성을 쫓는 경향이 채점자 집단에서 일반화된 경향으로 나타나는 가중치 편향 때문이라고도 해석할 수 있다. 가중치 편향(weighting biases)은 너무 많은 단서들을 고려하여 판단을 해야 하거나, 너무 적은 가중치를 두는 것과 관련하여 나타나는 현상인데, 이는 채점 과정에서 채점자가 특정한 수행 증거나 평가 준거 등을 고려하지 않는 경향을 갖는 것이 확률적으로 인식하는 결과에 대한 근거가 될 수 있다.

해 해당 영역 전문가가 타당화한 자료를 사용하거나 또는 이론적인 검토를 통해 내용 타당도를 검증해야 한다. 이러한 형식적인 접근 외에도 채점을 수행하는 교사들이 평가 준거와 척도를 사용하면서 발생한 문제를 공유하고, 의견을 반영하여 수정하는 실용적인 접근을 취할 수도 있다. 그런데 이와 같이 타당도를 확보한 평가 준거와 척도를 마련하는 것이 평가 준거와 척도 사용의 타당성을 보장하는 데에는 한계가 있다. 채점 과정에서 채점자가 평가 준거와 척도를 어떻게 이해하고 있으며, 어떻게 해석하여 적용할 것인지는 개별 채점자의 주관적인 영역으로 맡겨져 있으므로 그로 인한 채점의 모호성이 증가할 수밖에 없다.

이상에서 살펴본 채점 과정에 관한 추론과 추단의 관점은 채점 과정에 대한 상대적인 관점이라기보다 채점 과정에서 채점자가 획득한 정보를 어떻게 처리할 것인지에 대한 결정과 관련이 있다고 볼 수 있다. 채점자가 수집한 정보와 판단이 합리적이라고 생각할 때는 추론적인 접근이 가능할 수 있으나, 불확실한 부분이 많거나 잊고 있던 새로운 고려 사항이 떠오르는 등 복잡한 상황에서는 주관적인 추단에 의한 판단을 내릴 가능성이 있다.

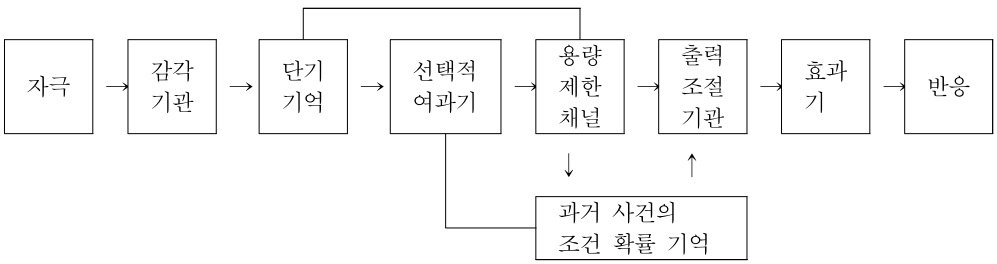
2) 말하기 평가 채점 과정의 구성

채점 과정은 측정하고자 하는 대상의 특징과 상호작용 요소 및 환경에 따라 다양한 형태로 이루어지지만, 기본적으로는 채점 대상에 대한 지각 및 인식, 그리고 점수 결정을 위한 판정 및 최종 점수 결정의 절차로 이루어진다. 말하기 평가는 채점자의 청각 기관을 통한 음성 지각을 바탕으로 한다는 물리적 특징을 갖고 있다. 쓰기 평가 연구에서도 채점 과정에 대한 인식의 기본적인 틀이 판정 대상 자료에 대한 채점자의 지각과 세부 정보에 대한 판별과 최종 점수 결정의 구조로 이루어지는 것으로 보고 있다(Lumley, 2000; 박종임, 2014). 그런데 학습자의 과제 수행 결과를 보면서 채점을 할 수 있는 쓰기 평가와 달리 말하기 평가는 채점자가 청취한 발화를 기억하면서 채점을 해야 한다는 차이가 있다.[8] 즉 학습자의 과제 수행 내용을 채점 과정 전반에 걸쳐 필

8) 준직접식 평가의 경우에는 녹음 자료를 들으면서 채점을 하기 때문에 반복적인 청취가 가능하지만, 실제적인 말하기 상황의 일반적인 특성상 반복 청취의 기회가 없다는 점과 가시적인 문자 언어에 대한 평가가 아니라 음성 언어에 대한 평가라는 점에서 재청취를

요할 때마다 직접적으로 확인을 할 수 있는 것과 기억 체계에서 해당 내용을 인출하여 떠올리는 차이로 인하여 정보의 정확성에 대한 차이가 있으며, 청취 정보 인출 과정에서 해당 정보에 대한 인상을 함께 고려할 수 있기 때문에 말하기 평가의 채점 과정에서는 이러한 모호한 정보나 인상적인 정보를 어떻게 처리하느냐가 판정과 점수 결정에 영향을 줄 수 있다. 따라서 말하기 평가 채점 과정의 독특성은 청각 기관을 통해 입력된 청취 정보에 대한 인지적 처리로 이루어진다는 점에 있다.

사람의 머릿속에서 입력된 정보(자극)를 어떻게 처리하는지(반응)에 관한 연구는 인지심리학에서의 정보처리이론(information processing theory)에서 접근이 이루어져 왔다. 정보처리이론에서는 자극과 반응 사이에 마음이라는 정보처리체계가 관여하며, 정보처리체계를 구성하는 정보처리구조와 정보처리과정의 작용 가운데 반응이 이루어지는 것으로 본다(이정모, 2001: 208). 이러한 인지적 처리 과정과 관련하여 브로드벤트(Broadbent, 1958: 299)는 특정 자극은 남기고 그 밖의 것들은 폐기하는 선택적 주의 개념을 제안한 바 있다.



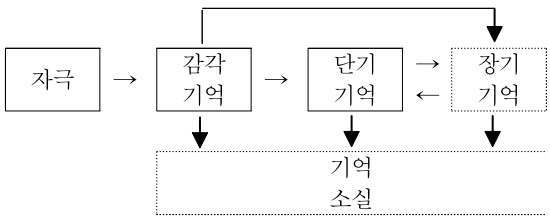
[그림 Ⅱ-1] 브로드벤트(1958: 299)의 선택적 주의 처리 과정

[그림 Ⅱ-1]에서 감각 기관으로 입력된 자극은 아주 짧은 시간 동안 감각을 저장하는 단기 기억을 거쳐서 자극의 유형에 따라 선택되어 가용량이 제한적인 중앙처리기를 통해 기억으로서 저장하거나 출력 조절 기관을 거쳐 반응으로 나타난다(이정모, 2001: 184). 브로드벤트의 인지적 정보의 처리에 대한 관점은 자극으로부터 반응이 나타난다고 보는 행동주의적 관점에서 간과하고 있는 인지적 처리 과정을 주목하여 제시하였다는 의의가 있으나, 정보에 대한 선

가 이루어지지 않는다.

택적 여과가 어떻게, 어느 정도로 이루어지는 가에 대한 규정이 모호하고, 주목하지 않은 자극도 약한 강도일지라도 기존의 인지 체계와 결부되어 저장될 수 있다는 점을 간과하고 있다.

앳킨슨과 시프린(Atkinson & Shiffrin, 1968: 17)은 인지적인 정보 처리의 과정과 관련하여 기억 체계 구조(the structrure of memory system)를 제시하였는데, 이는 정보를 저장하는 감각기억(sensory register)과 단기기억(short-term store), 장기기억(long-term store)으로 이루어져있다. 감각기억은 외부로부터 들어온 정보(자극)가 감각 기관을 통해 수용하면서 그것이 무엇인지에 대한 식별이 이루어지는 것과 관련이 있다. 감각기억은 순간적으로 이루어지지만, 본격적인 기억 체계에서의 처리 여부를 결정하기 때문에 결정점(decision-point, Brownell, 2007)이라고도 한다. 작업기억이라고도 부르는 단기기억은 기억 체계로 수용한 자극을 0～30초 동안 저장하고, 정보에 대한 통합과 변형이 이루어진다. 장기기억은 수분에서 평생 동안 유지되는 기억으로서, 과거 경험에 대한 기억이나 학습한 것과 관련이 있다([그림 Ⅱ-2] 참조).



[그림 Ⅱ-2] 앳킨슨과 시프린(1968)의 기억 체계 구조

이러한 인지 과정에 대한 정보처리적 관점은 인간을 인지능동적인 정보 처리자로 보며, 주어진 자극에 대한 지각과 사고의 과정이 복잡한 자연체계라는 점을 가정하고 있다(이정모, 2001: 214).

인지적인 정보 처리로 이루어지는 말하기 평가의 채점 과정에서 채점자는 청취를 통하여 수험자 발화를 지각하고, 지각한 청취 정보는 기억 체계와 연계하기 위하여 단기기억을 통해 통합 및 변형하며, 장기기억에 내재된 채점자의 채점 관련 전문성을 반영하여 채점을 수행한다. 채점자는 채점 과정 가운데 평

가를 구성하는 외적인 맥락 요소와 채점자의 내적 요소의 영향을 받으며, 이러한 요소들은 채점자의 인지적인 정보 처리의 수준 및 양상을 결정한다. 특히 언어 평가에서 수험자나 과제 이외에 채점자의 발화 수준 판정 및 점수 결정에 영향을 미치는 요인으로는 채점 척도(rating scale)에 대한 이해와 적용이 있다. 채점자는 채점 척도를 구성하는 평가 준거 및 기준을 학습자 발화를 연계시킴으로서 최종적인 점수 결정에 이를 수 있다. 인지적 정보 처리 과정으로서의 말하기 평가 채점 과정은 학습자 수행 정보에 대한 청취 및 판정, 점수 결정의 단계로 이루어진다.

(1) 수행 정보 청취 및 판정

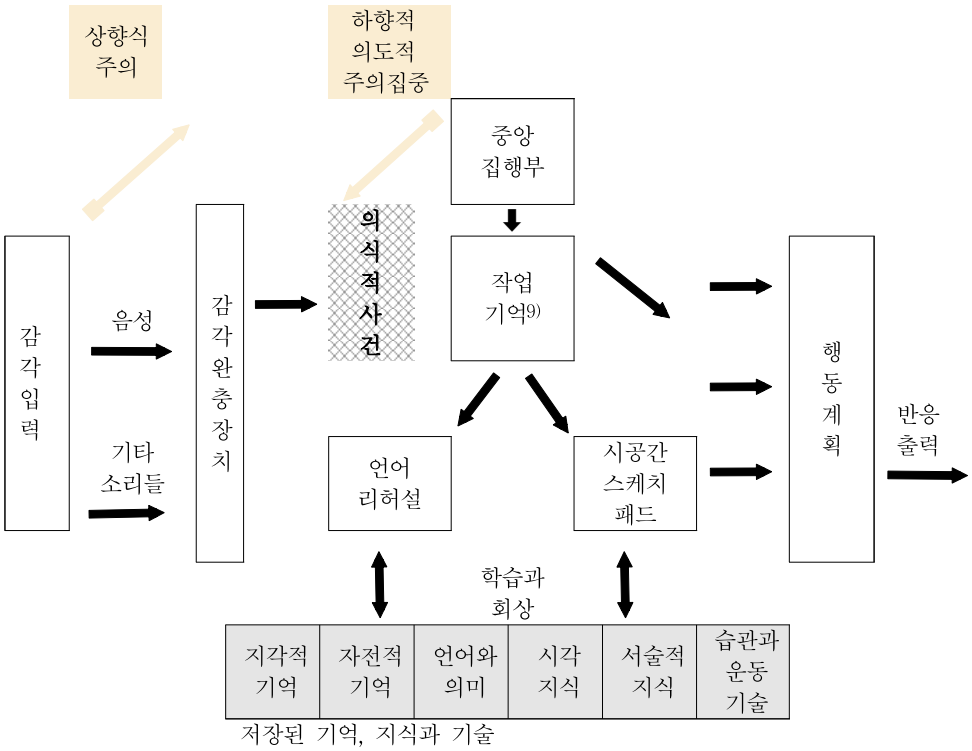
말하기 평가에서 채점은 수험자가 과제 수행을 통해 구성한 응답을 채점자가 청취하는 것으로 이루어진다. 청취는 의도적으로 듣고자 하는 소리를 찾아서 인식하는 수동적 청취와 자연스럽게 노출된 상황에서 들리는 소리를 인식하는 자동적인 청취로 나눌 수 있다. 채점자는 채점 과정에서 청취를 통해 수험자의 발화를 평가를 위한 정보 단위로 분류하거나 연합하는 등의 인지적 처리를 하는데, 이 때 선택된 정보는 이해와 해석, 평가의 중첩적인 처리를 거친다(Brownell, 2007: 21). 채점자의 머릿속에 입력된 정보는 선험적인 기억 체계와 연계하기 위한 변형과 통합의 과정을 거치게 되고, 그에 대한 결과로서 잠정적인 판정을 내리게 된다. 말하기 평가 채점 과정에서 응답 청취의 단계는 학습자의 발화에 대한 지각 및 인식, 그리고 인식한 발화에 대한 해석과 평가, 그리고 이후의 발화 정보 처리를 위한 발화 기억의 과정으로 이루어진다.

① 발화 지각 및 인식

말하기 평가의 채점 과정에서 채점자는 수험자의 응답을 들으면서 그 소리가 무엇이며, 어떤 특징을 갖고 있는지를 확인하게 된다. 앞서 살펴보았던 정보처리이론에서 감각 기관을 통해 수용한 자극은 모든 발화 가운데 선별된 것이며, 담화로 구성한 발화에서는 어떠한 단위로 끊어 이를 인식하는지의 차이가 인식의 차이를 가져올 수 있다(Anderson, 2010). 이러한 발화 지각과 인식에

는 수험자가 어떻게 말을 하였는지도 영향을 주지만, 채점자가 수험자의 발화에 나타난 특징을 어떤 범주로 처리하는지가 가장 직접적인 영향을 미친다고 볼 수 있다. 채점자가 수험자의 응답에 나타난 특징을 무엇으로 지각하고 어떻게 인식하는가를 이해하기 위하여 청각 정보 처리에 관한 뇌의 작동을 살펴볼 수 있다.

인간 뇌의 작동은 수천 억 개의 뉴런을 통해 전기적 작용 과정에서 신경전달 물질을 주고받는 것으로 이루어진다. 이러한 신경 전달 물질의 교환에서 뇌 안에 존재하는 다양한 수용체들과 경로들, 회로, 네트워크, 효과기들이 다양한 신경 활동을 나타낸다(Baars & Gage, 2007: 60-61).



[그림 Ⅱ-3] 청각 처리 과정의 기능적 구조 체계(Baars & Gage, 2007: 180)

9) Baars & Gage(2007)에는 'working storage'로 되어 있는 용어이며, 번역서에는 '작동 저장'으로 번역되어 있는데, 본 연구에서는 인지심리학적 접근을 따라 해당 용어를 '작업기억'으로 수정하였다.

[그림 Ⅱ-3]은 청각적인 자극을 수용하는 과정에서 뇌에서 일어나는 처리 과정을 보여준다. 소리는 물리적인 실체로서 진동에 의해 발생한 음파가 청각 기관으로 전달된 것을 말한다. 어떤 환경 속에서 청각 기관에 도달한 소리는 의도적 혹은 자연적으로 듣거나 들리면서 청취자의 하향식 또는 상향식 주의 집중 과정을 통하여 감각완충장치(sensory buffer)에 도달하여 일시적으로 저장되었다가 선택적 주의 집중 과정을 거치면서 이전의 기억 및 경험, 다른 감각과의 상호작용 속에서 부호화 및 저장, 파지(retention)하는 처리가 이루어진다 (Baars & Gage, 2007: 180).

말하기 평가 과정에서 채점자들은 시험장에서 의식적인 사건으로서 학습자 또는 수험자의 발화를 들으며 평가 실천 과정을 진행한다.[10] 그리고 청취 과정에서 선택적으로 주목한 발화의 요소들에 대한 작동 기억을 거치게 되는데, 이때 발화 정보를 저장하고 처리하는 과정에서 어떤 정보들은 삭제되거나 또는 장기적으로 보관될 수 있다. 또한 장기적인 처리와 관련하여 이전에 형성된 장기 기억의 인출이 일어날 수 있다. 청취로부터 발생하는 반응은 작동기억에서 일어난 일과 관련하여 나타날 수 있으며, 학습자의 발화, 수행 특징, 수행 특징에 대한 느낌, 수행과 관련하여 인출한 장·단기 기억 정보, 기타 채점자의 정서적 표현 등이 나타날 수 있을 것이다.

청취자는 청취 과정에서는 들리는 소리에 집중하게 되는데, 소리 집중을 유발하는 특성으로는 반복, 변화, 새로움, 강도가 있다(Brownell, 2007: 103). 반복적으로 들리거나, 먼저 들었던 것과 다른 변화가 나타나거나, 뜻밖의 행동이 나타나거나, 그 강도가 강할 경우에 집중이 이루어지는데, 말하기 평가 과정에 적용해 보면 집중을 통하여 학습자 발화를 꼼꼼히 잘 듣게 한다는 점에서는 도움이 될 수 있으나, 평가와 무관하거나 관련성이 적은 요소에 집중을 할 경우에는 평가에 부정적인 영향을 줄 수 있으므로 집중한 것의 가치를 잘 따져 볼 필요가 있을 것이다. 브라우넬(2007)의 청취에 관한 접근에서 특이한 점은 인지적인 청취에서 기억을 설명하면서 즉각 기억(immediate memory)[11]이 청취

10) 청취 과정에서 학습자의 얼굴을 마주하거나 평가 문항이나 채점 기준표 등의 시험 구성물을 보는 시각적인 입력도 주어질 수 있다. 본 연구에서는 말하기 평가의 채점자로서의 본질이 청취자(listener)에 더 가깝다고 보고, 청취 과정으로 한정하여 논의를 전개한다.

11) 브라우넬(2007)의 즉각 기억은 인지신경과학에서의 감각 여과 장치(sensory buffers)를

의 결정점으로서 “25%의 회상 가능한 내용을 결정한다.”고 하였다. 말하기 평가 상황에서 브라우넬이 언급한 즉각 기억은 발화가 물리적으로 전달되어 감지되는 것을 의미하며, 학습자가 발화하고 있다는 인식을 갖게 하는 것이라고 볼 수 있다. 발화 감지로부터 평가 실천 과정이 시작된다고 보았을 때, 평가 환경과 학습자 음성의 음향적 특징 등이 발화 감지를 방해하지 않도록 일관성을 확보할 필요가 있다.

② 발화 해석

청취 단계에서 인식된 소리에 대한 구체적인 인지적인 대응은 정보처리이론에서 제시한 작업기억을 통해 이루어지는데, 이는 중앙집행부(executive)로부터 조절된 청각적 입력을 수초에서 수분 동안 해석하고 평가하는 영역이다. 작업기억에서 어떤 정보들은 주변적인 간섭에 취약하여 즉각적 또는 매우 짧은 시간만 유지되다가 사라지기도 하며 반대로 며칠에서 몇 주 이상의 기간 동안 지속되는 영구화 현상이 일어나기도 한다.[12] 작업기억을 거친 후에는 다시 중앙집행부의 조절과 행동 계획을 따라 관련된 출력을 하게 된다.

먼저 채점자는 청각 기관을 통하여 지각한 발화가 무엇인지를 이해하는 해석을 한다. 수험자 발화에 대한 이해가 잘 이루어지기 위해서는 발화 해석을 위한 채점자의 선험적인 해석 기제가 갖추어져 있어야 한다. 효과적인 발화 해석에 대하여 브라우넬(2007: 245)은 청자의 의미 조율에 영향을 미치는 요소를 고려하는 사회적 민감성(sensitivity)을 강조하였는데, 이는 해당 맥락과 관련하여 고려할 필요가 있는 사항에 대해 인지할 수 있는 감각을 가리킨다. 말하기 평가의 채점 과정의 측면에서 보면 청취자인 교사 또는 시험관이 발화자인 학습자 또는 수험자의 과제 수행 발화에 나타난 어떤 특징을 지각하여 해석할 때 채점자의 발화에 대한 민감성의 영향이 있다는 점을 고려할 필요가 있다.

가리키고 있는 것이다. 즉각 기억을 감각 저장소로 보는 관점도 있지만, 단기 기억과 같은 경우로 보는 입장과 독립적인 기억으로 분류해서 보는 관점이 있다.

12) 장기기억과 관련하여 흥미로운 점은 장기기억이 뇌에 남겨지는 과정에서 활성화에 관여했던 영역에 저장이 된다는 사실(Baars, Gage, 2007: 53)이며, 이러한 점에 비추어 볼 때, 해당 감각 기관의 장기기억 활동이 반복 및 누적되었을 때 향후 인출될 수 있는 관련된 장기기억도 증가할 수 있다는 가설을 세울 수 있다.

채점자가 반복적으로 노출된 특정 음성적 자질이나 발화 특성에 대하여 더 잘 지각할 수 있다는 점과 반대로 낯선 특징에 대해서는 해석적인 한계가 나타날 수 있다. 채점자에게 익숙한 특징으로 인식된 정보에 대한 채점과 관련하여 선행 연구에서는 음성적 자질과 관련해서는 관대한 채점 경향이 나타난 바 있다. 발화 이해에 관한 채점자의 민감성으로 인하여 나타난 이러한 편향성은 채점 과정에서 체계적인 영향을 나타낼 수 있으므로, 이러한 특징을 고려한 채점 설계와 평가 결과 이해가 이루어져야 한다.

③ 발화 평가

발화 평가는 청취한 발화로부터 발화 목적과 목표를 바탕으로 발화에 대해 태도와 신념을 형성하는 것으로서, 청취한 발화를 다른 기관에서 처리하기 위하여 작업기억에서 유지하는 과정에서 이루어진다. 브라우넬(2007: 321)은 청취에서 일어나는 평가적인 발화 처리에 대하여 '비판적 듣기'와 '청자 설득 행위'라고 보았는데, 이는 결정에 초점을 기울이는 심판적인 평가가 아니라, 청취 과정에서 객관성을 유지하면서 개방적인 태도로 발화자의 주요 생각과 지지하는 논증의 타당성을 판단하는 것을 말한다. 이는 발화 이해와 마찬가지로 청취자가 청취한 정보를 들리는 대로 인식하는 것으로는 타당한 발화 평가가 이루어질 수 없으며, 따라서 자신의 청취 태도를 재고하면서 발화에 대한 판단을 내려야 함을 의미한다. 발화에 대한 채점자의 평가적 태도는 발화 특징에 대한 타당성을 결정하기 위하여 증거와 기반이 충분한가를 기준으로 접근할 수 있다. 발화 평가의 근거에 대하여 브라우넬(2007: 336)은 통계 자료, 인용, 증언, 비교, 이야기를 발화 타당성 평가의 증거로 보았으며, 이는 청취자들에게 발화에 대한 믿음의 강도를 결정하는 중요한 요소가 될 수 있다.

말하기 평가 상황에서 채점자는 수험자의 발화에서 자신의 선험적인 인지 체계와의 비교 활동이 일어날 수 있는 항목에 대하여 평가적 태도를 취할 수 있다. 이는 수험자 응답에 대한 즉각적인 반응으로서 나타나기 때문에 메모를 해 두거나, 기억 체계와 연계하여 이후의 점수 결정 과정에서 해당 항목에서의 양상을 다른 요소와 함께 고려하기 위하여 의미 중심으로 변형하여 재평가할 수 있다.

④ 발화 기억

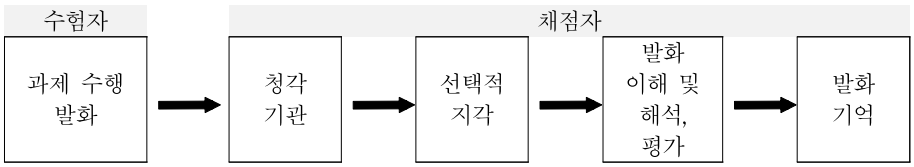
다음으로 발화 기억은 청취자의 머릿속에서 이해 및 해석, 평가적 처리를 거친 발화가 기억 체계와 연결되어 잠재적 가능성을 가진 것으로 처리되는 것을 가리킨다(이정모, 2009: 446). 감지된 발화는 작업기억을 거치면서 기억 체계와 상호작용하게 되는데, 이 때 메시지에 대한 이해와 메시지 상의 구조성 여부에 따라서 기억 수준의 차이가 발생할 수 있다. 예를 들어 청취한 메시지를 잘 이해하지 못하였을 경우에는 연결 자원의 부족으로 기억 체계에 저장하기 어려우며, 또한 구조가 잘 나타나지 않는 메시지도 기억 체계에 통합하기 보다는 "이전의 믿음에 따라 변화시킬" 가능성이 크다(Brownell, 2007). 발화에 대한 청취자의 기억 가운데 장기기억으로 처리된 것은 영구적으로 보관되는 것으로 알려져 있지만, 인출을 가능하게 하는 조건이 갖추어지지 못하였을 경우에는 기억할 수 없다는 특성이 있다. 다양한 종류의 기억이 있다는 것은 신경생리학에서 이루어졌던 여러 기억에 관한 연구들을 통해 드러난 바 있는데, 외적인 형태에 따라서는 명시적 기억(explicit memory)과 암묵적 기억(implicit memory)로 나눌 수 있다. 명시적 기억은 의식적인 자각과 서술이 가능(예: 일화적 기억)한 반면에 암묵적 기억은 의식적인 회상이 없으면서 행동 수행에 영향을 미치지만, 설명하기는 어려운 기억을 가리킨다(예: 절차적 기억)(이정모, 2009: 448). 말하기 평가 채점 과정의 측면에서 채점자가 수험자의 발화를 기억 체계에서 어떻게 저장하였는지를 알아보기 위해서는 기억을 인출하여 확인해야 하는데, 다양한 정보들 가운데 특정 정보는 암묵적 기억이기 때문에 인출이 가능하나 설명이 어렵고, 일화적인 기억은 직관적으로 설명할 수 있지만, 쉽게 변화 및 망각될 수 있기 때문에 반복적인 인출 경험이 필요하다. 따라서 채점자는 이러한 청취 과정에서 작동하는 기억 체계의 영향으로 인하여 수험자의 발화에 나타난 특징(자극, 정보)을 점수 결정 과정에서 회상할 때 정보의 특성에 따라 기억 여부 및 고려 방법에서 차이를 나타낼 수 있다.

말하기 평가의 채점 과정에서 이루어지는 채점자의 응답 청취는 앞서 제시한 인지적인 청취 과정을 기반으로 하면서, 점수 결정이라는 목표 하에 이루어진다. 주의해야 할 점은 채점자마다 개인적인 차이를 갖고 있는 것처럼, 응답 청취 과정에서도 인지적인 특성의 영향으로 인하여 차이가 나타날 수 있다는

점이다. 어떤 청취자이냐에 따라서 청취 과정에서 고려하는 강한 요소와 약한 요소, 혹은 요소 간의 특성에 대한 강한 관계와 약한 관계 인식 등의 차이가 나타날 수 있으며, 청취 과정 구성 요소에 대한 메타적인 인식 수준과 양상에도 차이가 나타날 수 있다. 이와 관련하여 브라우넬(2007: 22)은 청취자의 차이를 바라보는 관점으로 자기 점검(monitoring), 피드백(feedback)을 제시하였는데, 자기 점검을 잘 하는 청취자는 정확하게 상황을 평가할 수 있는 능력을 갖고 있으며, 타인에 대한 민감도가 높고, 상호작용 중에 자신이 반응을 수정할 수 있는 경우를 말한다. 반면에 낮은 수준의 자기 점검 상태에서는 상대방보다는 자신에게 집중하며, 의사소통 방식이 고정되어 있는 특징 등이 나타난다고 하였다. 청취에 대한 반응인 피드백은 구성적으로 제공될 때 청취력이 향상된다고 하였는데, 피드백이 특성에 초점을 두면서 평가적이거나 개괄적이지 않고, 행동에 초점을 두면서 설명적이고 구체적으로 이루어진다고 하였다. 말하기 평가의 채점 과정에서는 수험자의 응답으로부터 청취자의 반응이 나타나게 되는데, 대면 상황에서 이루어지는 직접식 평가에서는 채점자가 속으로 생각을 하면서 청취 정보의 인지적 처리를 보조하기 위하여 메모나 표시를 할 수 있으며, 비 면대면 상황에서 이루어지는 준직접식 평가에서는 메모나 표시 이외에도 채점자가 응답을 청취하면서 떠오르는 개인적인 사고를 발화 청취를 방해하지 않을 만한 의식적 혹은 무의식적 반응으로 발화하면서 접근할 수 있다. 이 때 채점자가 발화한 내용은 채점 과정에서 평가 결과를 도출하기 위하여 채점자의 사고를 유지 및 강화, 촉진시킬 수 있다는 점에서 채점 과정의 양상을 확인할 수 있는 중요한 정보가 된다.

또 다른 효과적 청취자의 특징은 청취 과정을 메시지 생성과 전달로만 보는 것이 아니라 발화자와 의미를 함께 구성해 가는 관계적 과정으로 보는 것이다(Howell, 1982). 이러한 관점에서는 청취 상황 변화에 따라 자신의 행동을 수정하는 경향이 나타나는 것이 특징인데, 이는 말하기 평가의 채점 과정에서 채점자의 채점 과정 접근의 차이를 가져오는 원인이 될 수 있다. 앞선 경험을 토대로 학습자의 발화 수준을 대략적으로 미리 결정하고 자신이 발견한 단서가 나타나지 않았을 경우에는 채점 경향이나 결과를 수정하지 않고 앞선 결과를 그대로 수용하는 채점자가 있는 반면에 수험자 응답에 나타난 다양한 요소에 대하여 각각을 구분하여 개별화된 접근을 취하는 경향이 나타날 수 있다.

말하기 평가의 채점 과정에서 수험자의 발화에 대한 채점자의 지각과 인식의 과정을 요약하면 [그림 Ⅱ-4]와 같다. 채점자의 청각 기관에 입력된 수험자의 응답은 선택적으로 지각되며, 이해와 해석, 평가 과정을 거쳐 기억 체계와 연계되어 청취 발화에 대한 판단으로서의 반응(언어)을 생성하는 것으로 이루어진다.



[그림 Ⅱ-4] 말하기 평가에서 채점자의 인지적 청취 과정

말하기 평가 채점 과정에서 채점자의 응답 청취 이후에는 청취를 통해 확보한 수험자 발화에 대하여 채점 척도를 바탕으로 점수를 결정하여 부여하는 일이 이루어진다. 일반적으로 채점자가 점수를 결정하는 일은 임의적으로 이루어지는 것이 아니라, 사전에 마련한 채점 척도나 루브릭을 기준으로 하여 점수를 부여한다. 채점자는 점수 결정 과정에서 수험자의 과제와 점수를 연계하기 위하여 다양한 내재적·외재적 자원들을 활용하기 때문에 점수 결정 과정의 차이가 발생한다.

(2) 수행 점수 결정

채점 과정에서 수행 점수의 결정은 채점 척도[13](rating scale)를 근거로 하여 이루어진다. 채점자는 채점 과정에서 청취한 정보와 채점 척도의 연관성을 파악하고, 여러 판단 근거들을 복합적으로 고려하여 척도 상에서 수험자의 수행

13) '채점 척도'는 수험자의 과제 수행에 대하여 점수를 부여하기 위하여 평가하는 준거와 그에 따른 수준을 의미하는 척도, 그리고 각각의 구체적인 내용 기술로 구성한다. 기존의 연구에서 채점 척도와 혼용하고 있는 용어로는 채점 기준표, 루브릭이 있으며, 채점 기준표는 '기준'의 의미가 채점 척도의 내용을 포괄하기에 추상적이라는 점과 루브릭은 리커트 척도와 같이 상대적인 정도의 차이를 나타내기 위한 것이라는 점에서 본 연구에서는 채점 척도라고 사용하였다.

을 가장 잘 나타내고 있다고 판단한 내용의 기술이 있는 점수를 선택한다. 점수 결정 과정은 채점 척도에 대한 해석과 적용, 그리고 결론으로서 점수 선택을 정당화하는 것으로 이루어진다.

① 채점 척도의 해석

채점자는 수험자의 과제 수행에 대하여 점수를 부여하기 위하여 주어진 채점 척도를 읽고, 구성하는 요소들의 특징을 파악함으로써 채점 척도를 해석해야 한다. 채점 척도는 평가 문항의 특성을 고려하여 작성하는데, 평가 준거를 구성하는 구인(준거)과 평가 수준을 의미하는 척도로 구성한다. 채점 척도의 유형은 채점 척도를 구성하는 주체와 구성하는 방법에 따라서 나눌 수 있다. 일반적으로 교실 평가에서 채점 척도 개발의 주체는 평가 개발자인 교사이며, 대규모 평가에서는 전문 평가 개발자에 의하여 이루어질 수 있다.

채점 척도를 구성하는 방법은 직관 및 경험 기반, 이론 기반, 자료 기반, 실증 기반으로 나눌 수 있다(Fulcher, 2003: 89-90). 직관 및 경험 기반은 채점 척도 개발에 관한 가장 일반적인 접근이라고 볼 수 있으며, 오랜 역사를 갖고 있는 방법이다. 직관 및 경험 기반 채점 척도는 경험이 충분한 교사 또는 전문 평가자가 기존의 채점 척도, 관련 교수 요목, 요구 분석, 이전 수행 결과물 등에 관한 자료를 바탕으로 작성한다. 직관 및 경험 기반 척도는 실제적인 측면을 고려하기 위하여 평가 전문가(교사)가 참여하거나 실제 수행 자료를 바탕으로 채점 척도를 구성한다는 점에서 실증 기반 채점 척도 개발과 유사해 보이지만, 전문가 평정과 협의를 중심으로 하는 실증 기반 접근과 달리 시험관들이 주관적으로 보고한 결과를 바탕으로 전문가가 정리하여 구성한다는 차이가 있다. 이론 기반 채점 척도는 평가자 기반 척도라고도 하는데, 평가를 통해 측정하는 구인에 대한 정의를 바탕으로 나타날 것으로 예상하는 수행의 수준에 대한 선행 연구 검토를 바탕으로 구성한다. 이론 기반 척도는 척도별 수행 특징에 관한 설명의 구조가 단순하고, 포괄적인 성격을 갖고 있어서 효율적인 채점 수행을 가능하게 한다는 장점이 있다. 자료 기반 척도는 이론 기반 척도에 실제적인 응답 결과에 나타난 수험자 특성 정보를 반영하여 개발한다. 채점 척도의 내용을 작성하는 과정에서는 평가를 통해 수험자로부터 어떤 과제 수행을

유도하였으며, 그것을 적확하게 평가할 수 있는 채점 방법을 선정해야 한다. 예를 들어 문법적 지식을 평가하는 과제와 그림을 보고 문제를 파악하여 설명하는 과제는 같은 채점 척도를 적용하여 평가할 수 없는 과제이기 때문에 과제에 맞는 채점 척도를 작성하여 사용해야 한다. 채점 방법은 평가 결과를 통해 확인하고자 하는 정보가 무엇인지에 따라 선택할 수 있는데, 종합적 척도의 경우에는 수험자들의 순위를 파악하기 위한 목적으로 사용할 수 있다면, 분석적 척도는 교육 목표 달성과 관련하여 준거별로 상세한 평가 결과를 얻기 위하여 사용한다.

채점 척도는 평가의 목적에 부합하는 유형을 선택하는 것이 중요하며, 말하기 평가의 경우 어떤 말하기 능력을 평가하고자 하는 것인지가 채점 척도의 내용에 나타날 수 있도록 해야 한다. 채점자는 채점 척도에 낯선 내용이나 모호한 부분이 없는지를 살펴보고, 채점 척도 사용이 점수 결정 과정을 방해하지 않도록 사전에 사용 연습을 통해 점검을 할 필요가 있다.

② 채점 척도의 적용

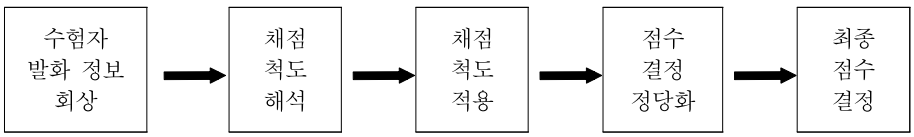
채점 척도를 이해하고 있는 채점자는 점수 결정을 위하여 수험자의 과제 수행에 대하여 채점 척도를 적용해야 한다. 말하기 평가의 경우에는 채점 척도와 과제 수행을 연결하기 위해서 청취 과정에서 주목한 내용들을 채점 척도를 따라 적용해야 하는데, 채점 척도에서 수험자의 과제 수행에 부합하는 수준과 그에 합당한 기술이 있는지를 확인해야 한다. 이 과정은 최종적인 결정 이전에 예비적인 수준의 채점을 하는 것이기 때문에 과제 수행 수준에 대한 판정을 평가 결과로서의 점수로 치환한다는 점에서 점수화(scoring)라고 부르기도 한다(Lumley, 2000: 291). 채점자는 채점 척도에 대한 이해를 수험자의 응답과 연결하기 위하여서 자신의 기억 체계와 과제 수행, 그리고 채점 척도를 연계하는 복잡한 정보 처리에 대하여 전략적인 접근을 취하게 되는데, 럼리(Lumley, 2000: 291)는 쓰기 평가의 채점 과정에서 채점자가 '중재(arbitrating), 균형(balancing), 비교(comparing), 보상(compensating), 해석(interpreting), 거절(rejecting)의 전략'을 사용하여 점수를 부여한다고 하였다. 이러한 전략들은 점수 부여의 기준이 되는 채점 척도의 사용 방법과 관련이 있는데, 럼리의 접근

외에도 써스톤(Thurstone)의 ‘대응비교법(paired comparison method)’을 적용하거나(Pollitt & Crisp, 2004) ‘맞춤 활동(matching)’을 하는 방법(Suto & Greatorex, 2008), 채점 척도를 자신만의 관점에서 재해석하여 사용한다는 입장(Quellmalz & Burry, 1983) 등도 채점 척도를 기준으로 한다는 점에서 공통점을 갖고 있다. 여기서 대응비교법이나 맞춤 활동은 형식성이 높은 접근을 취하기 때문에 높은 채점자 간 신뢰도를 기대할 수 있는 접근이라면 기술된 채점 척도의 내용을 재해석하는 것은 유연한 관점으로서 평가 결과 해석의 다양성을 수용하는 관점으로 볼 수 있다. 그 밖에 채점 척도 사용에 대한 다른 접근으로는 채점자가 채점 척도와 수험자의 과제 수행을 연결시키는 과정에 채점자가 갖고 있는 사전에 형성된 ‘내재된 인식(inbuilt perception)’(Brown, 1995: 13)의 작용이 있으며, 이는 채점 척도의 개선이나 일시적인 채점자 훈련으로 변화시킬 수 없는 것으로 보는 관점이 있다. 채점자의 내재된 인식은 채점 과정에 체계적으로 관여하면서 채점 수행과 결과에 영향을 나타낸다는 점에서 채점자 영향(rater effect)이라고 부른다. 채점자 영향은 채점자의 내재된 인식을 형성한 채점 수행과 관련하여 갖고 있는 이전 경험으로부터의 영향이라고 볼 수 있으며, 이는 교육 경험, 평가 경험, 상황에 대한 해석과 관련이 있다. 채점자가 가진 해당 분야의 교육 경험은 수험자의 수준을 가늠하거나 양상을 예측하는 기준으로서 작용한다. 채점자의 평가 경험은 교육문화적인 현상으로서 평가의 상황에서 수험자의 입장과 과제 수행 결과물에 대한 태도를 형성한다. 상황에 대한 해석은 특정 과제에서의 응답을 검토하면서 평가 기준이나 채점 지침 외에 언어적·반언어적 요소를 고려하여 점수화하는 것을 가리킨다.

말하기 평가의 채점 과정에서 채점자가 채점 척도를 적용하면서 이루어지는 복합적인 평가 요소에 대한 고려는 채점자의 점수 결정 과정을 불확실하게 하는 원인이 된다. 여기에 채점자가 갖고 있는 특정한 채점의 경향은 점수 선택의 정당성을 위협하는 원인이 된다. 이러한 상황은 매우 탁월하거나 매우 미흡하게 과제 수행을 한 발화같이 채점 수행 특성이 명료한 경우에 나타나기 보다는 주로 보통, 중간, 평범한 수준과 같이 상대적인 수준 고려 이외의 접근이 모호한 경우에 나타날 수 있다. 채점자는 이러한 채점 척도 적용에 따른 불확실함을 해소하기 위하여 채점 과정에서 고려하였던 요소에 대한 재고려 또는 새로운 증거에 대한 고려 등을 통하여 자신의 점수 결정을 정당화한다.

③ 점수 결정의 정당화

채점자가 점수를 결정하는 과정은 합리적으로 점수 선정이 이루어졌음을 확인하는 정당화를 통해 완성된다. 점수를 확정하는 단계에서 예비적으로 선택한 점수를 유지하거나 또는 다른 판정 근거에 대한 고려로 인하여 다른 점수로 교체하는 일이 일어날 수 있다. 평가 결과로서 부여하는 최종적인 점수는 과제의 요건이 단순하고, 채점 척도가 명확한 경우에는 직관적인 판정을 통하여 결정할 수도 있지만, 대부분의 수행 평가에서 측정하는 잠재적인 능력의 속성인 구인은 직접적인 관찰이 어렵거나 불가능하기 때문에 과제 수행 상에 나타난 여러 특징과 채점 척도에 대한 재검토 과정을 통하여 타당성을 확인하는 일이 필요하다(Suto & Greatorex, 2008). 이러한 정당화 과정은 채점 과정에서 채점자에게 선택한 점수가 불확실한 신념을 갖게 하는 특정 요소를 채점의 근거로 보는 것이 타당한지 아니면 부분적인 근거나 무시 가능한 것으로 볼 것인지, 이와 관련하여 고려할 수 있는 새로운 정보를 탐색할 것인지 등에 대한 판단을 요구한다. 이 단계는 채점 과정의 부족한 점을 보완하기 위하여 머릿속에서 이루어졌던 채점 과정을 확인하는 반성적인 절차인데, 카네만과 프레더릭(2002)이 추단의 두 번째 체계로 제시한 반성적 인지의 작동과 관련이 있다. 채점자의 채점 척도 이해 수준과 수험자의 응답에 대한 인지 수준의 다양성으로 인하여 발생하는 인지적인 초점의 차이는 반성적 인지의 작동 양상을 다양하게 하는 원인이다(Vaughan, 1991).



[그림 Ⅱ-5] 말하기 평가에서 채점자의 점수 결정 과정

[그림 Ⅱ-5]와 같이 말하기 평가에서 채점자는 채점 과정을 통해 점수를 결정하며, 이러한 과정은 채점 척도에 대한 이해와 수험자의 과제 수행에 나타난 특징을 연계하는 것과 채점자의 내재적인 평가적 인식의 영향, 점수 확정을 위한 근거의 초점화 방향과 수준에 따라서 다양한 양상으로 나타날 수 있다. 채

점자가 수험자의 과제 수행 발화를 청취하는 과정과 점수를 결정하는 과정에는 채점자의 평가 내·외적인 특성으로 인한 영향이 개입할 수 있으며, 이는 채점 과정을 복잡하게 하는 원인으로 작용할 수 있다. 점수를 결정하기까지 채점자가 고려하는 평가 요소들은 점수 결정의 근거로서 작용하며, 평가 결과를 통해 나타내고자 하는 채점자의 의도를 반영한 것이다.

2. 말하기 평가 채점 과정의 영향 요인

채점자의 채점 수행에 관여하여 채점 과정을 변화시키는 영향 요인은 채점자의 외부로부터 기인한 요인과 내부적인 요인으로 나누어 살펴볼 수 있다. 외적 요인은 평가 맥락을 구성하는 수험자와 과제, 채점 척도, 환경 등과 관련이 있다. 내재적 영향 요인은 채점자의 심리적·인지적 측면에서 기인하여 채점 과정의 결과에 영향을 미친다.

1) 채점자 외적 요인

(1) 수험자의 수준 및 배경

수험자의 사전적인 의미는 '시험을 치르는 사람'이며, 이는 시험을 경험하면서 필요한 정보나 자격을 얻기 위한 목적을 갖고 있는 학습자를 가리킨다. 말하기 평가에서 수험자는 평가의 목적과 실행에 있어서 최우선적으로 고려해야 하는 대상이며, 평가 결과에 영향을 미치는 가장 큰 변인으로 여겨진다(Matsugu, 2013). 전통적인 평가에서 수험자 변인을 통제하기 위하여 직관적으로 평가 가능한 능력만을 대상으로 하는 접근을 취하였으나(Spolsky, 1978), 심리측정학의 발달과 평가 타당도의 중요성이 강조되면서 평가 목적 세밀화 및 체계화 토대 위에 수험자 수준과 필요를 반영(통제)하여 평가하는 접근이 이루어졌다. 이러한 접근은 "수험자에게 너무 많은 자유를 주지 말라"는 휴즈(2003: 54)의 언급처럼 평가자인 교사 또는 평가 개발자 중심의 평가관을 바탕으로 한다. 수험자의 수준과 필요를 반영한 평가는 수험자가 역량을 충분히 발휘할 수 있는 평가 환경이 조성되지 못하였을 때는 능력을 과소평가하거나 구

인과 무관한 평가가 이루어질 수 있다는 문제가 있다. 수험자 영향을 고려하기 위해서는 평가 과정에서 수험자의 수준과 평가 목적을 바탕으로 체계적 접근을 위한 형식화를 통해 채점 기준이나 과제 특성, 평가 방법 등에 관한 지침을 따라 개별적으로 판정하는 접근이 필요하다. 채점자는 점수 결정 과정에서 수험자의 개인적인 수준이나 특성을 주목할 수도 있는데, 개인적 특성이 최종 수행 수준 결정에 대한 가정으로 관여할 수 있으나, 관찰한 수행이 아니라 선험적 신념을 바탕으로 한다는 점에서 유의해야 한다.

(2) 말하기 과제의 형식과 내용

말하기 평가의 과제는 수험자와 채점자의 수행을 연결하는 매개이며, 평가를 통해 판정하여 점수를 부여할 수 있는 언어 사례를 도출하는 도구이다(Fulcher, 2003). 과제는 평가 문항으로 구성하여 수험자에게 제시되는데, 채점자는 각 문항의 과제에서 다루는 내용이 무엇이며, 어떤 형식으로 제시되었는지를 파악하여 수험자 응답의 특성을 이해할 수 있도록 해야 한다.

말하기 평가 과제의 형식은 수험자가 말해야 하는 담화 상황에 따라서 유형을 구분할 수 있다. 이는 말하기 평가에서 과제의 역할이 특정한 상황과 목적, 성취 목표와 관련하여 목적 지향성을 갖고 있어 수험자가 무엇을 말해야 하는지를 안내하는 것(Pica et al., 1993; Bachman & Palmer, 1996)이라는 점과 관련이 있다. 가장 대표적인 말하기 과제의 형식은 대면 상황의 여부에 따라 독화와 대화로 구분할 수 있다. 독화(monologue)형 과제는 수험자가 주어진 화제에 대하여 주어진 시간 동안 답을 하는 것을 가리키는데, 간단한 문장을 읽는 것에서부터 복잡한 문제에 대한 의견을 말하는 발표의 형식까지 다양한 수준에서 이루어질 수 있다. 독화형 말하기 과제는 참고 자료의 사용 여부에 따라서 독립적 과제(integrated task)와 통합형 과제(integrated task)로 나누기도 하며, 독립적 과제는 주로 수험자의 경험이나 생각을 말하는 형식이라면 통합적 과제는 읽기·듣기 자료를 활용하여 주어진 과제에 답을 하는 형식으로 평가가 이루어진다. 독화형 과제는 과제 수행을 위해 면담자가 필요하지 않기 때문에 준직접식 말하기 평가에서 사용하기에 적합하다. 독화형 과제에 대한 채점은 과제를 통해 측정하고자 하는 것이 무엇이냐에 따라서 차이가 있을 수 있다. 독

립적 과제에서는 언어적인 측면에 초점을 둔다면, 통합적 과제에서는 복잡한 담화를 구성하는 것과 관련된 복합적인 능력을 측정하기 때문에 채점의 복잡성이 높다고 볼 수 있다. 다음으로 대화형(interview) 말하기 과제는 수험자가 면담자와 이야기를 주고받는 형식으로 이루어지는 것을 가리키는데, 실제적인 상호작용을 통해 평가가 이루어진다는 점에서 평가의 진정성을 확보할 수 있다는 장점이 있다. 대화형 과제는 면담자가 있어야 하기 때문에 직접식 평가로 이루어지는데, 면담자로 인한 영향이 개입할 수 있고, 면접관이 채점자의 역할을 동시에 해야 하는 경우에는 인상 중심의 평가에 그칠 수밖에 없다는 어려움이 있다. 따라서 기초적인 의사소통 능력을 평가하는 경우에는 상호작용 중심의 대화형 과제를 선택할 수 있으나, 수험자가 준비한 응답을 보고 또는 발표의 형식으로 말하는 상황에서는 준직접식 평가로 구성한 독화형 과제를 사용하는 것이 복합적이고 잠재적인 능력을 측정하는 데에 더 적합하다고 볼 수 있다.

한편, 말하기 평가 과제는 과제의 지향성, 상호작용성, 목적지향성, 관계성과 이를 수반하는 주제 및 상황을 고려하여 작성하는데, 상대적으로 수험자에게 부담이 적은 과제는 요구하는 응답이 간단하고 단순한 것이라면, 복잡하고 구성적인 응답을 요구하는 과제는 인지적인 부담이 크기 때문에 문항의 곤란도가 높아진다는 점을 유념해야 한다(Brown & Yule, 1983). 이와 같이 과제의 형식과 내용에 따른 곤란도의 차이는 채점자의 채점 과정에도 영향을 미치며, 단순한 구성의 과제에서는 직관적인 접근이 가능하다면, 복잡한 구성의 과제에서는 여러 과제 구성 요소에 대한 고려가 복합적으로 이루어져야 하므로 반성적인 접근이 필요하다. 따라서 채점자는 과제에 따라서 측정하는 준거를 어떻게 판단해야 할지 생각하고, 수험자의 응답이 과제의 특성과 관련하여 나타난 것이라는 점을 유념해야 한다.

(3) 채점 기준의 구성

채점 기준은 말하기 과제 수행을 측정 및 평가하기 위하여 기술하며, 이론적인 바탕에서 경험적인 요소와 실행적인 측면을 고려하여 구성한다. 채점 기준에 따라 측정하는 말하기 능력에 관한 설명이 달라지기 때문에 평가의 목적에

부합하도록 기술하는 것이 중요하다. 채점 기준은 평가하는 말하기 능력을 구성하는 잠재적인 특성과 변별적인 수준을 구성하는 척도를 바탕으로 그에 부합하는 내용에 관한 기술로 이루어진다.

채점 방법에 따라서 총체적 채점의 채점 기준은 통합된 말하기 능력을 대상으로 하기 때문에 단일 척도와 단일 특성으로 구성하며, 분석적 채점 기준은 말하기 능력 구성 개념을 분리하고, 각 척도 및 특성을 구별하여 기술한다. 분석적 채점 기준은 총체적 채점 기준에 비하여 작업량이 많다는 부담이 있지만, 인상 중심의 비체계적 평가를 방지할 수 있고, 수험자에게 유용한 여러 능력 특성 정보를 제공할 수 있어서 전반적인 효과를 알아보기 위해서는 총체적 채점 기준을 사용하고(Brown, 2000; Brown, 2006), 평가 준거별 정보를 얻기 위해서는 분석적 채점 기준을 사용해야 한다. 수험자의 수행을 채점 기준과 관련지어 해석하는 과정에서 채점자는 문서화된 채점 기준을 바탕으로 채점자 내에서 정신적으로 재구성한 기준을 적용한다(Bejar, 2006). 이상주의적 관점에서는 채점자는 채점 척도와 수험자 발화의 인접쌍 찾기 활동을 수행하여 채점을 하는 것으로 보고 있으나(Pollitt, 2004), 실제로는 채점자들은 자신만의 채점자 영향을 갖고 있으며, 평가 과제와 특성 등에 따라 다양하게 나타난다(Lazaraton, 1996; Brown & Hill, 1998).

2) 채점자 내적 요인

말하기 평가에서 채점자의 내재적 영향 요인은 채점 과정에서 평가 결과 도출에 영향을 미치는 채점자의 인지적·정의적 수준과 특성을 가리킨다. 채점자(rater/scorer/marker)는 채점 행위를 하는 사람이며, 사전에서 채점은 '시험 답안의 맞고 틀림을 살피어 점수를 매김'으로 답안에 대한 판정을 통하여 추상화된 점수를 부여하는 책임을 의미한다. 채점자는 평가 목적에 따라 평가 결과로 도출한 점수의 타당성을 보장하는 전문가로서, 교실 평가에서는 교사가, 대규모 평가에서는 전문 평가자가 채점자라고 볼 수 있다. 말하기 평가에서 채점자는 측정 대상인 구인을 바탕으로 과제 수행 발화에 대한 수준 판정과 점수 결정을 하면서 청취 내용 판별과 채점 척도 이해, 채점 자원 관리에 관한 능력을 발휘해야 한다.

(1) 청취 정보 판별의 정확성

말하기 평가 채점 과정에서 채점자는 수험자의 과제 수행 응답에 대한 청취를 통하여 수험자의 과제 수행 수준 판정과 점수 결정을 위한 직접적인 증거들을 수집한다. 채점자가 수험자의 응답의 특성을 적절하게 채점하기 위해서는 청취한 응답의 각 요소가 무엇이며, 어떤 상태인지에 관한 청취 정보를 정확하게 판별할 수 있는 능력이 요구된다.

말하기 평가의 수험자는 음성으로 응답을 산출하며, 채점자는 청각 기관을 이용하여 응답을 청취한다. 인간이 소리를 듣는 과정은 일종의 신경 전달 과정이라고 볼 수 있는데, 이 때 작동 기억의 영향으로 청취한 정보가 축소 또는 확장, 지속, 변형되는 일이 일어난다(Baars & Gage, 2007). 채점자가 수험자 발화 판별은 물리적인 파형인 소리가 청각 기관을 통하여 전기적인 신호로 변환되고, 뇌에서 청각 신경을 관장하는 부위를 통하여 입력된 신호를 처리하는 것으로 이루어진다. 이 때 수험자 발화의 특성과 이를 처리하는 채점자의 인지적 특성이 상호작용하는 수준에 따라 청취 정보 판별 결과가 달라진다.

말하기 평가에서 채점자는 수험자의 응답을 듣고 이해할 수 있어야 하며, 이는 청취의 책임이 청자에게 있다는 브라우넬(2007)의 관점과 통한다. 말하기 평가 상황에서 수험자는 평가 맥락을 고려하여 발화를 한 것이며, 채점자는 수험자가 평가 받는 상황에서 발화했다는 것을 고려하여 평가의 목표에 부합하는 가를 기준으로 이를 판별하여야 한다.

채점 과정에서 채점자가 수용한 수험자 응답의 음성 정보는 음성적 지각과 선택적 인식, 발화의 이해·해석·평가, 기억 체계로 처리가 이루어지는데, 채점자는 인지 활동을 통해 청취를 통해 발견한 신정보와 기존에 갖고 있던 구정보를 연계하여 어떤 의미를 구성하거나 모형화하여 점수 결정 활동으로 연계시킨다(Mislevy, 2006). 채점자가 청취한 정보를 기존의 체계와 연계시킨다는 것은 해당 정보를 장기기억 체계로 통합한다는 것을 의미하며, 채점 과정에서 이를 유효한 점수 결정의 단서로 인출하여 사용할 수 있다. 그러나 청취한 정보 가운데 주목하지 않았거나, 잘못 판별한 경우에는 채점의 정확성을 위협하는 원인으로 작용할 수 있다. 연속적으로 주목해야 하는 정보가 발생하여 단기

기억 체계의 용량을 초과하는 경우도 판별을 어렵게 하는 원인이 되는데, 이러한 인지적 한계를 보완하기 위하여 필요한 경우에는 채점 과정에서 발화 정보를 기록하는 활동을 병행할 수 있다.

(2) 채점 척도의 이해와 적용 능력

채점 척도[14]는 채점 과정에서 점수를 결정하는 절대적인 기준으로서 존재하며, 채점자는 평가의 목적과 과제의 요구를 바탕으로 채점 척도의 내용을 파악하고, 수준 판정과 점수 결정의 근거를 확보해야 한다. 채점 척도에 대한 채점자의 이해는 채점자의 특성에 따라 다양할 수 있으며, 일반적으로는 전문적인 교육을 받은 채점자의 경우에는 채점 척도에 대한 공통의 이론적 기초를 갖고 있는 것으로 본다. 그러나 채점 척도의 형식적인 면의 기저에 있는 평가 결과 도출의 원리를 채점자가 이해하지 못하고 있다면 결과적으로는 같은 점수라도 그 의미는 다른 것이다. 채점자는 채점 과정에서 사용하는 채점 척도의 유형이 무엇이며, 그에 따라 평가 결과 산출 과정에 영향이 있음을 고려하여 채점을 하여야 한다.

채점자가 갖고 있는 채점 척도에 대한 지식은 채점 척도 상의 정보를 쉽게 해석하고 적용하는데 기여한다. 채점 과정에서 사용하는 채점 척도는 채점 방법을 결정하는데 '총체적 채점 척도(holistic rating scale)'와 '분석적 채점 척도(analytic rating scale)'가 대표적이다. 총체적 채점 척도는 수험자의 과제 수행에 대한 종합적인 판정 결과를 제공하기 위한 것이며, 포괄적인 접근을 취한다는 점에서 전체적 척도(global scale)라고도 한다. 종합적 척도는 하나의 점수나 등급을 선택하도록 구성하기 때문에 직관적인 접근이 가능하며, 채점의 효율적이며 결론이 명료하다는 장점이 있다. 그러나 복합적인 요소로 이루어진 측정

14) '채점 척도(rating scale)'는 평가 결과 산출을 위한 평가 기준으로 평가 척도와 평가 준거에 따라 해당하는 수준에 대한 기술로 구성한다. 채점 척도는 평가하는 언어 능력에 대한 이론적 구성 개념인 구인을 기초로 하며, 각각의 구인들은 평가하는 언어 능력을 구성하는 것으로 본다. 예를 들어 한국어 말하기 능력을 평가한다고 하였을 때, 평가하는 구인으로 '발음', '문법', '담화 구성'을 정하였다면, 채점 척도는 각각의 구인에 대한 수준별 특성을 기술한 채점 기준으로 구성한다. 기존의 연구에서 채점 척도는 채점 기준, 채점 기준표, 루브릭으로 명명되기도 하는데, 본 연구에서는 수준 판정의 기준과 점수 결정을 위한 척도로 이루어졌다는 점을 강조하기 위하여 '채점 척도'라고 부른다.

대상을 지나치게 단순화하여 접근하면서 평가 결과의 의미를 훼손시킬 수 있고, 과제 수행의 여러 결과들을 종합하였을 때는 일반화 가능성이 낮다는 한계가 있다(Bachman & Palmer, 2010: 340). 분석적 채점 척도는 수험자의 응답을 여러 요소로 분리하여 측정하기 위하여 사용한다. 이러한 분석적 접근은 수험자 수행에 대한 상세한 정보를 제공한다는 장점이 있는데, 상대적으로 종합적 접근에 비하여 효율이 낮고, 채점자 신뢰도 확보를 위한 물리적·경제적 투입이 이루어져야 한다는 약점이 있다(Lane & Stone, 2006).

총체적 채점 척도를 사용하는 경우 채점자는 점수 결정 과정에서 여러 가지 응답의 특징들을 바탕으로 하나의 점수 또는 등급을 부여하기 위하여 다양한 요소들을 합성(合成)하는데, 이 과정에서 어떤 부족한 수행 특징을 다른 영역에서 보상하는 접근이 이루어진다. 일반적으로 총점 개념을 사용하는 평가에서 보상적인 접근이 나타나며, 말하기 평가에서는 '말하기 능력'이나 '의사소통 능력'과 같이 종합적인 개념의 구인을 평가할 때 이러한 접근이 나타난다. 분석적 채점 척도는 어떤 능력 개념으로부터 이론적으로 구성한 여러 평가 구인들이 서로 독립적인 성격을 갖고 있음을 가정한다. 그래서 채점을 수행할 때 각 채점 영역들이 결합 관계라는 점을 고려하여 구인을 중복 평가하거나 임의적인 심리적 가중치를 고려하는 현상이 일어나지 않도록 주의해야 한다.

(3) 점수 결정의 일관성

말하기 평가의 채점자는 채점 과정에서 여러 가지 복합적인 평가 요소들을 고려하면서 최종적인 수행 점수나 등급을 결정해야 한다. 이와 관련하여 채점자는 복합적인 채점 관련 요소를 고려하면서 점수를 결정해야 하기 때문에 평가 요소들을 중재하고, 채점이 일관되게 이루어지도록 하여 타당한 결론을 도출할 수 있도록 전략적인 접근을 취해야 한다.[15] 이는 복잡한 채점의 과정에서 채점자가 자신의 인지적인 부담을 조절하고, 결론으로 나아가기 위한 방책으로

15) 채점 과정에서 채점자가 문제를 해결하기 위하여 사용하는 전략(strategy)은 다양한 상황에서 나타날 수 있다. 채점 전략과 관련하여 럼리(2001)의 관점은 점수 결정을 위한 것이 중심이라면, 커밍 외(Cumming et al., 2001)에서는 채점 척도 이해를 위한 해석 전략(interpretation strategies)과 등급이나 점수 형성을 위한 판단 전략(judgement strategies)이 채점자가 문제 해결에 사용된다고 보았다.

서 개별 채점자의 채점 특성으로도 볼 수 있다. 채점 전략과 관련하여 럼리(2001: 240)의 연구에서는 채점자가 갈등 상황에서 사용하는 전략으로 '쥐어짜기(squeezing), 다듬기(shaping), 규정하기(defining), 중재하기(arbitrating), 비교하기(comparing), 거절하기(rejecting)'의 사용을 확인하였다. 그런데 이러한 채점에 관한 전략적인 접근은 해당 채점자가 어떤 조건과 어떤 맥락을 어떻게 고려하느냐에 따라서 사용 유무에 차이가 있을 수 있으며, 어떤 전략이 자신에게 더 효과적인지를 가늠하기 어렵다. 전략적 접근의 차이를 비교하기 위해서는 이를 사용함으로서 얻은 채점자의 효용이 무엇인지를 평가 결과와의 관계 속에서 해석하는 것이 필요하다. 평가 결과와 무관한 점수 결정 전략의 사용은 채점자 편의에 기반한 것일 수 있으며, 이는 편향된 평가 결과로 나타날 수 있다.

말하기 평가의 채점 과정은 평가 결과를 도출하는 직접적인 원인을 제공한다. 기존의 수행 평가에서의 채점 과정 연구에서 통합 방법 연구로 접근한 사례들에서는 채점 과정을 영향 요인으로 보면서 채점 과정을 분석하면서 보고 내용의 빈도로 경향을 파악하거나, 양적 분석과 연계하지 않거나 보충적인 차원에서 채점 과정에 대한 사후 면담 내용을 제시하였다. 이러한 접근은 채점 과정과 평가 결과에 나타난 채점자 영향과 양상을 파악하는데 도움이 되지만, 인과적 관계를 가정하지 않기 때문에 무엇으로부터 나타난 결과인지를 설명하는데 한계가 있다.

본 연구에서는 채점 과정이 평가 결과의 원인으로서 어떻게 작용하였는가를 파악하기 위하여 두 가지 조건을 고려하여 연구 자료를 수집하였다. 첫째, 말하기 평가 채점 과정의 특성을 파악하기 위하여 채점자의 채점 특성을 비교·대조할 수 있는 척도화된 정보가 수집되어야 한다. 말하기 평가의 점수는 특정 채점자가 특정 수험자의 응답에 대하여 특정 평가 맥락을 고려하여 특정 상황에서 부여한 것이다. 이러한 점수를 비교하기 위해서는 같은 척도 상에 위치할 수 있도록 점수를 조정해야 하며, 문항반응이론 가운데 MFRM은 평가 결과를 문항 곤란도를 바탕으로 척도화(logit 분포)하여 여러 평가 국면들을 비교 가능하게 하는 방법이다. 이와 관련하여 본 연구의 Ⅲ장에서는 MFRM 분석을 통해 한국어 말하기 평가의 평가 결과에 나타난 채점자 영향을 확인할 것이다.

둘째, 말하기 평가 과정의 자료 수집 및 분석의 타당성이 보장되어야 한다. 자료 수집 및 분석의 타당성은 과정 타당도(process validity)를 나타내며, 본 연

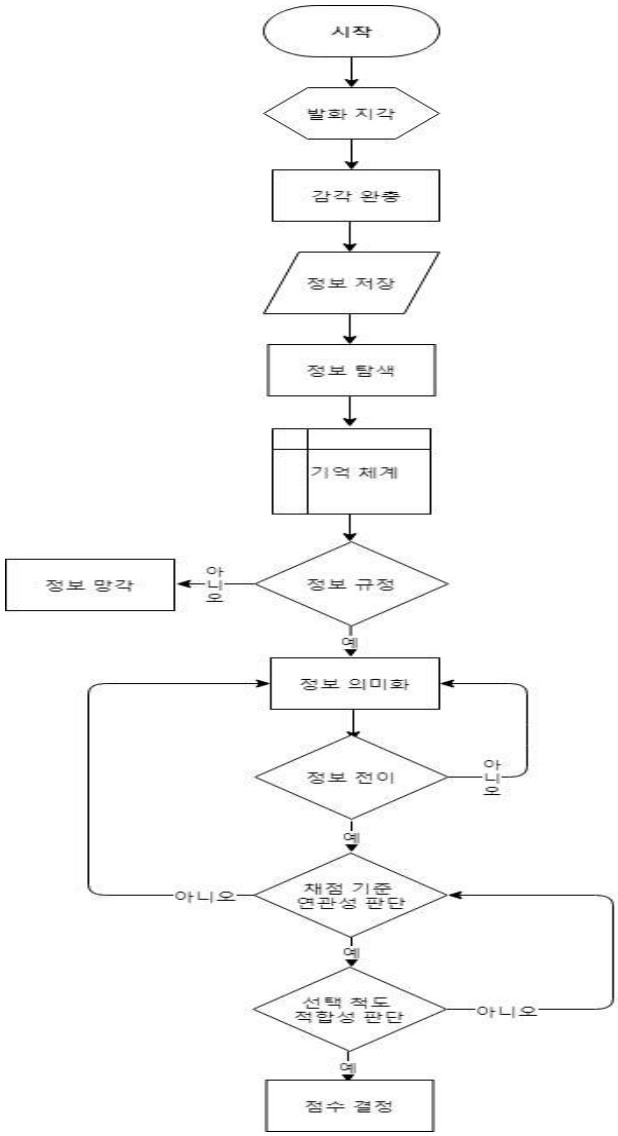
구에서는 채점자별 말하기 평가 과정이 검사 문항이나 채점 척도에 따른 양상의 차이를 알아보기 위하여 구어 담화 자료의 체계적인 수집 방법인 언어 보고 프로토콜(verbal report protocol, Ericssson & Simon, 1993)을 바탕으로 채점 과정에 초점을 맞추어 수정한 '채점 과정 보고법'을 적용하여 채점자가 채점 과정을 보고한 자료를 수집하였다.

3. 말하기 평가 채점 과정 모형과 유형

본 절에서는 앞서 살펴본 채점 과정에 관한 추론과 추단의 관점과 말하기 평가 채점 과정의 영향 요인, 그리고 채점 과정에 영향을 미치는 채점자의 외재적·내재적 요인에 대한 검토를 바탕으로 말하기 평가의 채점 과정을 설명하기 위하여 가설적인 모형을 제안하고자 한다. 또한 말하기 평가의 맥락에 따라 채점 과정이 어떻게 이루어질 수 있는지를 평가 방법과 채점 방법, 채점 척도 구성의 측면에서 살펴보고자 한다.

1) 말하기 평가 채점 과정의 가설적 모형

말하기 평가의 채점 과정은 채점자가 평가 결과를 도출하기 위하여 수행하는 일련의 활동으로 이루어진다. 이와 관련하여 본 연구에서는 말하기 평가 채점 과정에 대한 실증적인 접근을 위하여 말하기 평가 채점 과정을 구성하는 이론적 가설을 설정하고, 각각의 과정 구성 요소의 관계를 나타내기 위하여 순서도를 사용하여 모형으로 제시하였다([그림 Ⅱ-6] 참조).



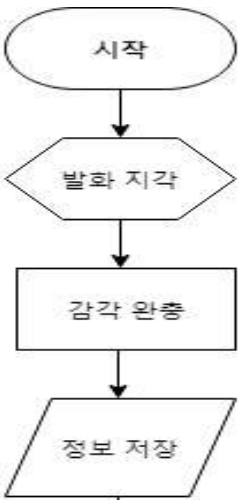
[그림 Ⅱ-6] 말하기 평가 채점자의 인지적 채점 과정 모형

[그림 Ⅱ-6]은 채점자의 인지적인 채점 과정을 직관적으로 나타낸 것이다. 제시한 모형은 채점 과정에서 이루어지는 인지적인 처리에 따라 세 개의 세부적인 과정으로 구성된다. 첫 번째인 '정보 수집 과정'은 수험자의 말하기 평가 응답에 대한 지각이 이루어진 것에서부터 단기 기억에 선택된 정보를 저장하기까지의 과정에 해당한다. 두 번째는 수집한 정보가 평가와 관련하여 어떠한

의미를 갖고 있는 것인지를 판단하여 상위 인지 체계로 전이시키기까지의 '정보 판단 과정'이다. 세 번째 과정은 채점 척도를 바탕으로 점수를 부여하는 '점수 결정 과정'이다.[16]

(1) 지각 체계를 통한 채점 정보의 수집 과정

말하기 평가의 채점 과정에 관한 첫 번째 가설은 채점자가 청각적 지각 활동을 통하여 채점 실행에 필요한 정보들을 수집한다는 것이다([그림 8] 참조).



[그림 Ⅱ-7] 말하기 평가 채점자의 정보 수집 과정

[그림 Ⅱ-7]은 채점자가 응답 청취를 통해 채점 정보를 수집하는 과정을 제시한 것이다. 채점자는 채점을 시작한 후에 발화에 대한 지각을 통해 음성 자료의 상태를 점검하고, 채점이 가능한 지를 확인한다. 다음으로 지각한 발화 중에 감각 완충 장치를 통해 저장할 정보에 대한 선택적 집중이 이루어지며, 이 때 인식한 정보는 단기 기억으로 저장하여 판단 과정을 거치도록 한다.

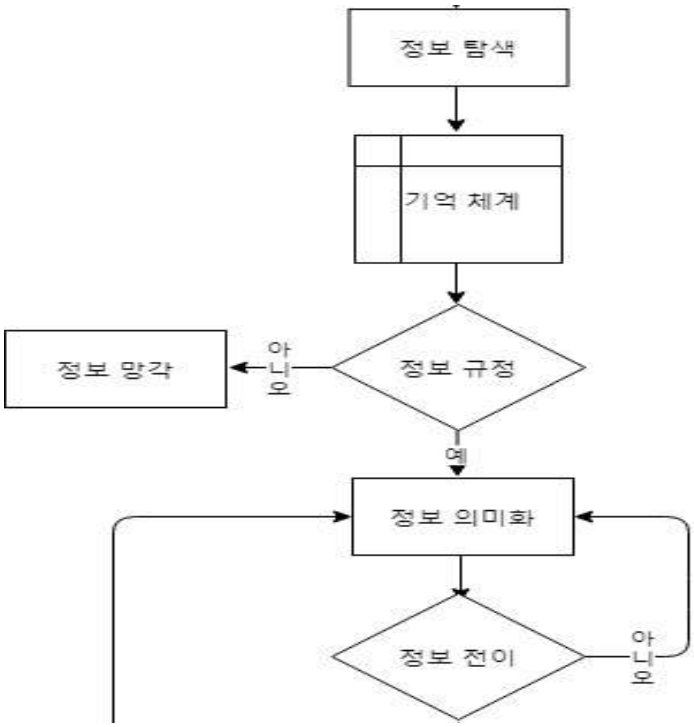
말하기 평가에서 채점자는 수험자의 발화를 청취 활동을 통해 채점의 근거

16) 각각의 과정은 채점자의 머릿속에서 이루어지는 것이기 때문에 단계적이기 보다는 연쇄적으로 일어난다고 볼 수 있으며, 채점자의 특성에 따라서 특정 과정을 축소 또는 확장하여 진행될 수 있을 것이다.

를 확보해야 한다. 채점자의 청각 기관을 통하여 수집된 정보는 청각적 지각 체계의 구동 방식에 따라 자동적으로 들리는 상향식 음성 정보와 의도적으로 들은 하향식 음성 정보로 나눌 수 있다. 수험자의 응답은 채점자의 청각 체계를 거치면서 선택적인 지각과 인식의 영향을 받기 때문에 감각의 완충 과정을 거치면서 정보의 소실이 일어날 수 있다. 채점자가 응답에서 기대하는 바가 있거나, 수험자가 반복해서 나타내는 특징들은 그렇지 않은 정보에 비해 상대적으로 기억 체계에 잘 저장될 수 있으며, 이는 채점 과정에서 수험자의 말하기 능력에 대한 전반적인 인상으로 작용할 수 있다.

(2) 인지 체계를 통한 정보의 판단 과정

말하기 평가의 채점 과정에 관한 두 번째 가설은 채점자가 청취한 정보를 자신의 인지 체계와 연계하여 판단한다는 것이다([그림 Ⅱ-8] 참조).



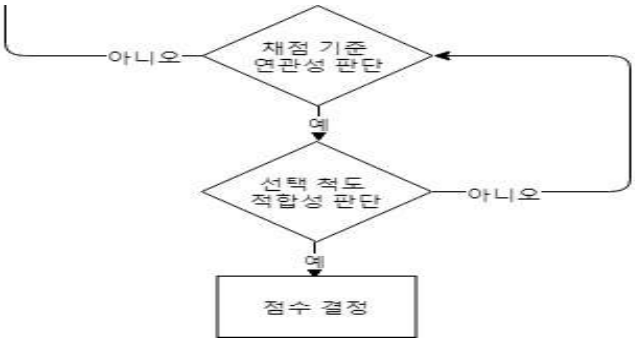
[그림 Ⅱ-8] 말하기 평가 채점자의 정보 판단 과정

[그림 Ⅱ-8]에서 제시한 정보 판단 과정은 먼저 단기 기억에 저장된 응답 정보에서 주목해야 할 정보를 탐색하고, 남은 채점 과정에서 사용할 수 있도록 장기 기억 체계에 저장하는 것으로 이루어진다. 다음으로는 해당 정보가 무엇인지를 판단한 후에 그러한 특징을 나타내는 것의 의미를 구체화하여 점수 결정을 위한 정보로 전이하는 과정이 이루어진다.

채점 과정에서 작동하는 채점자의 인지 체계는 채점 수행과 관련이 있는 말하기 교육 및 평가의 경험, 이론적 지식, 평가 훈련 및 교육의 경험 등으로부터 형성한 기억 체계를 바탕으로 정보를 판단한다. 이 과정에서 의미가 없거나 중요하지 않다고 인식한 정보는 단기 기억에 저장한 것이었기 때문에 수 초에서 수 분 안에 사라질 수 있다. 인지 체계를 바탕으로 의미가 있는 정보로 인식한 응답 정보는 점수 결정 과정에서 활용하기 위하여 특성을 규정하고, 관련된 정보를 바탕으로 의미를 구체화하게 된다. 구체화된 정보는 점수 결정을 위한 메타인지 체계로 '전이(transfer)'시키게 되는데, 같은 정보의 반복적인 처리에 의해 전이가 강화될 수 있으며, 반대로 특정 정보에 대한 전이가 강화(lateral transfer)되거나 약화(negative transfer)되어 나타날 수 있다.

(3) 상위 인지 체계를 통한 점수 결정 과정

채점 과정에 관한 세 번째 가설은 채점자가 인식한 정보에 대한 반성적 처리를 통해 점수를 결정한다는 것이다. 상위 인지는 이전에 수행한 인지적 활동에 대한 성찰과 평가를 위한 것으로서 인지에 관한 지식과 조절로 이루어져 있으며(Schraw, 1998), 이는 인지적으로 처리한 정보에 대한 재인식과 해석, 평가 등을 통해 채점자가 점수를 확정해 가는 사고 과정과 관련된다([그림 Ⅱ-9] 참조).



[그림 Ⅱ-9] 말하기 평가 채점자의 점수 결정 과정

[그림 Ⅱ-9]는 채점자가 점수 결정을 위하여 정보 판단 과정에서 전이된 내용을 바탕으로 채점 기준과 척도를 적용하여 점수를 결정하는 형식적인 과정을 나타낸 것이다. 채점자는 점수 결정을 위하여 인지적으로 처리한 정보가 채점 척도의 여러 정보 가운데 어떤 정보와 어떻게 연관되는지를 파악하여 점수를 부여한다. 공식적으로 제공하는 채점 척도의 내용 이외에 채점자는 선택하지 않은 점수가 수험자의 말하기 능력을 설명할 수 있는 근거가 될 가능성은 어떠한지를 검토할 수 있다. 또한 채점을 수행하면서 형성한 평가 문항에 대한 전반적인 수험자의 응답 수준이나 특성에 대한 주관적인 참조틀을 지침으로 적용하여 점수를 결정할 수도 있다. 끝으로 채점자는 채점 과정에서 인식한 정보에 대한 메타 인지 활동을 통해 내린 결론을 재검토할 수 있으며, 이 과정에서 고려하지 않았던 점수 결정의 근거를 표상할 경우 수집한 정보의 타당성을 재검토한 후에 점수 결정 과정을 반복할 수 있다.

이상에서 살펴본 말하기 평가의 채점 과정 가설과 그에 따른 모형은 이론적인 검토를 바탕으로 기본적인 모형으로서 구안한 것이며, 따라서 실제 평가 상황에서는 예상하지 못한 여러 평가적인 변수의 개입으로 인하여 채점 과정의 전개가 달라질 수 있을 것으로 예상된다. 이와 관련하여 다음 항에서는 말하기 평가의 채점 방식에 따라 채점 과정이 어떻게 달라질 수 있는지를 살펴보고, 각각에 해당하는 유형을 살펴본다.

2) 말하기 평가의 형식에 따른 채점 과정 유형

말하기 평가의 채점 과정은 청취 정보 판별과 직관적 접근을 통한 점수 선정, 그리고 선정한 점수를 정당화하여 결정하는 체계로 이루어진다. 채점 과정은 말하기 평가의 평가 방법과 채점 방법, 채점 척도 구성에 따라 절차와 상호작용 요소의 차이가 있다.

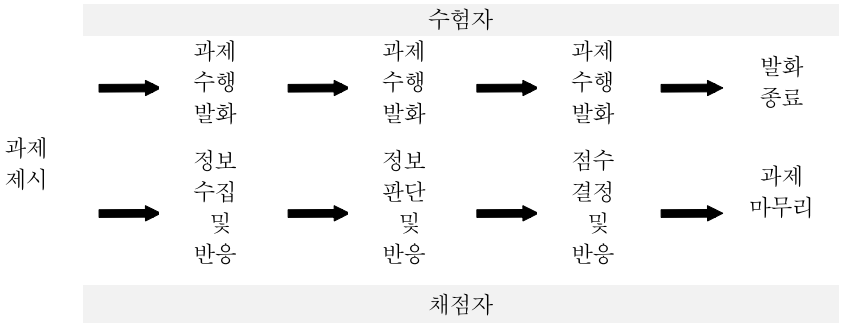
(1) 평가 방법에 따른 채점 과정 유형

말하기 평가의 채점 과정은 평가의 방식(mode)에 따라 접근의 차이가 나타나며, 이는 평가의 목적과도 관련이 깊다. 말하기 평가의 방법은 평가 상황에서 수험자가 시험관과 직접 마주보면서 이루어지는 직접식 평가와 직접 만나지는 않지만 녹음 또는 촬영한 자료를 바탕으로 이루어지는 비대면 준직접식 평가, 그리고 직접적으로 말하기 능력을 측정하는 것이 아니라 문장 완성하기 활동과 같이 관련된 내용으로 평가를 하는 간접식 평가가 있다. 이 중에서 본 연구에서 다루고 있는 수행 평가로서의 말하기 평가에는 직접식 평가와 준직접식 평가가 있다.

직접식 평가의 채점 과정에서 채점자는 면담자로서의 역할도 수행해야 하기 때문에 수험자 발화에 대해 응답하거나, 다시 질문을 하는 등의 반응을 해야 한다([그림 Ⅱ-10] 참조). 면대면 대화 과제와 같이 채점자가 수험자와 직접적으로 상호작용을 해야 하는 경우에 채점자는 언어적 반응을 통해 수험자의 과제 수행 발화에 직접적으로 관여하게 된다. 이와 동시에 채점자는 수험자의 과제 수행 발화에 나타난 채점 관련 정보들을 파악하여 점수를 부여해야 한다. 이러한 직접식 말하기 평가의 채점 과정에 발생하는 채점자의 인지적 부담을 해소하기 위해서 채점자는 수험자와 상호작용하면서 점수를 결정하기 위해 필요한 주요 발화 정보를 기록하거나 채점 척도에 관련된 특징을 표시하였다가, 수험자가 퇴장한 후에 해당 정보를 바탕으로 점수를 부여할 수 있다.[17] 그러나 평가 상황에서 채점자가 기록할 수 있는 정보가 제한적이기 때문에 많은 부분

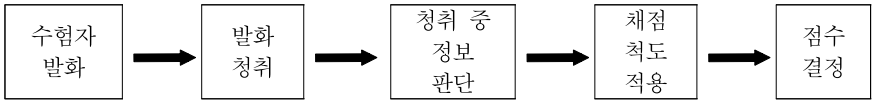
17) 채점자가 수험자와 마주 보고 이루어지는 말하기 평가 상황에서 면담자의 역할도 함께 해야 하기 때문에 채점 수행에 많은 방해를 받을 수 밖에 없는데, 이와 관련하여 영국 캐임브릿지 대학의 UCLES에서 개발한 영어 능력 평가 가운데 CPE의 경우에는 시험장에 면접관과 채점자를 함께 배치하고, 면접관은 종합적인 채점을 하고, 전문 채점자는 분석적인 채점을 하여 상호 보완적이면서, 실제적인 접근을 취하고 있다.

은 수험자 발화를 회상하면서 발화에 대한 인상을 중심으로 채점을 수행할 가능성이 높다.



[그림 Ⅱ-10] 직접식 말하기 평가의 채점 과정

준직접식 평가의 경우에는 채점자가 녹음된 수험자의 발화를 들으면서 채점을 하기 때문에 직접식 평가에 비해 상대적으로 채점 과정의 방해 요소가 적다([그림 Ⅱ-11] 참조). 대신에 준직접식 평가의 채점 과정에는 청취 자료와 채점 규칙이 큰 영향을 끼칠 수 있으므로, 그에 대한 수립이 이루어져야 한다. 먼저 평가 가능한 수준의 수험자 발화 녹음이 이루어질 수 있도록 품질이 확인된 녹음 장비로서 평가 진행을 방해하지 않는 것을 선택하여 사용해야 한다. 또한 시험장의 환경도 녹음을 방해하지 않도록 점검이 이루어져야 한다. 다음으로 준직접식 평가를 위한 채점 규칙은 채점자의 채점 일관성과 경제성 확보에 영향을 미칠 수 있는데, 채점이 어떤 환경에서 이루어져야 하며, 수험자 응답을 청취하는 과정과 방법은 어떠해야 하는지, 그리고 채점 중에 발생한 문제에 대한 해결 방법 등의 내용을 포함한다. 특히 준직접식 말하기 평가의 채점에서는 불필요한 반복 청취로 인한 평가 결과 영향을 통제하기 위하여 사전에 청취에 관한 규칙을 분명하게 안내해야 한다.



[그림 Ⅱ-11] 준직접식 말하기 평가의 채점 과정

본 연구에서는 직접식 말하기 평가의 채점 과정에 관여하는 채점자와 수험자의 상호작용에 따른 영향을 통제하고, 인지적인 정보 처리 과정으로서 채점자 중심의 채점 과정에 초점을 두기 위하여 준직접식 말하기 평가에서의 채점 과정 연구를 수행하고자 한다.

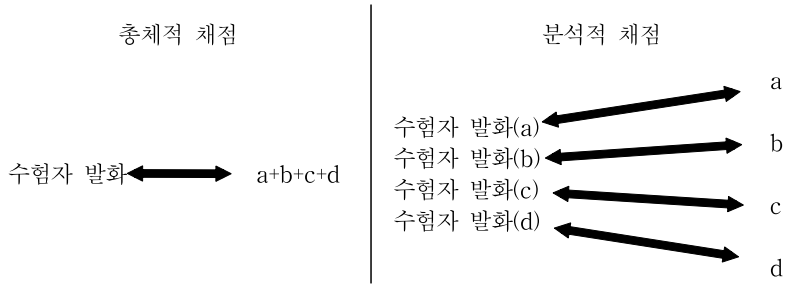
(2) 채점 방법에 따른 채점 과정 유형

수행 평가로서 이루어지는 말하기 평가의 채점은 일반적으로 평가의 목적과 결과 활용의 측면에서 총체적 채점과 분석적 채점으로 나눌 수 있다. 총체적 채점(holistic rating)은 순위 결정이나 전반적인 특징 파악, 또는 특정 능력의 측면에 한정하여 평가를 하는 상황에서 주로 사용하는 채점 방식이다. 분석적 채점(analytic rating)은 평가를 통해 측정하는 요인이 다양한 세부 구성 개념으로 이루어져 있으며, 이들을 각각으로 구분하여 채점하였을 때 얻을 수 있는 교육적인 효용이 큰 경우에 사용한다. 분석적 채점 방식은 평가 결과로서 얻을 수 있는 정보가 다양하고, 총체적 채점에 비하여 상세한 내용을 포함하고 있기 때문에 평가 이후에 수험자의 교수·학습을 어떻게 할 것인지에 관하여 유용한 정보를 얻을 수 있다.

총체적 채점의 채점 과정은 청취, 정보 인식, 채점 척도 연계, 점수 결정으로 이루어지는데, 과제 수행에 나타난 전반적인 측면을 고려해야 하기 때문에 수행 정보 가운데 어떤 특징을 중심으로 점수 결정이 이루어졌는가에 따라 평가 결과에 차이가 나타날 수 있다. 분석적 채점도 기본적인 채점 과정은 종합적인 채점과 비슷하지만, 이러한 과정을 측정하고자 하는 구인의 수만큼 반복해야 하고, 반복의 과정에서 발생한 인지적인 변화가 여러 요소에 대한 채점 과정을 거듭하면서 서로 상호작용이 일어나며, 그에 따라 발생한 영향이 평가 결과 전반에 나타날 수 있다.

[그림 Ⅱ-12]는 총체적 채점과 분석적 채점 방법에 따른 채점 과정의 차이를 나타낸 것이다. 총체적 채점은 수험자 발화에 대하여 평가 준거 a, b, c, d를 합하여 한 번의 종합적인 점수를 결정하는 것으로 채점 과정이 이루어진다면, 분석적 채점은 평가 준거에 따라서 수험자의 과제 수행 발화에 나타난 해당 준거에 관한 측면을 각각 분리하여 채점을 해야 한다. 채점 과정에서 고려하는

평가 준거들 사이에 공유하는 특징이 있거나, 영향 관계가 있을 경우에 특정 준거가 다른 준거에 영향을 끼칠 수 있으며, 이러한 현상을 확인하기 위해서는 채점 과정에서 채점자가 평가 준거에 따라서 점수 결정 과정에서 나타낸 특징이 무엇인지를 비교하는 것을 통해 접근할 수 있다.

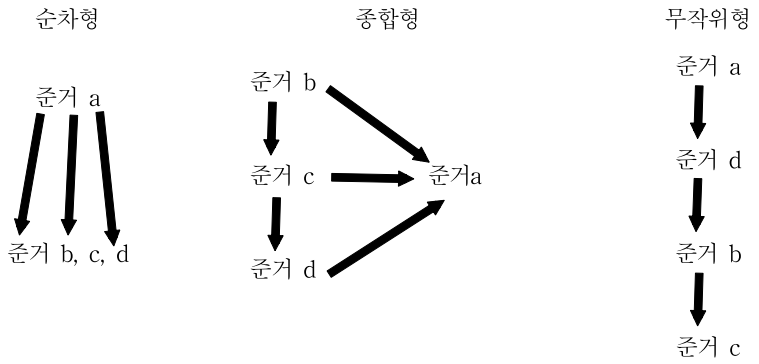


[그림 Ⅱ-12] 총체적 채점과 분석적 채점의 채점 과정

(3) 채점 척도 사용에 따른 채점 과정 유형

다음으로 말하기 평가 채점 과정의 구성에 관여하는 것은 채점자의 채점 척도 사용 방법이다. 채점자가 채점 척도를 어떻게 사용하느냐는 채점자의 채점 경향과 관련이 있으며, 이는 평가 결과 도출에도 영향을 끼칠 수 있다. 채점 척도는 평가 대상의 과제 수행으로부터 측정하고자 하는 심리적인 구인이나 영역인 평가 준거(criteria)와 그에 대한 수준을 구성하는 평가 척도(scale), 그리고 각 준거의 척도별 내용의 기술어(descriptor)로 구성한다(Fulcher, 2003). 채점 척도의 구성은 평가 목적에 따라 달라지는데, 교육적 활용을 고려하여 분석적 채점 방식을 선택한 경우에는 여러 요소를 구분하여 채점해야 하는 상황에서 어떤 평가 준거를 우선적으로 고려하는가, 또는 어떤 순서로 평가 준거를 채점하는가에 따라서 순차형과 종합형, 무작위형으로 나눌 수 있다. 순차형은 채점 척도에서 제시하고 있는 순서를 따라서 그대로 채점을 하는 경향을 말하는데, 대체로 앞서 제시하는 준거가 뒤에 배치한 준거에 비하여 포괄적이거나 상위의 것으로 볼 수 있다. 예를 들어 채점 척도에서 평가 준거로 a, b, c, d를 제시하고 있을 때, 순차형 채점자는 항상 a, b, c, d의 순으로 채점을 하는 것을

말한다. 다음으로 종합형 채점자는 평가 준거의 범주에 따라 상향식 접근을 취하며, 미시적이거나 하위에 해당하는 준거를 먼저 채점하고, 그 결과를 종합하여 전반적인 측면과 관계가 되는 준거에 대한 점수를 부여하는 것을 말한다. 무작위형은 채점 척도에서 제시하는 순서나 평가 준거의 범주 특성보다는 채점자가 회상할 수 있는 수험자 응답의 특징이나 두드러진 어떤 특성을 따라서 채점하는 것을 말한다([그림 Ⅱ-13] 참조).



[그림 Ⅱ-13] 채점 척도 사용에 따른 채점 과정 유형

본 연구에서는 말하기 평가 채점 과정을 구성하는 추론과 추단의 관점, 그리고 채점 과정 영향 요인에 대한 검토를 바탕으로 구안한 말하기 평가의 채점 과정 모형에서 채점자의 지각 체계와 인지 체계, 메타 인지 체계의 작동과 특성을 확인하기 위하여 채점자의 채점 과정에 대한 실증적인 접근을 취하고자 한다. 이를 위하여 통합 방법 연구 설계를 바탕으로 실제 말하기 평가 채점 과정을 통해 도출한 평가 결과(양적 자료)와 평가 결과를 도출하기까지의 채점 과정 자료(질적 자료)를 수집한다. 양적 · 질적 자료는 각각 평가 결과와 평가 결과를 도출하는 과정이라는 점에서 인과적 관계를 맺고 있는데, 이를 연계하여 해석하기 위해서는 먼저 양적 자료 분석을 통하여 평가 결과에 나타난 채점자의 전반적인 채점 경향과 편향성에 대한 파악이 이루어져야 한다. 이러한 양적 분석 결과는 질적 자료 분석 설계의 틀로서 채점자별 채점 과정의 차이를 알아보기 위한 도구를 설정하는 기초가 된다.

III. 말하기 평가의 채점자 영향 분석

본 장에서는 한국어 말하기 평가의 평가 결과를 분석하여 채점자가 채점 과정을 통해 나타낸 채점자 영향을 알아보고자 한다. 이는 채점 과정을 연구하는 이유가 점수의 변화나 측정의 오류를 일으키는 요인을 확인하는 것이라는 관점과 관련이 있다(Van Moere, 2013: 1364). 채점자 영향[18]이란 채점자가 채점을 수행하는 과정에서 나타낸 주관적인 특성이라는 점에서, 채점 과정을 통해 발생하는 것으로 인식되어 왔다.말하기 평가의 채점자 영향을 알아보기 위하여 18명의 중·고급 한국어 학습자를 대상으로 구성형 응답 문항으로 이루어진 '한국어 말하기 능력 시험'을 실시하였다. 채점자 영향의 분석은 전반적인 평가 결과의 특징을 파악하기 위한 CTT 분석과 구체적인 채점자 영향을 알아보기 위하여 MFRM 분석을 통해 접근하였다.

1. 채점자 영향 측정을 위한 말하기 평가 도구의 구안

한국어 말하기 평가의 평가 결과에 나타난 채점자 영향을 확인하기 위하여 구성형 응답 문항으로 이루어진 말하기 시험을 실시하고, 수험자들로부터 응답 자료를 수집하였다. 응답 자료 수집을 위하여 개발한 말하기 시험은 '한국어 말하기 능력 시험'이다. 본 시험은 성취도 평가로서 수험자인 중·고급 한국어 학습자들이 해당 수준의 말하기 기술에 어느 정도로 도달하였는가를 알아보는 것을 목적으로 하며, 그 중에서도 종합적으로 말하기 능력을 평가하기 위하여

18) '채점자 영향(rater effect)'은 수행 평가에서 평가 결과를 도출하는 데에 채점자가 관여함으로 인하여 나타나는 경향성으로서, 지금까지의 선행 연구에서는 채점자의 전공이나 경력과 같은 간접적인 배경변인을 주목하여 왔다. 그러나 채점자의 배경변인이 직접적으로 채점에 어떻게 관여하였는지를 규명하는 것은 교육과 평가 경험 속에서 자리 잡은 채점자의 특성을 고려하지 못한다는 한계가 있다. 이와 관련하여 미포드와 울프(Myford & Wolfe, 2003)는 채점자 영향을 평가 결과에 직접적으로 영향을 미치는 체계적인 채점 경향으로 보아야 한다고 하였으며, 그 증거로서 복합적인 통계적 증거를 종합하여 채점자 영향을 분석하는 사례를 제시하였다. 본 연구에서는 한국어 말하기 평가 채점 과정에 나타난 직접적인 채점자 영향을 확인하기 위하여 채점자의 채점 경향에 관한 증거들을 바탕으로 채점자 영향을 확인한다.

주제 말하기나 과제 기반 말하기 문제와 같이 수험자가 구성한 과제 응답 담화에 대해 채점자인 교사가 점수를 부여하는 학기말 말하기 시험의 상황을 바탕으로 개발하였다.[19] 시험에 사용한 구성형 응답 문항은 문법이나 발음의 정확성과 같이 지식을 평가하는 단답형 말하기 문항과 달리 실제적인 말하기 능력을 측정하여 구체적인 학습자의 말하기 능력 특성을 파악할 수 있다는 장점을 갖고 있다. 수행 평가로 이루어지는 말하기 시험의 기본 요건은 유도한 응답(responses)과 수행의 조건(condition), 그리고 제시한 자료(stimulus)(Slater, 1980)로 이루어지며, 본 연구에서는 수험자가 4개의 과제별 응답 조건과 자료를 바탕으로 응답을 구성하도록 문항을 개발하였다.

1) '한국어 말하기 능력 시험' 문항의 개발

(1) 말하기 평가 방법의 선정

말하기 평가의 방법은 평가의 형식에 따라 달라진다. 일반적인 수행 평가로서의 말하기 평가는 교사 또는 평가자와 학습자 또는 수험자가 면대면으로 진행하는 직접식 평가로 이루어진다. 직접식 평가와 유사한 상황을 추구하면서도 평가 맥락의 일관성을 보장할 수 있는 방식으로는 녹음기나 컴퓨터를 활용한 준직접식 말하기 평가가 있는데, 평가의 공정성을 보장해야 하거나 대규모 평가를 시행해야 하는 경우에는 준직접식 평가를 하는 것이 일반적이다 (Stansfield & Kenyon, 1992; Fulcher, 2003). 준직접식 평가는 양방향 소통이 가능한 직접식 평가와 달리 일방향 소통만 이루어지기 때문에 산출 중심의 말하기 평가 방법이라고 볼 수 있다.

본 연구에서는 수험자들의 말하기 평가 응답에 대한 채점자의 채점 경향을 알아보기 위하여 준직접식 평가를 통하여 학습자 음성 자료를 수집하였다. 준직접식 평가 방식을 선택한 이유는 준직접식 평가가 지문과 제시 자료의 표준

19) 본 연구에서 수험자의 말하기 과제 수행 응답을 수집하기 위한 평가 도구로 성취도 평가로서의 말하기 시험을 설정한 것은 성취도 평가가 한국어교육에서 가장 보편적인 말하기 평가라는 점을 고려한 것이다. 다른 측면에서는 본 평가 도구에서 사용한 구성형 응답 문항이 TOEFL이나 IELTS와 같은 대규모 말하기 능력 평가에서 사용되고 있다는 점에서 향후 이루어질 대규모 한국어 말하기 평가의 문항 및 채점 관련 연구와 연계될 수 있을 것으로 보았다.

화가 가능하며(Stansfield & Kenyon, 1992), 평가 실천의 일관성을 확보할 수 있으면서(Qian, 2009), 직접식 평가와 유사한 수준의 결과를 확보할 수 있다(Kiddle, & Kormos, 2011)는 장점을 갖고 있다. 본고는 말하기 평가 채점 과정에서 기인한 평가 결과에 대한 채점자 영향을 확인하는 데에 초점이 있으며, 따라서 채점자와 수험자의 불분명한 상호작용의 영향을 통제할 수 있는 준직접식의 말하기 평가를 실시하였다.

(2) 말하기 평가 문항 유형의 선정

본 연구에서는 채점자의 채점자 영향과 점수 결정의 관계를 알아보기 위하여 '구성형 응답 문항(constructed-response item)'으로 평가 과제를 구성하였다. 준직접식 한국어 말하기 평가 문항 유형은 발화 길이에 따라 단답형 문항과 구성형 문항으로 나눌 수 있는데, 단답형 문항이 발화 맥락에 대한 고려보다는 어휘나 문법 사용을 중심으로 평가하는 유형인 반면에 구성형 문항은 과제에 맞추어 응답 내용을 구성하여 응답하는 유형으로 학습자가 보유한 지식과 능력을 측정하는 대표적인 수행 평가 문항 유형이다(Cizek & Bunch, 2007: 27).

구성형 문항은 응시자들에 대한 제약이 적고, 주도적으로 응답해야 하도록 책임감을 갖게 한다는 장점도 갖고 있는데, 복합적인 언어 능력 산출을 평가하기 때문에 문항의 형식이 복잡하고 채점 요소가 다양하여 복수 채점을 필요로 하기 때문에 평가 결과의 타당성 확보가 필요하다는 점, 산출 비용 부담이 큰 문항 유형으로 알려져 있다(McNamara, 2000: 30-31). 단답형 문항은 정오에 대한 판단으로 채점이 이루어질 수 있어 직관적이며, 결과 산출의 의미가 분명하여 초급 수준의 말하기 능력 평가에서 주로 사용하는 반면에 구성형 문항은 담화를 구성하는 여러 잠재적 능력 특성 요소들을 고려하여 채점해야 하기 때문에 중·고급 수준의 말하기 평가에서 주로 사용하고 있다.

준직접식 말하기 평가에서 구성형 응답 문항은 과제와 수험자의 거리를 나타내는 맥락화 정도에 따라서 구분할 수 있으며, 이는 발화 곤란도(difficulty)를 높이는 요인이 된다(Brown & Yule, 1983).[20] 본 연구에서는 채점 과정에서 채

20) Brown & Yule(1983)에서 언급한 인지적 부담이란 과제 수행 과정에서 고려해야 하는 요소들과 관련이 있는데, 시간 순서나 이야기 순서를 따라서 말을 하는 것보다 다양한

점자가 평가 문항에 따라 고려해야 하는 맥락화 요구 수준의 차이를 두어 문항 변별도를 확보하고자 하였으며, 평가 결과에서 실제적인 차이를 확인한다.

(3) 말하기 평가 문항의 작성

본 연구에서는 중·고급 수준의 한국어 학습자가 해당 수준의 학습을 종료한 후에 학습 성취도를 파악하고 이후의 학습 목표를 설정하기 위한 말하기 평가를 개발하는 것을 목적으로 한다. 이를 위하여 중·고급 한국어 학습자용 구성형 말하기 평가 문항을 작성하기 위한 참조틀로 국제통용한국어표준교육과정(국립국어원, 2017) 및 2010년 이후 발간된 한국어 교재의 말하기 능력 관련 내용을 분석하였다. 그 가운데 말하기 화제 및 과제와 관련하여 제시하고 있는 내용들을 정리하고, 이를 바탕으로 평가 문항의 화제와 기능을 선정하였다.

① '국제통용한국어표준교육과정'의 말하기 교수·학습 목표 및 화제 분석

'국제통용한국어표준교육과정'(이하 표준교육과정)은 한국어교육의 표준화를 목표로 국가 수준에서 발간한 교육과정 모형이며, 2010년에 발간된 1단계 표준 모형을 바탕으로 현장 적합성을 높이기 위하여 수정·보완한 자료이다(국립국어원, 2017:1). 표준교육과정은 6개 등급 체계로 구성되었으며, 한국어 기능과 수준별로 언어 지식과 언어 기술, 텍스트, 문화, 평가에 관한 내용을 제시하고 있다. 본 연구에서 표준교육과정 가운데 말하기 평가 문항의 내용과 맥락 구성과 관련하여 중·고급 학습자용 언어 기능 및 과제와 말하기 교수·학습 관련 내용, 수준별 텍스트 자료 활용 내용, 그리고 말하기 평가 체제에 관한 기술을 참조하였다.

ⓐ 기능 및 과제

정보를 고려하면서 설명하는 것이 어렵고, 담화 차원으로 확장하여 담화를 구성해야 하는 경우에는 더 큰 어려움이 따른다는 것이다.

표준교육과정에서는 한국어교육을 위한 언어 기능을 정보 요청 및 전달, 설득과 권고, 태도 표현, 감정 표현, 사교적 활동의 범주로 구분하고, 각각의 세부 기능들을 수준별로 제시하고 있다(국립국어원, 2017: 44-48). 이 중에서 중·고급에 해당하는 3~6급의 중점 기능은 36개였다(<표 Ⅲ-1> 참조).

<표 Ⅲ-1> '표준교육과정(2017)'의 중·고급 말하기 기술의 범주

	수준			
범주	3급	4급	5급	6급
정보 요청 및 전달	서술, 확인, 대조	설명, 묘사, 비교, 수정	보고	진술, 기술
설득 및 권고	권유, 조언	경고, 충고, 지시	주의	-
태도 표현	추측, 거절 표현	동의, 반대, 부인, 의도 표현	문제 제기	-
감정 표현	만족 표현, 걱정 표현, 위로 표현, 후회 표현, 놀람 표현, 선호 표현	고민 표현, 불평 표현, 안도 표현	심정 표현	-
사교적 활동	칭찬	-	-	-

<표 Ⅲ-1>의 수준별 목록을 살펴보면, 정보 요청 및 전달하기에서 최고 수준에 해당하는 세부 기술로 '진술하기'와 '기술하기'를 포함하고 있으며, 반대로 사교적 활동 기능에서는 '칭찬하기'가 유일하였다. 전체 범주 가운데 준직접식 평가에서 활용하기 적합한 것은 정보 요청 및 전달과 설득 및 권고, 태도 표현하기였으며, 감정 표현, 사교적 활동은 대인 의사소통 및 친교적 의사소통 상황에서 주로 사용하기 때문에 준직접식으로 이루어지는 평가의 특성을 고려하여 제외하였다.

본 연구에서는 과제 구성을 위하여 수준별 중점 언어 기술 가운데 핵심으로 각 언어 기술의 변별적인 특성을 고려하여 중급 수준에서는 '서술하기'와 '조언하기', 고급에서는 화제의 수준이 높은 '보고하기'와 '진술하기'를 선정하였다. 중급 수준 과제로 선정한 '서술하기'는 순서적인 언어 수행에 관한 것으로, 말하기 평가에서는 경험 말하기와 관련이 있다. '조언하기'는 대상에 대한 특

징을 고려하여 알맞은 내용의 조언을 하거나 구하는 내용의 언어 수행과 관련이 있는데, 말하기 과제로는 가까운 대상에게 생활 관련 조언을 하는 것은 낮은 수준으로, 가깝지 않는 대상에게 작업과 관련하여 조언을 구하는 것은 높은 수준에서 선택할 수 있다(국립국어원, 2017: 55). 고급 수준 중점 기술인 '진술하기'는 '설명하기'를 바탕으로 전문성과 추상성이 높은 화제를 다루는 능력을 요구한다. 그 다음 수준인 '보고하기'는 분석적인 사고를 바탕으로 체계적인 언어 수행 능력을 요구한다.

ⓑ 말하기 교수·학습 관련 내용 검토

다음으로 말하기 성취도 평가 개발을 위하여 수준별 교수·학습 목표와 내용에 대한 검토를 실시하고, 과제의 세부적인 요건을 구성하였다. 중·고급 수준의 말하기 교육의 목표와 내용에 대한 표준교육과정에 나타난 차이를 살펴보면, 화제에 대한 발화자의 친숙성과 전문성이 중급(3·4급)과 고급(5·6급)을 구분하고 있는 기준임을 알 수 있다(국립국어원, 2017: 150-151). <표 Ⅲ-2>를 보면, 중급에 해당하는 말하기 교육의 목표에서 주제에 대한 친숙성을 포함해야 한다고 기술한 반면에 고급에서는 친숙하지 않은 주제를 다룰 수 있어야 한다고 기술하고 있었으며, 직업과 학문 영역과 같은 개인적·사회적으로 전문성을 요구하는 영역에 대한 말하기 능력을 요구하고 있음을 확인할 수 있다. 이와 같은 중급과 고급 한국어 말하기 교육에서 요구하는 말하기 능력에 관한 차이는 전나영 외(2007: 283-290)에서도 확인할 수 있는데, 이 연구에서는 중급에 해당하는 3·4급 수준의 말하기 능력에 관한 화용적 기능으로 일상생활에서의 자유로운 말하기를 강조하였다면, 고급 수준인 5·6급에서는 전문적인 업무 수행에서의 정확하고 유창한 언어 사용을 강조하고 있다.

<표 Ⅲ-2> 표준교육과정의 중·고급 한국어 말하기의 목표와 내용

수준	목표	내용
6급	●친숙하지 않은 사회적·추상적 주제 및 자신의 직업이나 학문 영역에 대한 의견을 논리적으로 주장할 수 있음. ●자신의 전문 분야에 대해 상세하고 유창하게 말할 수 있음.	- 친숙하지 않은 주제나 자신의 전문 분야 관련 입장 발화의 논리성과 체계성 - 다양한 주제에 대한 토론이나 대담에서의 논리적으로 타당한 근거 말하기 - 적절한 한국어 대화 및 담화 구조와 전략 사용 - 대부분의 상황에서 한국인과 같은 자연스러운 발음 구사
5급	●친숙하지 않은 사회적·추상적 주제 및 자신의 직업이나 학문 영역에 대해 어려움 없이 설명할 수 있음. ●자신의 의견을 유창하게 말할 수 있음.	- 친숙하지 않은 주제 및 업무, 학문 관련 유창하고 타당한 설명과 주장 - 업무 및 학문 관련 공식 상황에서의 격식체 사용 - 다양한 매체 활용 상황에서의 적절한 발화 - 대부분의 상황에서 한국인과 같은 자연스러운 발음 구사
4급	●친숙한 사회적·추상적 주제와 자신의 관심 분야에 대해 비교적 유창하게 묻고 답할 수 있음. ●자신의 직업 관련된 업무 상황에서 요구되는 비교적 간단한 의사소통 가능	- 친숙한 주제와 관련된 생각의 유창한 발화 - 간단한 보고, 요청, 지시 - 주변 인물 및 상황의 사실적 묘사 - 친숙한 업무 상황 및 공식적 자리에서의 격식 표현 구분 - 원어민이 이해 가능한 한국어 발음 구사
3급	●친숙한 사회적·추상적 주제와 자신의 관심 분야에 대한 간단한 대화 가능. ●대화 상황을 어느 정도 구분하여 말할 수 있음.	- 친숙한 주제에 대한 간단한 대화 - 일상적 주제에 대한 유창한 발화 - 개인적인 경험이나 생각에 대한 간단한 담화 - 적절한 경어법 사용 - 원어민이 이해 가능한 한국어 발음 구사

이와 관련하여 본 연구에서는 표준교육과정에서 제시한 중·고급 수준의 한국어 말하기 교수·학습 내용 가운데 대표적으로 중급 수준에서는 화제의 친숙성과 발화 맥락의 단순성이, 고급 수준에서는 화제의 비친숙성과 발화 맥락의 격식성 및 체계성이 핵심이 된다는 점을 고려하여 본 평가 도구의 과제 구성도 이를 반영하고자 하였다.

ⓒ 유형별 텍스트 자료 검토

표준교육과정에서 제시하는 텍스트 자료 유형은 목적을 따라 '정보 전달과 이해', '문학적 반응과 표현', '비판적 분석과 평가', '사회적 상호작용'으로 나뉘는데, 본 연구에서 구성형 문항에 대한 말하기 평가 문항에서 활용할 자료

유형은 교육과정에서 ‘분석과 평가’를 목적으로 하는 자료 유형 가운데 ‘신문 기사’에 해당하는 것으로 ‘그래프, 도표’유형과 ‘사건/사고 보도기사’를 선정하였다. 이들은 모두 4~5급 수준에 해당하는 자료로서 형식성이 높지 않고, 일상적으로 학습자들이 접할 수 있는 자료 유형이라는 공통점을 갖고 있다. 자료의 수준을 최고급 수준인 6급이나 중급의 낮은 수준인 3급에서 고려하지 않은 것은 수준의 편차가 심하였기 때문인데, 가령 3급 수준의 텍스트로는 사진, 광고, 인터뷰 등이 있었는데, 이들은 친숙성이 높은 자료들이지만, 평가 상황에서는 다양한 정보를 함축하고 있어 해석의 일관성 확보에 문제가 발생할 수 있어 사용에 주의가 필요하다. 반대로 6급에 해당하는 자료로는 학술 보고서, 논문, 연구 계획서, 논설문, 평론 등이 있었는데, 모두 형식성이 매우 높고, 글쓰기나 읽기 능력 학습을 위한 용도로 사용하기에는 적합하나, 말하기 과제에 사용할 경우 읽기 부담이 과도하게 높아져 과제 수행에 심각한 영향을 미칠 수 있기 때문에 제외하였다.

ⓓ 평가

표준교육과정에서는 교실 평가 상황을 고려하여 대규모 평가인 TOPIK에 비하여 평가 수준을 낮게 설정하였음을 밝히고 있으며(국립국어원, 2017: 212), 각 1~6급의 수준별 평가 목표와 초·중·고급의 평가 체제를 제시하고 있다. 그 중에서 중·고급 수준의 평가 내용 중 말하기 평가에 관한 내용을 살펴보면, <표 Ⅲ-3>과 같다(국립국어원, 2017: 219-221).

<표 Ⅲ-3> 표준교육과정에서 제시한 중·고급 말하기 평가 체제

구분	중급	고급
평가 총괄 목표	• 친숙한 사회적 맥락 및 직업 관련 업무 처리 • 친숙한 주제에 대한 유창한 의사소통 • 격식체의 적절한 사용	• 덜 친숙한 사회적 맥락 및 업무나 학업 관련 기능 수행 • 친숙하지 않은 주제에 대해 비교적 유창하게 기능 수행 • 한국인이 자주 사용하는 담화 구조 이용
주제	• 친숙한/일상적 화제 • 일과 직업 맥락에서의 사회적 관계	• 개인 신상을 중심으로 한 사회적·추상적 화제 • 덜 친숙한 사회적·추상적 주제, 예술, 전문 분야 • 직장 생활, 업무, 전문 맥락에서의 소통
과제	• 공식적인 칭찬 • 선호에 대해 말하기 • 걱정과 안도한 경험 말하기 • 후회, 불평, 위로 표현하기 • 고민 이야기하기 • 거절하기 • 의도 말하기 • 개념이나 의미 추측하기 • 부인하기, 반대하기, 동의하기, 조언하기, 충고하기, 경고하기, 수정하기, 비교·대조하기, 서술하기, 묘사하기, 설명하기	• 친구에게 심정 표현하기 • 사회 문제 제기하기 • 사적/공적인 상황에서 주의 주기 • 관심 분야 관련 기술하기 • 추상적 내용 관련 기술하기 • 결과 및 조사 내용 보고하기 • 관심 분야 진술하기
언어 지식	• 친숙한 주제 관련 어휘 • 확장된 형태의 문장 구조 • 사회문화적 이해 기반 표현 • 음운 변동 및 억양 변화	• 친숙하지 않은 사회적·추상적 주제 관련 어휘 • 전문 분야 관련 어휘 • 확장된 문장 구조 • 특정 영역 사용 문법 표현 • 복잡한 수준의 음운 변화 발화 • 억양에 따라 달라지는 화용 의미 이해
텍스트 유형	• 친숙한 대화, 간단한 공문서, 설명문, 광고, 공고문 등	• 공문서, 서류, 시사 프로그램, 기사, 각종 계획서, 학술문

<표 Ⅲ-3>에서 제시한 표준교육과정의 말하기 평가의 목표와 내용을 보면, 앞서 살펴보았던 말하기 기능 및 과제, 교수 · 학습 관련 내용과 마찬가지로 말하기 평가에서 중·고급 수준을 화제의 친숙도를 기준으로 구별하고 있었다 (중급: “친숙한 사회적 맥락” / 고급: “덜 친숙한 사회적 맥락”). 과제의 측면에서는 중급에서는 특정한 맥락에서 사용하는 표현이나 말하기 기술에 대한 측면을 다루도록 하고 있는 것에 반해 고급에서는 기술적인 말하기, 보고하기,

진술하기 등의 기본적인 형식을 요구하는 말하기 기능 목록을 제시하고 있어 형식성에 따라 차이가 있음을 알 수 있다. 또한 과제에서 요구하는 언어 지식의 측면에서는 전문적인 언어를 사용하도록 요구하고 있음을 알 수 있다. 수준별로 과제와 관련하여 사용할 수 있는 자료도 차이가 있었는데, 중급에서의 자료는 실생활에 가깝고 단순한 것(대화문, 간단한 공문서)이라면, 고급의 자료는 형식성이 높고, 분석적인 접근이 필요한 자료(기사, 학술문)라는 점이다.

본 연구에서는 표준교육과정에서 말하기 평가와 관련하여 제시하고 있는 평가의 목표와 과제, 활용하는 텍스트 유형 등에 관한 내용 가운데 중급과 고급을 구분하는 기준으로 화제와 발화 맥락에 대한 거리를 고려하고 있다는 점과 수준별로 말하기 과제 수행을 위해 활용하는 자료를 다르게 구성하도록 제시한 점이 말하기 평가 문항의 수준을 차별화 하는데 도움이 될 것으로 판단하였다. 이에 '한국어 말하기 능력 시험'의 말하기 과제의 문항별 화제의 범주 및 활용 자료 선정 시에 이를 적용하여 중급 학습자와 고급 학습자가 모두 응답할 수 있는 과제이면서, 동시에 화제와 자료의 차이를 두어 수험자의 능력을 변별할 수 있도록 과제를 설계하였다.

(2) 2010년 이후 발간된 한국어교육 교재의 말하기 활동 분석

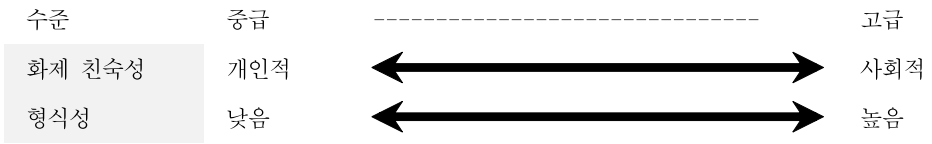
본 연구에서는 교육과정 검토를 통하여 중·고급 학습자용 구성형 말하기 평가 문항의 기능으로 서술하기, 조언하기, 보고하기, 진술하기를 선정하였다. 그리고 문항별 응답 수준에 따라 읽기 자료로 그래프와 신문 기사를 제공하여 기능 통합형 말하기 과제를 구성하였다. 실제적인 말하기 과제 선정과 관련하여 최근 10년간 발간된 중·고급 한국어 교재 4종 13권에 제시된 단원별 말하기 활동과 자료를 검토하였다.

중급 한국어 말하기 교재의 말하기 활동과 자료를 알아보기 위하여 검토한 교재는 모두 기능 통합형 교재였으며, 말하기 활동은 단원 주제를 따라서 기초적인 활동과 심화적인 활동으로 나뉘어 나타났다. 중급 교재 말하기 활동에서 기초 활동과 심화 활동 모두에서 개인적인 경험과 친숙한 화제에 대하여 말하는 활동이 중심을 이루고 있었으며, 교재에 따라서 심화 활동에서는 친숙하지 않은 화제에 대해 조사하거나 분석하여 발표하는 활동이나 사회적인 문제에

대한 토의, 문제 상황에 대한 의견 말하기 등의 구성형 응답 활동이 나타나고 있었다. 중급 교재의 말하기 활동 가운데 구성형 평가 문항 과제로서 학습자에게 친숙한 화제이면서 준직접식 평가에 적합한 독화 활동과 구성형 문항으로서 담화를 구성할 수 있는 활동 유형을 탐색하였다. 중급 교재에 제시된 활동 가운데 자신의 진로에 관하여 이야기를 하는 활동(이화 한국어4, 세종한국어5)이나 한국어를 잘 몰라서 실수한 경험 이야기하기(재미있는 한국어4, 세종한국어 6), 사회적 문제에 관해 말하기(재미있는 한국어 4, 서울대 한국어 4B, 이화 한국어4) 등이 공통적으로 나타나고 있었다.

고급 한국어 교재의 말하기 활동에 대해서도 말하기 기초 활동과 심화 활동으로 나누어 살펴보았다. 고급 교재에서는 중급 교재의 말하기 활동에 비하여 구체적인 요구 사항을 따라 응답해야 하는 말하기 과제들(예: '면접을 본 경험과 준비 방법에 대해 이야기하기', 세종한국어7)이나 '도표를 보고 최근의 경제 상황에 대해 설명하기'(재미있는 한국어 5), '환경 문제 기사를 바탕으로 최소화 방안 이야기하기'(이화 한국어5), '모국과 한국의 대학 진학률 비교하여 설명하기(서울대 한국어 5B)와 같이 자료 분석을 바탕으로 응답해야 하는 활동들을 제시하고 있었다. 이와 같은 중·고급 한국어 교재의 말하기 활동에 나타난 전반적인 차이는 어떤 화제를 어떤 형식을 통해 말하도록 할 것인가에 대한 것이며, 이는 표준교육과정의 말하기 평가 체제 관련 내용(<표 Ⅲ-3> 참조>)과 유사한 것이었다.

본 연구에서는 중·고급 한국어 교재의 말하기 활동에 나타난 수준별 특징을 바탕으로 중·고급 수준의 성취도 평가로서 '한국어 말하기 능력 시험'의 과제에서 화제의 친숙성을 고려하여 개인적인 것에서부터 사회적인 것으로 배치하고, 서사적인 말하기와 같이 개방적인 것부터 문제 해결적인 말하기와 같이 형식성이 높은 과제까지 배치하여 중·고급 학습자의 말하기 평가 참여에 대한 수월성을 확보하고자 하였다([그림 Ⅲ-1] 참조).



[그림 Ⅲ-1] '한국어 말하기 능력 시험'의 수준별 화제 및 형식성

③ '한국어 말하기 능력 시험'의 문항 구성

본 연구에서는 한국어 학습자들의 말하기 과제 수행 자료를 수집하기 위하여 표준교육과정 및 2010년 이후 발간된 한국어 교재에 제시된 중·고급 학습자용 말하기 활동의 화제와 과제에 대한 검토를 바탕으로 준직접식 한국어 말하기 평가의 구성형 응답 문항을 작성하였다.[21] 그리고 개발한 문항을 바탕으로 '중·고급 학습자용 한국어 말하기 능력 시험'으로 구성하고, 평가 실천 체재를 구성하였다. 본 평가의 목표는 한국어 학습자들의 구성형 말하기 과제에 대한 말하기 능력을 측정하는 것이다. 시험은 응시 방법 안내, 연습 문제, 본 시험의 과정으로 이루어졌으며, 빈 교실에서 학습자가 헤드셋을 착용하고, 컴퓨터 화면에 제시된 구성형 응답 문제를 읽고 문제에 답을 하는 것으로 진행하였다.

실시한 말하기 시험의 전체 문항은 모두 4 문항이며[22], 과제 수준에 따라 말하기 중심의 독립적 말하기 문항과 응답을 위해 관련 자료를 활용하는 기능 통합적 말하기 문항의 두 부분으로 나뉘어졌으며, 전체 과제는 언어 사용 기능을 중심으로 경험적 말하기, 상황적 말하기, 분석적 말하기, 설득적 말하기 기능으로 구성하였다. 경험적 말하기는 학습자의 개인적인 경험을 바탕으로 응답을 하는 과제를 제시하였으며, 상황적 말하기는 주어진 상황을 고려하여 응답하는 화용적 과제를, 분석적 말하기 문항은 제시된 도표를 정리하여 설명하는 과제를, 마지막 설득적 말하기 문항은 신문 기사를 읽고, 문제를 파악하여 해결방안을 제시하는 과제를 제시하였다(<표 Ⅲ-4> 참조).

---

21) 사용한 평가 도구의 실제 화면은 <부록 1>에서 확인할 수 있다.

22) 본 연구에서 학습자 응답 자료를 수집하기 위하여 문항은 자료 활용 여부와 말하기 세부 기능을 따라 작성하였으며, 문항의 수는 학습자들의 응시 부담을 최소화하면서, 말하기 시험에서의 신뢰도 최적화 최소 조건(Lee, 2006)을 고려하여 4개로 구성하였다.

<표 Ⅲ-4> 평가 문항 구성

문항	1	2	3	4
기능	경험적	상황적	분석적	설득적
맥락	개인적	관계적	사회적	사회적
화제	한국어 학습	직업	경제	환경
말하기 기능	경험 말하기	조언하는 말하기	분석하여 설명하기	문제 해결 방안 말하기
자료	없음	없음	그래프	신문 기사
준비 시간(초)	30“	45“	120“	300“
응답 시간(초)	60“	90“	150“	180“
수준	중급	중급	고급	고급

2) 채점 척도의 구성과 채점 방법의 선정

본 연구에서 실시한 말하기 평가의 평가 결과 확보를 위하여 이론적 검토를 바탕으로 채점 척도를 구성하고, 이를 활용하기 위한 채점 방법을 선정하였다. 채점 척도의 구성은 먼저 평가 준거를 선정하고, 평가 준거별로 척도에 따라 평가 기준을 기술하는 일로 이루어졌다. 그리고 채점의 방법은 구안한 말하기 평가 도구의 특성을 중심으로 채점자의 채점 수행 양상을 심층적으로 파악할 수 있어야 한다는 점을 고려하여 선정하였다.

(1) 채점 척도의 구성

본 연구에서는 한국어 말하기 평가 연구에서 제시한 한국어 말하기 능력의 평가 구인에 대한 검토를 바탕으로 평가 준거를 선정하였다. 또한 수험자의 능력 수준을 의미하는 척도는 상대적인 평가 척도로서 보편화된 총화평점 척도(summated rating scale)인 리커트 척도 방식을 선택하였다. 각 준거별 척도에 따라 기술한 채점 기준은 국내외 말하기 평가 연구에서 제시한 기술문의 내용을 바탕으로 작성하였다.

① 평가 준거의 선정

본 연구에서는 한국어 말하기 평가 선행 연구에서 제시한 한국어 말하기 능력 구인을 바탕으로 '한국어 말하기 능력 시험'의 채점을 위한 4개의 평가 준거를 선정하였다. 여기서 '평가 준거(criterion)'는 평가하는 구체적인 특성을 가리키며, 본 연구에서는 구안한 평가의 특성을 고려하여 중·고급 한국어 학습자의 독화형 말하기 능력을 측정할 수 있는 구성 요인(구인, construct)을 중심으로 선정하고자 하였다.[23)]

평가 준거 선정을 위하여 먼저 지금까지 한국어 말하기 평가 연구에서 제시한 말하기 능력의 구인을 살펴보았다. 검토 결과, 선행 연구에서 제시하고 있는 말하기 능력의 구인은 어떤 상황에서의 말하기 능력을 측정하고자 하느냐에 따라서 설정한 구인의 항목과 범주에 차이가 있었다(<표 Ⅲ-5> 참조).

연구자들이 반복해서 언급한 말하기 능력의 구인은 과제 수행에 관한 능력(전은주, 1997; 이영식, 2004; 지현숙, 2006; 김정숙, 2014; 원미진·김지영, 2018)과 담화를 구성하는 내용 및 조직에 관한 능력(김정숙·원진숙, 1993; 전은주, 1997; 이영식, 2004; 김정숙, 2014), 그리고 면대면 상호작용 상황에서의 의사소통 전략과 상호작용 능력(전은주, 1997; 전나영 외, 2006; 지현숙, 2006)이었다. 여기서 연구자 간 구인 설정의 차이는 주로 언어적인 영역을 넘어서는 구인인 '과제 수행 능력'과 '상호작용 능력'을 포함시키느냐에 관한 것으로 나타났다. 이와 관련하여 본 연구에서는 준직접식 말하기 평가를 실시하였다는 점과 수험자의 응답이 과제를 수행하는 것으로 구성된다는 점을 고려하여 말하기 능력 구인 가운데 평가 준거로 '과제 수행 능력'을 포함시키고, '상호작용 능력'은 제외하였다. 또한 언어적인 측면과 관련하여 선행 연구에서 반복적으로 나타났으며, 구인 간의 변별이 이루어지고 있는 최소한의 항목[24)]을 포함하여 최종적으로 '전반적 수행 능력', '발음 구사 능력', '어휘와 문법 사용 능력', '담화 구성 능력'을 평가 준거로 선정하였다.

---

23) 구인은 평가를 통해 수험자에 대해 측정하는 핵심적인 능력 요소 또는 개념을 의미한다. 구인의 설정은 능력의 개념을 구체화하고, 이론적인 정의를 바탕으로 이루어진다.

24) 평가 구인의 수는 채점의 효율성과 경제성에 직접적인 영향이 있는데, 5개 이하에서 경제적인 접근이 가능한 것으로 알려져 있다(Matsugu, 2013). 본고에서는 선행 연구의 차별적인 구인 설정 결과를 종합하고, 채점 효율성 확보를 위하여 4개로 하였다.

<표 Ⅲ-5> 한국어 말하기 평가 구인의 선정

연도	2019	1983	1993	1997	2004	2006	2006	2014	2017	2018
연구자	본 연구	노대규	김정숙 원진숙	전은주	이영식	지현숙	전나영 외	김정숙	민병곤 외	김지영 원미진
평가 구인	전반적 수행 능력	-	-	과제 수행력	과제 실현성	-	-	과제 수행 능력	-	과제수행 적절성
	발음 구사 능력	자연성 유창성	문법적 능력·사회언어학적 능력	발음 능력	유창성, 정확성,	발음	언어적 능력	발음 능력	발음과 유창성	유창성 및 발음
	어휘·문법 사용 능력	정확성 다양성		문법·어휘·사회언어학적 능력	언어의 풍부함	어휘, 문장 구조 규칙	언어적 · 사회언어적 능력		어휘와 문법	정확성 및 범위
	담화 구성 능력	-	담화구성 능력	구성력	대화의 적절성	내용 조직, 담화 운용	화용적 능력	내용 구성·담화 능력	내용, 조직	-
	-	-	-	의사소통 전략과 상호작용	의사소통 능력	상호작용 태도, 전략	-	상호작용 능력	-	-

② 채점 기준의 기술

채점 기준은 채점 척도에서 각 수준별 특성을 어떻게 측정할 것인가에 관한 내용을 기술한 것이다. 본고에서는 채점 기준의 기술을 위하여 선행 연구에서 제시한 말하기 평가 채점 척도 및 루브릭의 기술문을 검토하였으며, 이를 바탕으로 '한국어 말하기 능력 시험'의 채점 기준을 기술하였다(김정숙 외, 2006; 민병곤 외, 2017; Xi & Mollaun, 2006; Matsugu, 2013, IELTS, 2019).

먼저 '전반적 수행 능력'에 관한 채점은 과제 수행을 중심으로 총체적으로 응답을 평가하는 준거라는 점을 고려하여 채점 기준을 작성하였다. 여기서 가장 핵심인 과제 수행에 관해서는 문항별로 과제를 어느 정도 수준으로 수행하였으며, 응답 내용에 과제의 핵심적인 내용이 잘 나타나는 가를 기준으로 평가하도록 하였다. 다음으로 '발음 구사 능력'은 외국어 발음 평가의 핵심인 발음의 오류가 없음의 나타내는 정확성과 언어적인 전달력을 나타내는 유창성을 중심으로 기술하였다. '어휘·문법 사용 능력'은 응답에서 사용한 어휘와 문법이 정확하고, 또한 적절하게 이루어졌는가를 기준으로 기술하였고, 끝으로 '담화 구성 능력'은 응답 내용이 논리적으로 매끄럽게 연결되며, 문장과 담화가 긴밀하게 연결되어 있는지를 중심으로 기술하였다(<표 Ⅲ-6> 참조).[25]

채점 기준 기술 과정에서 각 평가 준거별 척도에 따른 세부 기술어는 선행 연구에서 사용한 어휘와 수준별 배치를 참고하였으며, 기술어들이 서로 상충되거나 수준 배열의 오류가 없도록 구성하였다. 척도별 채점 기준은 리커트 척도를 따라 1~5점마다 기술하였으며, 0점은 응답 내용이 과제와 무관하거나 제시한 문제나 자료를 따라서 읽기만 한 경우에 부여할 수 있는 것으로 하였다.

25) 본 연구에서 사용한 전체 채점 척도는 <부록 2>에 제시하였다.

<표 Ⅲ-6> '전반적 수행 능력'에 대한 채점 기준

점수 (수준)	전반적 수행
5 탁월	• 응답에 사소한 실수가 있지만 과제를 충족하며, 적절한 세부 설명을 포함함. • 표현의 실수가 거의 없고, 잘 이해할 수 있으며, 담화의 응집성이 있음.
4 우수	• 응답이 완벽하지는 않지만 과제를 상당히 적절하게 다루고 있음. • 표현의 몇 가지 실수가 있지만 명료하고 유창하게 응집성 있는 담화를 구성함.
3 보통	• 응답에서 과제와 관련하여 어느 정도 적절한 내용을 다루고 있음. • 표현에 실수가 지속적으로 나타나고, 유창성이 다소 부족하지만 이해가능한 편임.
2 미흡	• 응답에서 과제를 다루고 있지만, 제한적으로 주제를 다루고 있음. • 표현을 이해할 수는 있으나, 전반적으로 발화 전달과 담화 응집성에 문제가 있음.
1 부족	• 응답이 과제와 거의 관련이 없음. • 표현을 대체로 이해하기 어렵고, 응집성이 매우 부족함.
0 채점 불가	• 응답한 내용이 과제와 전혀 관련이 없음. • 다른 응답 내용이 없이 문제나 자료의 내용을 그대로 읽기만 하였음.

작성한 채점 기준의 유용성을 확보하기 위하여 실사용자로부터 피드백을 받을 필요가 있다고 판단하여 채점자들에게 연습 채점 과정에서 피드백을 요청하였다(Wolfe et al., 1998; Joe, Harmes, & Hickerson, 2011). 채점자들은 제공한 채점 기준을 사용할 때에 수험자 응답에 대한 수준별 채점에 문제가 없고, 이해 가능하다고 하였으며, 이에 구성한 채점 기준을 본 채점에도 사용하였다.

(2) 채점 방법의 선정

본 연구에서는 연구의 목적과 앞서 선정한 평가 준거를 고려하여 채점 방법으로 총체적 채점과 분석적 채점을 결합한 '혼합형 채점'을 선정하였다. 말하기 평가의 채점 방법은 평가 문항과 구인의 특성을 고려하여 선정하는 것이 일반적인데, 평가하는 능력을 하나로 볼 것인지, 아니면 여러 개로 나누어서 볼 것인지에 따라서 총체적 채점과 분석적 채점으로 나누기도 하고[26], 과제 특

26) 총체적 채점은 학습자의 수행으로부터 학습자 능력을 종합적으로 추론하여 하나의 점수 또는 등급을 부여하는 것을 가리킨다. 총체적 채점은 채점 효율성과 결론의 명료성

성을 기준으로 나눌 때는 주 특성 채점과 복합 특성 채점으로 구분하기도 한다(Hamp-Lyons, 1991; Fulcher, 2003). 이 중에서 '분석적 채점(analytic scoring)'은 수험자의 응답을 여러 개의 준거로 분리하여 채점하는 방법을 가리키는데, 구성형 응답 문항에서 수험자의 말하기 능력을 구성하는 여러 특성에 관한 개별적인 정보를 얻을 수 있다는 장점이 있다는 점에서 성취도 평가에 적합한 채점 방법이라고 판단된다. 그런데 분석적 채점은 채점자가 평가하는 각 특성을 완벽하게 분리하여 평가하기가 어렵고, 평가 결과에서 종합적으로 수험자 수준의 변별이 모호하다는 한계가 있다(이영식, 2006). 이러한 점을 보완하여 본 연구에서는 전반적 수행 능력을 중심으로 하는 총체적 채점과 그 밖의 언어적인 능력에 관한 분석적 채점을 병행하는 혼합형 채점 방법을 선택하였다.

연구에 참여한 모든 채점자들은 혼합형 채점 방법을 적용하여 본 채점 대상인 12명의 수험자의 응답을 모두 채점하는 '완전 교차형' 채점을 실시하였다. 이는 채점자 영향과 채점 과정 분석을 위한 전체 채점자의 비교 가능한 자료 확보와, 교실에서 이루어지는 말하기 성취도 평가의 형식적인 유사성을 고려하한 것이었다([그림 Ⅲ-2] 참조)[27]. 채점자의 수험자 응답 청취는 말하기 성취도 평가로서의 실제성을 고려하여 1회로 제한하였고, 불가피한 문제가 발생한 경우를 제외하고는 재청취를 불허하였다.[28]

등의 장점을 갖고 있지만, 평가 결과에 대한 의미를 정확히 파악하기가 어렵고, 인상에 의한 평가라는 비판을 받기도 한다. 분석적 채점은 학습자 수행으로부터 학습자의 능력을 세분하여 영역별로 점수를 부여하는 방법이며, 채점자 수행과 잠재된 능력에 관한 상세한 정보를 제공할 수 있다는 장점이 있다. 하지만 분석적 채점은 효율이 낮고, 경제적으로 부담이 큰 평가 방식이며, 채점자가 평가 준거를 구별하지 못할 수 있다는 약점을 갖고 있다.

27) 완전 교차형 설계는 신뢰도 확보가 용이하다는 장점이 있지만, 경제성과 효율성이 떨어져 '고급 채점 체재'로 여겨진다(Myford & Wolfe, 2004: 198). 본 연구에서는 채점자 영향에 대한 양적 분석 결과의 비교 가능성 확보를 위하여 완전 교차형으로 채점을 설계하였다.

28) 채점자들은 수험자 응답을 녹음한 음성 자료를 재청취한 경우 점수 기록표에 재청취 횟수를 기록하였다. 전체 13명의 채점자 가운데 재청취를 한 경우는 R02 1회(E12), R04 1회(E02), R07 1회(E10), R08 1회(E05)였으며, 청취를 중단했던 해당 문항에 대해서 재청취가 이루어졌음을 보고하였다.

수험자 / 채점자	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	E11	E12	문항수
R01	할당	할당	할당	할당	할당	할당	할당	할당	할당	할당	할당	할당	48
R02	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R03	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R04	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R05	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R06	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R07	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R08	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R09	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R10	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R11	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R12	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48
R13	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	48

[그림 Ⅲ-2] 완전 교차형 채점 설계도

2. 말하기 평가 결과의 채점자 영향 분석

1) 채점자 영향 분석의 설계

(1) 평가 결과 자료의 수집 과정

말하기 평가의 결과는 채점자가 수험자의 각 문항별 응답을 듣고, 채점 척도를 바탕으로 부여한 점수를 가리키며, 본 연구에서는 평가 결과 자료 수집을 위해 채점자들에게 수험자들의 응답이 녹음된 음성 자료와 엑셀 파일로 된 점수 기록표를 제공하였다.

평가 결과 자료의 수집은 연습 채점과 본 채점 과정에서 이루어졌다. 연습 채점은 '한국어 말하기 능력 시험'의 평가 문항과 채점 척도 및 채점 규칙에 대한 적응을 위하여 실시하였다. 연습 채점은 3명의 수험자가 응답한 12개의 음성 자료를 바탕으로 이루어졌다. 채점자들은 채점을 마친 후에 엑셀 파일에 점수를 입력하여 제출하였다. 연습 채점 결과, 채점자 간 내적 일관성 신뢰도를 의미하는 Cronbach α가 .89로 우수한 수준으로 나타났으며, 이에 채점자들 간에 평가 도구에 대한 일관된 이해가 마련되었다고 판단하였다.

다음으로 본 채점에서 13명의 채점자는 무선적으로 배열한 12명의 수험자가 4개 문항에 대하여 응답한 48개의 음성 자료[29]를 듣고, 점수 기록표에 평가 결과를 입력하여 제출하였다.

(2) 평가 결과 자료의 분석 설계

본 연구에서는 말하기 평가의 채점자 영향을 파악하기 위하여 평가 결과에 대한 CTT 분석과 IRT 분석을 실시하였다. 관찰 점수를 바탕으로 이루어지는 CTT 분석에서는 기술 통계 분석과 문항 양호도 분석, 신뢰도 분석, 배경변인 분석을 실시하였다. 다음으로 척도화된 점수를 사용하는 문항반응이론 분석에서는 비교 가능한 측정값을 바탕으로 채점자 영향과 그 의미를 파악하고자 하였다.

① 관찰 점수 기반 접근

채점자 영향의 파악을 위하여 먼저 관찰 점수를 기반으로 하는 고점검사이론 분석을 실시하였다. 분석 자료는 채점자들은 12명의 수험자가 4개 평가 문항에 응답한 것을 듣고 채점 척도를 바탕으로 점수를 부여한 것이다. CTT에 따른 분석은 기술 통계를 기본으로 평가 문항의 양호도와 채점 신뢰도, 배경변인 분석으로 이루어졌다.

먼저 채점 자료의 기본적인 특성을 파악하기 위하여 기술 통계 분석을 실시하였다. 기술 통계 분석은 관찰 점수의 평균과 표준 편차 및 최대·최솟값과 첨도와 왜도를 확인하는 것으로 이루어졌다. 문항 양호도 분석은 수험자들이 응시한 말하기 평가 문항의 품질에 관한 종합적인 정보를 확인하기 위하여 실시하였다. '문항 양호도(item quality)'는 문항이 학습자의 능력을 적절하게 평가하는가에 관한 정보로서, 본 연구에서는 수행 평가 특성에 따라 문항 곤란도와

29) 채점에 사용한 자료는 중·고급 한국어 학습자들이 '한국어 말하기 능력 시험'에서 응답한 것을 녹음한 것이다. 수험자들은 약 20분간 말하기 평가에 응시하였으며, Windows(v. 10)에 내장된 '음성 녹음기'를 통해 녹음하였다. 녹음한 응답 자료는 음향적인 최적화를 위하여 Goldwave(v5.70)로 편집한 후에 채점용 응답 자료로 제공하였다.

변별도를 살펴보았다.[30] 다음으로 신뢰도 분석은 채점자들 간에 채점의 일관성에 관한 정보인데, 이와 관련하여 관찰 점수의 분산과 관찰 점수에서 오차를 뺀 진 점수 분산의 비로 산출하는 크론바흐 알파(Cronbach α)를 확인하였다. 그리고 채점자 간 일치도를 알아보기 위하여 상관 계수를 확인하였다. 끝으로 분산 분석을 통해 관찰 점수에 나타난 수험자의 특성 및 채점자 경력, 전공, 외국어 능력, 평가 훈련 등에 따른 집단별 차이를 알아보았다(<표 Ⅲ-7> 참조).

<표 Ⅲ-7> 관찰 점수 기반 평가 결과 분석 방법

분석 방법	측정 항목
기술 통계	총점, 평균, 표준편차, 최댓값, 최솟값, 첨도, 왜도
문항 양호도	문항 곤란도, 문항 변별도
신뢰도	채점자 간 신뢰도(점수 신뢰도), 상관 분석
배경 변인	수험자 모국어, 경력, 외국어 능력, 평가 교육 경험

② 척도 점수 기반 접근

관찰 점수에 의한 CTT 분석은 접근이 간단하고 결과가 명료하다는 장점을 갖고 있으나, 점수가 모두 동일한 오차를 갖고 있는 것으로 가정하기 때문에 정확한 능력 추정을 할 수 없다는 한계를 갖고 있다. 이러한 관찰 점수 기반 접근의 한계를 극복하기 위하여 본 연구에서는 문항반응이론 분석을 통해 척도 점수를 바탕으로 측정값의 객관화와 비교 가능성을 확보하고자 하였다.

IRT는 CTT에서 고려하지 않는 문항 양호도와 피험자 능력 수준과 같은 '잠재적 특성(latent trait)'에 대하여 확률 모형으로 접근한다는 점에서 차별점을 갖고 있다. CTT에서는 각 문항이나 채점 영역의 특성을 고려하지 않는 관찰점수와 총점을 바탕으로 접근을 하는 것과 달리 IRT에서는 총점을 학습자 능력 추정을 위한 정보로 취급하며, 학습자와 문항을 같은 척도 상에 위치하도록 하여 비교 가능성을 확보할 수 있다는 장점을 갖고 있다(Thissen & Orlando,

30) 고전검사이론에서 문항 곤란도는 학습자가 각 문항에서 획득한 점수의 평균으로 산출하며, 수치가 높을수록 쉬운 문항이다. 문항 변별도는 각 문항이 학습자의 능력 수준을 잘 구별해 주는 기능을 갖고 있는가에 관한 정보이며, 고전검사이론에서는 문항별 점수와 총점의 상관을 통해 산출한다. 추측도는 학습자의 능력과 무관하게 점수를 획득할 수 있는 확률에 대한 것이며, 선택형 문항에서 능력과 무관한 점수를 획득할 확률을 알아보는 것으로, 본 연구에서는 수행 평가를 기반으로 하기 때문에 다루지 않았다.

2001). IRT는 문항 양호도 모수 숫자에 따라 1~3모수 모형으로 나눌 수 있으며, 각각의 모형들은 각 검사별로 적합성을 따져서 적용해야 한다(곤란도를 기준으로 하는 1모수 모형, 곤란도와 변별도를 함께 고려하는 2모수 모형, 추측도까지 고려하는 3모수 모형). 초기 IRT 모형은 정오 판단이 분명한 이분반응 모형으로 적용되었으나, 주관식 문항에서 부분 점수를 부여하는 것과 관련된 다분 문항반응이론(polytomous IRT)이 개발되면서 이론 적용의 폭을 확장하게 되었다.

본 연구에서는 IRT에서 곤란도 모수를 중심으로 하는 1모수 라쉬모형을 바탕으로 채점자의 채점자 영향을 분석하였다. 라쉬모형은 변별도 모수를 1로 고정한다는 특징을 갖고 있는데, IRT 모형 가운데 산출 결과의 명료성을 장점으로 한다. 2모수 모형이나 3모수 모형을 고려하지 않은 까닭은 본 연구의 목적이 채점자의 채점자 영향에 관한 것이며, 수험자에게 민감한 변별도 정보와 달리 채점자의 채점자 영향에는 과제의 곤란도가 미치는 영향이 절대적이기 때문이다. 또한 3모수 모형은 객관식 문항에 대한 접근에 유효한 추측도 정보를 고려하고 있으므로, 본 연구에서는 채점자 영향 분석을 위하여 라쉬분석모형 가운데 채점 과정에 관여하는 과제, 평가 준거 등의 국면을 고려하는 MFRM을 선택하였다.

MFRM은 채점자 영향으로 인한 영향이 수행 판정에 미치는 영향을 알아보고, 공정하고 타당한 측정 결과 제공 및 해석을 목표로 하며, 관찰 점수로부터 보다 정확하고 적합하며 통계적 한계로부터 자유로운 객관적 측정을 지향한다(Linacre, 1994: iii). MFRM 모형에서 고려하는 '국면(Facets)'이란 평가에서 판정 상황과 관련이 있는 채점자, 문항, 과제, 수험자 등의 다양한 측면을 가리키며, 평가의 목적에 따른 정확하고 유용한 판정 결과를 얻기 위하여서 여러 국면에 대한 반응으로부터 구별하여 수험자의 능력 추정이 이루어져야 한다(Linacre, 1989: 20). Facets 프로그램(Linacre, 2019)은 이상의 관점을 바탕으로 MFRM 모형을 적용하여 수험자의 능력 수준, 문항의 곤란도, 채점자의 엄격성, 채점 척도 구조 등의 평가 결과 산출에 직접적으로 관여하는 요소들에 대한 추정치를 확인하고, 채점자 효과에 대한 객관적인 접근을 가능하게 한다. Facets을 활용하여 언어 능력 평가에서 채점자 영향을 구체적으로 분석한 미포드와 울프(Myford & Wolfe, 2003, 2004)는 Facets에서 추정하는 채점자 영향으

로 채점 경향(rater severity/leniency), 집중 경향(central tendency), 무작위성(randomness), 후광성(halo), 차별적인 채점자 엄격성/관대성(differentiate severity/leniency)을 언급하면서, 각 영역에서 얻을 수 있는 채점자 영향에 관한 구체적인 정보를 제시하였다. 본 연구에서는 Facets 분석 결과를 따라 채점자 영향을 집단 특성과 개인 특성으로 구분하고, 엄격하거나 관대한 채점을 하는 채점 경향성과, 집중 경향성과 무작위성, 후광성, 편향성 등을 살펴본다 (<표 Ⅲ-8> 참조).[31]

<표 Ⅲ-8> 문항반응이론을 통한 채점자 영향 분석 체제

분석 대상		채점 경향성	집중 경향성	무작위성	후광성	상호작용 편향성
검증 방법	집단	채점 척도 사용 빈도, 고정된 카이제곱 검정, 채점자 분리비(분리 신뢰도)	채점 척도 사용 빈도, 고정된 카이제곱 검정, 채점자 분리비(분리 신뢰도)	고정된 카이제곱검정, 채점자 분리비 (분리 신뢰도)	고정된 카이제곱검정, 채점자 분리비 (분리 신뢰도)	-
	개인	채점자별 엄격성 분포 및 측정값, 능력 특성 척도별 빈도,	특성 범주별 채점 빈도, 채점 적합도, 채점 척도,	채점 적합도, 채점자 간 신뢰도	채점자×평가 준거의 편향-상호작용 분석	채점자×과제×준거에 따른 편향-상호작용 분석

<표 Ⅲ-8>은 Facets 분석을 통해 채점자 집단 및 채점자 개인의 영향을 파악하여 검증할 수 있는 채점자 영향의 유형과 분석을 위한 측정값을 정리한 것이다. 먼저 전반적인 채점 경향의 차이의 검증은 카이제곱 값과 자유도 및 유의도, 채점자 집단의 분리비(separation ratio)와 분리지수의 신뢰도로 이루어지며, 채점자 내 신뢰도와 채점자 간 신뢰도에 관련된 평정 일치도(exact agreements)와 기대 일치도(expected agreements)에 관한 정보도 이용할 수 있다. 집중 경향은 채점자들이 척도의 특정 점수를 집중적으로 사용하는 경향으로, 주로는 중앙값 주변으로 채점 범주를 제한하는 것을 말한다. 집중 경향성은 채

31) 차별적인 엄격성에 관한 분석은 피험자의 집단에 따른 채점의 영향을 고려하려는 것이며, 본 연구에서는 채점자의 채점자 영향 연구에 초점을 두고 있기 때문에, 피험자 특성의 영향을 고려하는 차별성 변수는 결과로 제시하지 않기로 하였다.

점자들이 척도의 차이를 이해하지 못하고 있다는 의미이거나 또는 평가 구인을 잘 이해하지 못하고 있기 때문에 나타날 수 있으며(Myford & Wolfe, 2004: 198), 따라서 채점자들이 수험자의 수행 능력 수준을 잘 구별할 수 없음을 나타낸 것으로 볼 수 있다. 이와 관련하여 평가 결과의 부담이 큰 고부담 평가 혹은 평가 결과를 비교 받는다는 심리적 부담이 있는 상황에서의 평가에서는 위험 부담을 줄이기 위하여 척도를 제한적으로 사용하면서 집중 경향성이 나타날 수 있으므로 평가 상황을 고려하여 해석할 필요가 있다(Myford & Mislevy, 1995; Wolfe, Chiu, and Myford, 1999). 다음으로 채점자 척도 사용의 비일관성에 관한 무작위성은 채점 척도 사용에 능숙하지 않은 채점자에게서 나타날 수 있다. 이는 Facets 분석 결과에서 수험자 국면에서의 고정된 카이제곱 검정의 유의도와 분리비, 분리 신뢰도를 통해 확인할 수 있으며, 개별 채점자의 내·외적합도 지수가 1보다 훨씬 크게 나타나는 경우, 채점자 간 신뢰도 분석 결과를 통해서도 확인할 수 있다. 후광성(halo effect)은 채점자가 어떤 특성을 평가할 때 다른 특성을 평가한 결과가 영향을 미치는 것을 말한다(신동일·장소영, 2004). 후광성의 영향을 확인하는 방법은 Facets 분석에서 문항 곤란도 모수를 고정한 후에 평가 준거별로 유사한 경향을 나타내는 것을 찾는 방법을 이용한다. 끝으로 Facets에서는 채점자가 특정 집단이나 평가 특성에 편향적인 채점을 하였는지에 관한 정보를 제공하며 이는 채점자와 과제, 평가 준거의 상호작용을 분석하여 파악할 수 있다.

2) 채점자 영향에 대한 통계 분석 결과

(1) 관찰 점수 기반 분석 결과

① 평가 결과 점수 분포의 적절성

언어 평가에서의 기술 통계 분석에서는 문항별 평균과 표준편차, 최대・최솟값에 관한 기본적인 정보와 채점 자료의 정규성 확인을 통해 점수 분포의 안정성을 확인한다(Purpura et al., 2015). ‘한국어 말하기 능력 시험’의 평가 결과에 대한 기술통계는 평가 결과로 부여한 점수에 관한 종합적인 정보로서 사례

수, 최대-최솟값, 평균-표준편차, 첨도와 왜도 등을 제공하며, 이를 바탕으로 채점자들이 결정한 점수(관찰 점수)의 전반적인 경향성과 다양성, 분포 특성을 알 수 있다.

<표 Ⅲ-9>로 제시한 기술 통계 분석 결과, 13명의 채점자는 12명의 수험자의 응답에 대해 채점하였으며, 결측은 없었다. 문항 중에서 2번, 3번 문항에서는 최솟값으로 0점을 받은 경우가 있었으며, 이는 과제 수행의 요건을 갖추지 못하여 점수를 부여할 수 없는 경우에 해당한다. 각 문항의 평가 준거별 평균값을 보면, 3번 문항의 전반적 수행 평균 점수가 2.76으로 가장 낮았고, 1번 문항의 발음 구사 능력에 대한 평균이 3.90점으로 가장 높았다. 그리고 2번과 3번 문항에서는 1번과 4번 문항에 비해 전체 특성에서 표준편차가 큰 편이었다. 자료의 첨도와 왜도를 통하여 정규성을 확인한 결과, 전체 문항 및 평가 준거의 분포가 ±1 사이에 위치하여 정규성 조건을 충족하는 것으로 나타났다 (Bachman, 2004: 74).

<표 Ⅲ-9> 기술 통계

문항	평가 준거	사례수	최대	최소	평균	표준편차	첨도	왜도
1	전반적	156	5.00	1.00	3.86	1.05	-0.49	-0.70
	발음	156	5.00	1.00	3.90	0.95	-0.44	-0.55
	어휘·문법	156	5.00	1.00	3.72	1.04	-0.44	-0.53
	담화	156	5.00	1.00	3.85	1.00	-0.51	-0.45
2	전반적	156	5.00	0.00	3.20	1.38	-0.67	-0.20
	발음	156	5.00	0.00	3.40	1.30	-0.82	0.05
	어휘·문법	156	5.00	0.00	3.30	1.40	-0.68	-0.26
	담화	156	5.00	0.00	3.18	1.35	-0.69	-0.16
3	전반적	156	5.00	0.00	2.76	1.25	-0.02	-0.50
	발음	156	5.00	0.00	3.03	1.27	-0.22	-0.85
	어휘·문법	156	5.00	0.00	2.88	1.29	-0.02	-1.00
	담화	156	5.00	0.00	2.85	1.26	0.01	-0.75
4	전반적	156	5.00	1.00	2.85	1.13	0.25	-0.67
	발음	156	5.00	1.00	2.88	1.19	-0.02	-0.84
	어휘·문법	156	5.00	1.00	2.97	1.20	0.15	-0.92
	담화	156	5.00	1.00	2.95	1.20	0.08	-0.94

기술 통계 분석을 통하여 전체 평가 결과 분포의 특징을 확인한 결과, 수험자의 말하기 문항과 평가 준거에 따른 문항별 점수 분포(첨도와 왜도)가 정규성을 충족하는 것으로 나타났다. 문항별로 나타난 특징을 살펴보면, 문항 1 ~ 3의 첨도에서 문항3의 담화 구성 능력을 제외하고 모두 음수로 나타났으며, 이는 중앙값을 기준으로 우측으로 치우친 분포가 나타나고 있음을 말한다. 4번 문항에서는 발음을 제외하고 모두 양수로 나타나 상대적으로 낮은 점수를 받은 학습자 분포가 많았던 것으로 나타났다. 첨도는 2번 문항의 발음 구사 능력을 제외하고 나머지 문항의 평가 준거에서 음수로 나타났는데, 이는 척도 상에서 중앙값에 집중된 경향이 나타났음을 의미한다. 평가 결과 자료의 종합적인 특징 및 채점자 영향과 맺는 상호작용은 MFRM 분석을 통하여 구체적으로 확인하고자 한다.

② 평가 문항 및 준거별 채점의 적합성

채점이 평가 문항 및 평가 준거와 관련하여 적합한 수준으로 이루어졌는가를 알아보기 위하여 CTT에 따른 문항 양호도 분석을 실시하였다. 문항 양호도(item quality)는 평가 문항이 평가의 목적에 맞는 측정을 하고 있는지에 관한 정보이며, 이와 관련하여 구성형 응답 문항의 곤란도와 변별도 정보를 확인하였다. 문항 곤란도(item difficulty)는 선택형 문항에서는 관찰 점수의 정답률로 확인하며, 본 연구에서는 구성형 문항으로 평가가 이루어졌기 때문에 채점자들이 부여한 평균값으로 확인하였다. 문항 변별도(item discrimination)는 문항이 평가하고자 하는 능력 수준을 잘 변별하고 있는가에 대한 정보이며 각 문항의 수험자 점수와 총점 간의 상관을 통해 확인하였다.

<표 Ⅲ-10> 고전검사이론에 의한 문항 양호도

문항	평가 준거	사례수	총점	곤란도	변별도
1	전반적	156	602	3.86	0.96
	발음	156	609	3.90	0.91
	어휘·문법	156	581	3.72	0.96
	담화	156	601	3.85	0.94
	평균	156.00	598.25	3.83(1.01)	0.94
2	전반적	156	499	3.20	0.96
	발음	156	530	3.40	0.92
	어휘·문법	156	515	3.30	0.96
	담화	156	496	3.18	0.96
	평균	156.00	510.00	3.27(1.36)	0.95
3	전반적	156	431	2.76	0.96
	발음	156	472	3.03	0.92
	어휘·문법	156	450	2.88	0.95
	담화	156	445	2.85	0.96
	평균	156.00	449.50	2.88(1.27)	0.95
4	전반적	156	444	2.85	0.96
	발음	156	449	2.88	0.93
	어휘·문법	156	463	2.97	0.96
	담화	156	460	2.95	0.97
	평균	156.00	454.00	2.91(1.18)	0.96

<표 Ⅲ-10>으로 제시한 문항 양호도 분석 결과에서 전체 문항 중 곤란도가 가장 높은 문항은 3번 문항이었다(평균=2.88, 표준편차=1.27). 다음으로는 4번 문항(평균=2.91, 표준편차=1.18), 2번 문항(평균=3.27, 표준편차=1.36), 1번 문항(평균=3.83, 표준편차=1.01)순으로 곤란도가 높게 나타났다. 문항 변별도는 4번 문항(ID=.96)이 가장 높았으며, 가장 낮은 것은 1번 문항(ID=.94)이었는데, 전체 문항 평균 변별도가 0.95로 나타나 평가 전반에서 능력 특성과 총점의 상관이 매우 높은 것으로 나타났다.

문항 변별도 분석 결과에서 .9 이상의 높은 수치가 나타난 것은 기본적으로 '한국어 말하기 능력 시험'의 모든 평가 문항과 평가 준거가 적절한 기능을 하였다는 의미로 해석할 수 있다. 그러나 문항 및 준거 점수와 총점의 상관으로 확인한 통계적 변별도가 평가 문항이 수험자 집단에 따라 편향적으로 작용하였을 가능성(Wilson & Masters, 1993)이 있으므로, 이와 관련하여 MFRM 분석을 통해 각 문항 및 준거 점수에 나타난 채점자의 편향을 확인할 필요가 있다.

③ 채점자 신뢰도의 일관성

본 연구의 채점 설계는 완전 교차형이었으며, 모든 채점자가 모든 본 채점 대상 수험자 응답을 결측 없이 채점하였다. 본 연구에서 살펴본 신뢰도는 말하기 평가 결과로 부여한 점수가 일치하는 경향인 내적 일관성 신뢰도(Cronbach's α)(수식 (1.2) 참조)와 채점자 간 신뢰도(inter-rater reliability)이다. 채점자 간 신뢰도는 여러 채점자가 독립적으로 부여한 점수가 어느 정도로 일치하고 있는지를 알아보는 것으로, pearson의 상관계수법을 사용한다. 이와 관련하여 채점에 참여한 13명의 채점자의 평가 결과에 대한 상관 행렬 분석을 실시하였다.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma S_i^2}{S_x^2}\right) \qquad (1.2)$$

α : $Cronbach$의 내적일치도 계수
n : 평가문항의 수
S_x^2 : 전체 평가결과 점수의 분산
S_i^2 : 개별 문항들의 분산

분석 결과 전체 결과의 동질성을 의미하는 내적 일관성 신뢰도인 크론바흐 알파는 0.969로 매우 높게 나타났다. 채점자별로 상관행렬을 통해 신뢰도를 비교하였을 때는 R02과 R13 사이의 상관(.467)이 가장 낮았으며, 전체적으로는 R02와 R13이 다른 채점자들에 비해 상대적으로 낮은 상관성이 나타났다(<표 Ⅲ-11> 참조).

<표 Ⅲ-11> 채점자 간 총점의 상관 행렬

	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R13
R01	1.000												
R02	.586	1.000											
R03	.747	.675	1.000										
R04	.729	.675	.745	1.000									
R05	.725	.684	.750	.810	1.000								
R06	.766	.716	.872	.831	.815	1.000							
R07	.710	.638	.793	.780	.697	.802	1.000						
R08	.663	.550	.660	.754	.709	.695	.610	1.000					
R09	.674	.532	.688	.763	.738	.742	.721	.624	1.000				
R10	.820	.578	.771	.774	.722	.800	.804	.637	.708	1.000			
R11	.749	.616	.766	.762	.696	.780	.724	.668	.689	.778	1.000		
R12	.765	.618	.768	.802	.746	.796	.740	.706	.638	.717	.686	1.000	
R13	.688	.467	.674	.667	.643	.711	.749	.534	.648	.710	.673	.630	1.000

신뢰도 분석 결과에서 내적 일관성 신뢰도가 매우 높게 나타난 점은 신중한 해석을 필요로 하는데, 채점자들이 무엇을 합의하였으며 무엇에 동의한 것인지가 불분명하기 때문이다(Reed & Cohen, 2001). 이와 관련하여 채점 척도 사이의 강한 상관으로 인하여 다중공선성(multicollinearity)이 나타났을 가능성이 있다. 이는 말하기 평가 채점 과정에서 채점자들이 평가 준거들을 잘 변별하지 못하는 가운데 중복 측정 또는 가중 채점하면서 나타날 수 있다. 또한 척도 사용에서도 응답 수준 구별을 포괄적으로 하거나, 점수 결정에 대한 부담을 줄이기 위하여 척도의 중간 점수를 부여하는 경향이 나타나 높은 신뢰도가 나타났을 것으로 판단된다.

이와 관련하여 MFRM 분석에서 채점자 영향 가운데 평가 준거 및 척도 구별의 문제에 대한 비교 가능한 채점자 영향을 비교할 필요가 있다. 또한 채점자 간 신뢰도가 낮은 경향을 나타낸 R02와 R13에 대해서는 신뢰도를 저하시킨 원인에 대한 탐색이 필요하며, 구체적으로 특정 문항이나 평가 준거와 관련하여 채점 과정에서 어떤 상호작용 영향이 있었는지를 확인해야 한다.

④ 수험자 및 채점자 특성에 따른 변별성

학습자 배경 변인이나 채점자 배경 변인이 평가 과정에서 상호작용할 수 있기 때문에 배경 변인 영향에 관한 일변량 분산분석(One way ANOVA)을 실시하였다. 선행 연구를 바탕으로 학습자 변인으로는 '한국어 수준'과 '국적'[32], 채점자 변인으로는 '성별', '학력', '전공', '경력', '평가 교육', '외국어 능력'을 고려하여 분석하였다(<표 15> 참조).[33]

<표 Ⅲ-12> 수험자 배경 변인에 따른 분산 분석

구분		사례수	평균	표준편차	평균비교	
					검정통계량 (유의확률)	사후비교
전체		156	12.90	4.13	-	-
한국어 수준	4	39	8.96	3.68	63.677 (.000)	4 < 5 < 6 < 7
	5	39	11.13	2.53		
	6	39	14.56	2.69		
	7	39	16.94	1.89		
국적	일본	78	12.88	4.41	5.085 (.007)	기타 < 중국
	중국	26	14.96	2.34		
	기타	52	11.88	4.07		

수험자 배경 변인 분석 결과, 구성형 말하기 평가 문항의 평가 결과에서 학습자들의 한국어 수준에 따라 유의한 차이가 나타났으며, 이는 유의 수준에서 세 가지 사후 검정법(Tuckey HSD, Duncan, Scheffe)을 모두 만족하였다. 학습자 국적에 따른 차이에서는 사후 검정에서 중국 수험자와 기타 국적 수험자의 차이가 유의한 것으로 나타났다(<표 Ⅲ-13> 참조).

32) 학습자 성별도 배경 변인으로 조사하였으나, 남학생이 1명뿐이었기 때문에 분석 결과를 함께 제시하지 않았다. 분산 분석 결과에서는 성별에 따른 차이가 유의한 것으로 나타났다(F=14.844(.000)).

33) 채점자들이 부여하는 점수의 가변성과 관련하여 쓰기 평가 선행 연구에서 나타난 주장은 구인과 무관하게 평가자의 언어적 배경이 영향을 미친다는 것이었다(Cumming, Kantor, & Powers, 2001: 68).

<표 Ⅲ-13> 채점자 배경 변인에 따른 분산 분석

구분		채점 사 례 수	평균	표 준 편차	평균비교	
					검정통계량 (유의확률)	사후비교
전체		156	12.90	4.13	-	-
성별	남	24	13.00	3.97	0.18 (.894)	-
	여	132	12.88	4.17		
학력	석사 졸업	84	12.84	4.16	1.378 (.255)	-
	박사 수료	60	13.32	4.18		
	박사 졸업	12	11.17	3.39		
전공	한국어교육	132	12.94	4.23	.079 (.779)	-
	기타	24	12.68	3.56		
경력	5~7년	60	13.37	4.15	.751 (.474)	-
	8~10년	48	12.81	4.27		
	11년 이상	48	12.40	3.96		
평가 교육	없음	24	14.85	3.50	2.383 (.072)	0 < 1 < 2
	1회	36	12.38	4.30		
	2회	60	12.37	4.29		
	3회	36	12.99	3.80		
외국어 능력	1개 외국어	72	13.02	4.06	.390 (.678)	-
	2개 외국어	72	12.64	4.25		
	3개 외국어	12	13,69	3.93		

채점자 배경 변인 분석 결과에서 유의한 차이가 있는 변수는 나타나지 않았다. 다만 평가 교육과 관련하여서는 덜 엄격한 사후검정법(LSD)을 사용하였을 때에만 교육을 받지 않은 채점자 집단과 1회 받은 집단, 그리고 2회 받은 집단 간에 유의한 차이를 확인할 수 있었다(<표 Ⅲ-13> 참조). 이러한 결과는 말하기 평가에 관한 채점자 교육의 긍정적인 영향으로 볼 수도 있을 것이다. 다만 연구에 참여한 분석 대상 채점자 수가 적기 때문에 이러한 결과를 일반화시킬 수 없으므로, 이와 관련하여 추가적인 연구가 이루어져야 할 것이다.[34]

이상의 관찰점수를 바탕으로 하는 고전검사이론 분석을 통해 확인한 구성형 말하기 평가 결과에 나타난 채점의 특성은 다음과 같다.

34) 특히 채점자들의 평가 교육에 관한 분산 분석에서 관대한 검정법을 적용하였으므로 참여한 평가 교육의 빈도와 품질을 구별하여 해석이 이루어져야 할 것이다.

첫째, 채점자들이 부여한 점수는 정규성을 충족하고 있었으며, 일부 문항과 준거에서 중앙값에 집중된 경향(platykurtic)이 나타났다.

둘째, 구성형 말하기 평가의 문항 양호도 분석에서 곤란도가 3.22(SD=1.20)로 다소 쉬운 수준으로 나타났으며, 변별도의 평균은 0.95로 매우 높게 나타나 평가 문항과 과제가 적절하게 기능하였음을 확인하였다. 그러나 관찰 점수에 의한 문항 양호도 분석에서 고려하지 않는 오차 변량으로 인하여 채점자의 편향성이 개입할 수 있다.

셋째, 말하기 평가의 내적일관성 신뢰도 분석 결과에서 높은 신뢰도(α=.97) 수준으로 나타난 것은 과대 추정된 신뢰도의 원인이 된 채점 사례수 외에 채점 척도를 중복하여 고려할 때 개입할 수 있는 다중공선성이 나타났을 가능성이 있다.

넷째, 평가 결과에 영향을 준 수험자 및 채점자 배경 변인 분석에서 '한국어 말하기 능력 시험'이 수험자의 한국어 능력 수준 변인(4 < 5 < 6 < 7)과 국적 변인<기타 < 중국)을 변별하는 기능을 확인하였으며, 채점자의 평가 관련 교육 경험(0 < 1< 2)에 따라서도 낮은 수준에서 차이가 있음을 확인하였다. 이와 관련하여 본 연구에서는 평가 결과를 도출하는 데 관여한 채점자 영향을 구체적으로 알아보기 위하여 MFRM 분석을 실시할 것이다.

(2) 척도 점수 기반 분석 결과

관찰 점수에 기초한 CTT 분석을 통하여 수집한 평가 결과 자료의 분포적 특징과 평가 도구, 채점자, 배경 변수에 따른 특징을 살펴보았다. IRT에 따른 MFRM 분석은 평가에 영향을 미치는 오차 변량을 통제하여 비교 가능성을 확보하고, 평가 국면별로 logit(log odds unit; logit) 척도 상의 위치 및 측정값을 제공하며, 이를 바탕으로 구체적인 채점자 영향에 관한 다양한 정보를 얻기 위하여 실시하였다(Eckes, 2005).

MFRM 분석은 전용 프로그램인 Facets(Linacre, 2019)으로 이루어졌는데, 분석을 위하여 평가 결과 자료를 프로그램에서 요구하는 형식에 맞추어 변환한 후에 확인하고자 하는 정보의 성격에 따라 분석 모형을 달리하여 구동하였다. Facets에서 다분 반응 문항에 대하여 선택할 수 있는 분석 모형은 채점척도모

형(Andrich, 1978)(rating scale model; RSM)과 부분점수모형(Masters, 1982)(partial credit model; PCM)이 있다. 채점척도모형은 문항 곤란도와 함께 채점 척도에서 점수를 획득할 확률을 고려하는 모형이며, 부분점수모형은 채점 척도모형에서 고려하지 않는 문항별 점수 획득 확률의 차이를 고려하는 것이다. 이와 관련하여 본 연구에서는 두 모형이 MFRM에서 고려하는 국면 및 상호작용 요소에 따라 다양한 하위 분석 모형을 구성한다는 점을 고려하여 분석 내용에 맞는 분석 모형의 코드를 입력하여 진행하였다.[35]

<표 Ⅲ-14> MFRM 분석의 세부 모형

분석 모형	하위 국면			Facets Code
	채점자	평가 문항	평가 준거	
공통	RSM	RSM	RSM	?, ?, ?, ?
채점자	PCM	RSM	RSM	?, #, ?, ?
문항	RSM	PCM	RSM	?, ?, #, ?
준거	RSM	RSM	PCM	?, ?, ?, #
채점자×문항	PCM	PCM	RSM	?, #, #, ?
채점자×준거	PCM	RSM	PCM	?, #, ?, #
채점자×문항×준거	PCM	PCM	PCM	?. #, #, #

본 연구는 <표 Ⅲ-14>에서 제시한 분석 모형 외에도 각 평가 국면에 따른 편향을 확인할 수 있는 편향-상호작용분석모형(코드: B)을 적용하여 말하기 평가 결과에 나타난 채점자 영향을 심층적으로 살펴보았다.

① 분석 국면별 종합적인 경향

MFRM의 종합적인 분석 결과를 평가 국면별 종합도를 통하여 확인하였다. 분석 대상 자료는 13명의 채점자가 학습자 12명이 4개 문항의 과제를 수행한 것에 대하여 4개의 평가 준거별로 점수를 부여한 것이다.

35) Facets에서 고려하는 평가 국면과 관련하여 다양한 단일 국면 분석 모형 및 융합 모형(Hybrid Model, Linacre, 2001)을 구성할 수 있다. 단일 국면 분석 모형을 선택할 경우에는 해당 국면에 따른 상세한 정보를 얻을 수 있으며, 융합 모형 분석에서는 국면 간의 상호작용성을 반영한 분석 결과를 얻을 수 있다는 장점이 있다. Facets 분석에서 단일 국면 모형과 융항 모형은 분석 결과에서 수치 정보의 차이가 있으며, 제공하는 관련 도표의 차이가 있으므로, 적절한 모형에 해당하는 코드를 입력하여 분석을 진행하여야 한다.

[그림 Ⅲ-3]에서 제시한 Facets의 기본 분석 모형인 RSM을 적용하여 전체 국면을 분석한 결과를 살펴보면, logit 분포에서 영향이 없음을 의미하는 0을 기준으로 응시한 12명의 학습자 중에 5명은 0 logit 보다 위에 위치하여 높은 능력 수준인 것으로 나타났으며, 7명은 0 logit 보다 아래에 위치하여 능력이 부족한 것으로 나타났다. 채점자 국면에 대한 영향 분석에서는 12명의 채점자가 0 logit 보다 아래에 위치하여 전반적으로 관대한 채점을 한 것으로 나타났다. R09는 0 logit으로 전반적인 평가 결과에서 채점자 영향이 없는 채점을 한 것으로 나타났고, R13은 이상치(outlier)로서 지나치게 관대한 채점을 한 것으로 나타났다. 과제에 따라서는 그래프를 해석하는 문항3과 기사를 읽고 문제에 의견을 말하는 문항4가 같은 logit 선상에 위치하였는데 상대적으로 어려웠던 것으로 나타났고, 반대로 가장 쉬운 것은 경험을 말하는 문항1이었다. 설정한 4개의 평가 준거들은 모두 0 logit에 근접하여 나타나 편차가 작은 것으로 나타났으며, 그 중 전반적 수행에 대한 점수가 0 logit 보다 조금 높은 곳에 위치하여 다른 기준들보다 점수를 받을 확률이 다소 낮은 것으로 나타났다. 반대로 발음은 0 logit보다 조금 아래 위치하였으며, 다른 기준들보다 좋은 점수를 받을 확률이 높은 것으로 나타났다. 그 밖에 채점 척도별 사용에 있어서는 0 logit에 해당하는 값이 2.5점에 해당하는 것으로 나타나, 수행을 하였을 때 받을 수 있는 1 ~ 5점의 중앙값과 일치하는 것으로 나타났으며, 극단 값인 0점과 5점은 ±3 logit 밖에 위치하는 것으로 나타나 받을 수 있는 확률이 매우 낮은 것으로 나타났다.

측정값 (logit)	분석 국면				
	학습자	채점자	평가 문항	평가 준거	평가 척도
	높은 능력	엄격함	어려움	낮은 점수	5점
3					
	E03				4
2					
	E04				
	E08				3.5
1	E10				
					3
			문항3 문항4		
	E12				
				전반적	
0		R09	문항2	담화 어·문	2.5
	E01 E11			발음	
	E02				
		R05			
	E07	R04			2
		R03 R11			
-1		R06	문항1		
	E06				
		R01			
		R10			1.5
		R07 R12			
-2		R02			
		R08			
	E09				
					1
		R13			
-3	E05				
	낮은 능력	관대함	쉬움	높은 점수	0점

[그림 Ⅲ-3] MFRM에 의한 전체 분석 국면의 척도 분포도

Facets을 통해 산출한 채점자 특성에 관한 상세한 내용을 이해하기 위하여

Facets에서 제공하는 15가지 종류의 정보가 무엇을 의미하는지 확인하였다. 먼저 총점(Total Score)은 채점자들이 부여한 모든 점수를 더한 값이며, 채점 횟수(Total Count)는 채점자들이 점수를 부여한 전체 횟수를 말한다. 관찰 평균(Observed Average)은 채점자별로 부여한 점수의 평균값을 말하며, 조정 평균(Fair average)은 채점자, 학습자, 과제가 표준적일 때의 경우를 가정하여 산출하는 값으로, 관찰값과의 차이를 비교하여 수집한 자료가 완전한지를 확인할 수 있는 수치이다. 측정값(measure)은 라쉬모형분석을 통해 채점자의 채점 경향성을 상대화하여 나타낸 것으로, 0 logit을 기준으로 멀리 떨어질수록 엄격하거나 관대한 것을 나타낸다(<표 Ⅲ-15> 참조).

<표 Ⅲ-15> 채점자별 MFRM 분석 결과표

채점자	총점	채점 횟수	관찰 평균	조정 평균	Rasch Model		내적합도		외적합도		변별도	상관분석		채점자 간 일치도 %	(R)채점자 간일치도 %
					측정값	오차	제곱 평균	표준 점수 (Z)	제곱 평균	표준 점수 (Z)		관찰 -측정	(R)관찰-측정		
R13	755	192	3.93	4.1	-2.71	0.11	1.82	6.1	1.48	3.3	0.32	0.66	0.76	32.4	31.8
R08	703	192	3.66	3.79	-2.14	0.1	1.14	1.3	1.22	1.8	0.87	0.75	0.78	33.5	35.4
R02	671	192	3.49	3.6	-1.81	0.1	1.35	3.1	1.46	3.8	0.53	0.69	0.79	35.3	36.7
R07	657	192	3.42	3.52	-1.67	0.1	1.04	0.4	1.04	0.4	0.93	0.79	0.79	39.8	37
R12	653	192	3.4	3.49	-1.63	0.1	0.87	-1.2	1.05	0.4	1.05	0.81	0.79	40.4	37
R10	634	192	3.3	3.38	-1.44	0.1	0.94	-0.5	0.94	-0.5	1.07	0.83	0.8	42.1	37.1
R01	626	192	3.26	3.34	-1.36	0.1	0.75	-2.6	0.79	-2.2	1.21	0.8	0.8	42	37
R06	596	192	3.1	3.17	-1.08	0.1	0.68	-3.5	0.68	-3.5	1.38	0.88	0.8	42.6	36.4
R03	591	192	3.08	3.14	-1.03	0.1	0.84	-1.6	0.84	-1.6	1.17	0.83	0.8	41	36.2
R11	591	192	3.08	3.14	-1.03	0.1	0.85	-1.5	0.82	-1.8	1.18	0.82	0.8	41.4	36.2
R04	557	192	2.9	2.94	-0.72	0.1	0.88	-1.1	0.9	-1	1.14	0.88	0.8	38.4	34.6
R05	536	192	2.79	2.83	-0.53	0.1	0.63	-4.3	0.67	-3.7	1.3	0.82	0.8	36.4	33.2
R09	477	192	2.48	2.48	0.00	0.09	1.16	1.5	1.15	1.5	0.85	0.77	0.8	30.9	28.1

모형, 모집단: 표준오차평균 .10 표준편차(조정값). .67 분리비 6.82 층위 9.43 분리신뢰도 .98
모형, 표본: 표준오차평균 .10 표준편차(조정값). .70 분리비 7.11 층위 9.81 분리신뢰도 .98

모형, 고정된 엄격성 카이제곱값: 596.9 자유도.: 12 유의도(확률): .00
모형, 무선적 엄격성 카이제곱값: 11.8 자유도.: 11 유의도(확률): .38

채점자 간 일치 빈도: 14976 실제 일치도: 5717 = 38.2% 기대 일치도: 5262.3 = 35.1%

② 관대한 채점 경향성의 영향

전체적인 채점 경향성을 파악하기 위하여 Facets 분석에서 채점자가 수험자의 과제 수행에 대하여 부여한 점수가 상대적으로 엄격하거나 관대한지를 알아보았다. 채점 경향성에 대한 집단 수준 분석에서는 척도별 빈도, 고정된 카이제곱 검증, 분리비와 신뢰도를 확인하며 개인 수준에서 logit 척도 상의 위치와 각 특성별 척도 판정 빈도, 엄격성 측정값, 이상치를 활용하였다(Myford & Wolfe, 2004: 217).

ⓐ 집단 수준 영향

먼저 전체 0~5점으로 이루어진 전체 척도에서의 사용 빈도를 살펴보면, 수행이 이루어졌을 때 부여하는 1~5점의 빈도 2457회 중에 1점의 빈도가 9%(232회)로 가장 적었고, 3점이 27%(686회)로 가장 많았다. 척도 상에서는 낮은 점수나 높은 점수의 빈도가 많은 것으로 나타나지는 않았기 때문에 척도로 인한 엄격성/관대성의 영향은 없는 것으로 판단된다.

다음으로 <표 18>의 하단에 제시한 채점자 집단에 나타나는 엄격성/관대성의 영향에 관하여 분포를 바탕으로 확인하여 보았다. 채점자들의 엄격성을 고정한 카이제곱 검정에서는 카이제곱값이 569.9(df=12)로 나타났으며, 유의수준은 $p<.005$로 영가설을 기각하여 채점자들의 엄격성/관대성 수준에 차이가 있는 것으로 나타났다.

채점자별 엄격성의 차이가 집단 내에서 어느 정도로 구분되어 있는지와 관련하여 분리비(separation ratio)를 확인하였을 때, 모집단(population)에서의 분리비가 6.82로 나타났으며, 이는 채점자 간 엄격성 차이가 측정 오차보다 약 7배 크다는 것을 의미한다. 층위(strata)를 확인하였을 때는 9.43 개의 층으로 나뉘어지는 것으로 나타났으며, 채점자들 간의 차이의 신뢰도도 .98로 1에 가까운 것으로 나타나, 서로 매우 다른 채점을 수행하고 있는 것으로 나타났다.

ⓑ 개인 수준 영향

개인별 엄격성 수준의 차이를 알아보기 위하여, Facets의 종합 분석표와 채점자 영향 분석표를 확인하였다. 종합 분석표에서 채점자 엄격성/관대성은 0 logit ~ -3 logit 사이에 위치하고 있었으며, 그 중 가장 관대성이 높은 채점자는 R13이었고, 관찰 점수를 기준으로 하였을 때는 전체 평균보다 0.71점이 높았으며, MFRM 분석으로 환산한 조정 점수에서는 0.80점이 높은 것으로 나타났다. 이는 R13이 다른 채점자들에 비해서 평균적으로 0.8점을 더 부여할 수 있다는 의미이고, 1~5점을 부여하는 수행 측정 기준에서 평균적으로 16%의 점수를 더 받는다는 뜻이다. R09는 평가 결과에 나타난 채점자 영향이 0 logit에 해당하여 보통 수준으로 나타나 영향이 나타나지 않았다. 이러한 경우에는 관찰 점수와 보정 점수의 차이가 없으므로, 종합적으로 보았을 때 부여한 점수에 대한 엄격성/관대성 영향이 없다는 것을 의미한다. 전체 채점자의 측정값을 비교하였을 때, 엄격성 표준 오차 변량을 기준으로 R13은 24.63 표준 오차만큼 관대하였으며, 표준 오차 변량이 크다는 것은 그만큼 능력 측정에 대한 신뢰성이 떨어짐을 의미한다(<표 Ⅲ-16> 참조).

<표 Ⅲ-16> 채점자 영향 분석 결과

채점자	관찰평균	조정평균	Rasch Model		
			측정값	표준오차	엄격성 표준오차
R13	3.93	4.10	-2.71	0.11	-24.63
R08	3.66	3.79	-2.14	0.1	-21.4
R02	3.49	3.6	-1.81	0.1	-18.1
R07	3.42	3.52	-1.67	0.1	-16.7
R12	3.4	3.49	-1.63	0.1	-16.3
R10	3.3	3.38	-1.44	0.1	-14.4
R01	3.26	3.34	-1.36	0.1	-13.6
R06	3.1	3.17	-1.08	0.1	-10.8
R03	3.08	3.14	-1.03	0.1	-10.3
R11	3.08	3.14	-1.03	0.1	-10.3
R04	2.9	2.94	-0.72	0.1	-7.2
R05	2.79	2.83	-0.53	0.1	-5.3
R09	2.48	2.48	0.00	0.09	0.0
평균	3.22	3.30	-1.32	0.10	-13.2

채점 경향성은 그 원인이 채점 대상으로 인하여 나타나는 존재론적 영향인지 아니면 채점 방법으로 인하여 나타나는 인식론적 영향인지를 구별할 필요가 있다. 채점 방법의 영향에 관하여서는 보정 평균(fair average score)의 차이를 비교하여 추정할 수 있는데, 가장 관대한 채점을 한 R13은 채점자 경향성이 0인 R09보다 1.62점을 더 부여하고 있었으며, 이는 채점 과정에서 주목한 정보와 점수 결정 과정의 차이 등으로 인하여 나타난 것으로 판단된다. 채점 경향성이 나타나는 과정과 영향 요인에 대해서 Ⅳ장의 채점 과정 분석을 통하여 알아보기로 하였다.

③ 제한적인 척도 사용의 영향

채점 과정에서 채점자 영향은 점수 결정과 관련하여 채점 척도 사용의 문제로 인하여 발생할 수 있다. 척도 사용의 문제 가운데 집중 경향성은 척도의 중앙에 위치한 점수를 중심으로 점수를 부여하는 현상을 가리킨다. 전체 채점자에 대한 중심 경향성 분석은 척도별 사용 빈도 분석, 채점자 영향별 빈도 분석, 그리고 수험자의 분포 및 적합도 분석으로 이루어지며, 채점자별 중심 경향성 분석은 각 채점자의 특성별 빈도 분석, 채점 적합도 및 채점 척도 사용에 나타난 특징을 바탕으로 확인하였다.

ⓐ 집단 수준 영향

집중 경향성에 관한 집단 수준의 분석 결과, 채점자들이 부여한 점수 가운데 가장 높은 비율을 차지한 것은 1~5점 척도의 중앙값인 3점(27%)이었으며, 집중 경향이 있음을 확인할 수 있었다. 채점자들이 중앙값 선택을 통하여 얻을 수 있는 효용은 안전성과 효율성이라고 볼 수 있으나, 과제나 능력 특성을 충분히 고려하지 못한 상황에서는 과대 혹은 과소 추정이 이루어질 수 있다.

자세한 분석을 위하여 수험자 분석 결과표에 제시된 수험자 분포의 특징을 확인할 필요가 있다. 수험자 분석표에서 수험자의 수행 능력을 고정한 카이제

곱값은 2989.1(df=11)이었으며, 유의도는 p<.005이므로 차이가 나타났으며, 집단 수준에서의 집중 경향성은 나타나지 않았다. 다음으로 MFRM 분석에 따른 수험자 능력 수준의 변별을 의미하는 분리비와 층위, 분리신뢰도를 확인하여 보았을 때, 분리비는 16.34였으며, 그에 따른 층위는 22.12로 12명의 수험자 사이의 집중 경향성은 나타나지 않았다. 수험자 분석 결과에서 나타난 분리신뢰도는 1.0으로 매우 높게 나타났으며, 이는 채점자들이 수험자의 수행 수준을 충분하게 구별할 수 있었음을 나타낸 것으로 볼 수 있다. 각 평가 준거에서의 적합도 분석을 통해서도 집중 경향성 양상을 확인할 수 있는데, 평균제곱값이 1보다 작은 경우에는 특정 척도 범주를 더 많이 사용하였을 때 나타날 수 있다. 분석 결과에서 1보다 작은 제곱평균값이 나타난 말하기 특성은 발음(..90), 담화(.98), 전반적 수행(.97)이었으며, R05는 3점을 사용한 비율이 36%로 가장 높았으며, 채점 영역별로는 전반적 수행에서 33.3%, 발음의 41.6%, 어휘·문법의 33.3%, 담화의 35.4%에서 3점을 부여한 것으로 나타났다.

ⓑ 개인 수준 영향

집중 경향성에 대한 채점자의 개인 수준 분석을 위하여 먼저 채점자 영향 분석표에 제시된 적합도를 살펴보았다. 적합도의 의미는 채점자가 수험자의 능력 특성에 알맞은 점수를 부여하는 가와 관련이 있기 때문에, 적합도가 지나치게 높은 경우(overfit, MnSq < 1)에는 집중 경향성을 예상할 수 있다. 분석 과정에서는 평가 준거나 수험자 집단 특성에 따라 적절하게 채점을 하였을 때 나타나는 긍정적인 집중 경향성도 나타날 수 있으므로 적합도와 함께 관찰점수와 조정점수의 차이, 그리고 신뢰도를 함께 고려하였다.

<표 Ⅲ-16>에서 적합도 분석 결과를 살펴보면, 적합도가 1보다 작아서 집중 경향성을 나타내는 채점자는 8명이었으며, 그 중 R05의 과적합도가 가장 높았으며, 그 다음으로 과적합 양상을 나타낸 R06은 점수대역이 한정되었을 것으로 추정되었으나, 채점자 간 신뢰도 측면에서 우수한 것(관찰점수-측정값 상관=.88)으로 나타나 집중 경향성이 중앙값이 아닌 다른 점수대에서 나타나고 있는 것을 확인하였다. 적합도가 1보다 큰 채점자 중에 R13은 과제×채점 영역 분석 결과에서 '경험 말하기' 문항에서 모든 점수를 두 가지 척도(4점과 5점)

에만 부여한 것으로 나타났다. 채점자별로 빈도 확인을 하였을 때, 3점에 대한 집중 경향성이 가장 큰 채점자는 R05이었으며, 전체 192개 중 69개(36%)에 3점을 부여한 것으로 나타났다. R05는 5점은 한 번도 부여하지 않았으며, 척도를 제한적으로 사용한 것으로 나타났다. R01도 중심경향성이 나타났으며, 3점을 부여한 비율이 33%로 나타났다. R13과 R08은 척도의 중앙이 아닌 높은 척도에 집중 경향성이 나타났으며, 이는 곧 관대한 채점자 영향과 연결되어 있다(<표 Ⅲ-17> 참조).

<표 Ⅲ-17> 채점자의 척도별 점수 분포 비율(단위: %)

척도	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R13	누적
0	0	0	0	4	2	2	2	0	4	3	0	2	2	21
1	6	7	16	13	6	13	6	3	26	6	13	5	3	123
2	16	15	15	26	30	14	14	13	18	18	19	16	8	222
3	33	23	26	23	36	32	24	33	29	28	29	25	16	357
4	35	31	32	16	27	23	34	18	19	24	27	32	29	347
5	10	23	11	19	0	16	20	33	5	21	13	20	42	233
누적	100	100	100	100	100	100	100	100	100	100	100	100	100	1,300

집중 경향성과 관련하여 등급 척도(0~5점) 범주의 한계치(thresholds)를 확인하여, 한계치는 수험자의 점수 척도별 확률 곡선이 교차하는 지점으로서, 수험자가 50%의 확률로 점수를 받을 수 있다는 의미이다(Andrich, 1988). 한계치에서는 집중 경향이 나타날수록 곡선이 퍼져있는 모양으로 나타나는데, 채점자별로 척도에 따른 한계치를 비교해보면, R01은 한계치 간 평균 거리가 가장 넓었으며(2.07), 반대로 R13은 가장 좁게 나타났다(1.02)(<표 Ⅲ-18> 참조).

<표 Ⅲ-18> 채점자의 척도별 한계치

범주	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R13
0-1	-	-	-	-2.90	-2.16	-3.44	-2.56	-	-4.16	-2.36	-	-2.32	-1.62
1-2	-2.78	-2.20	-1.69	-1.50	-1.84	-.88	-1.52	-2.74	-.67	-1.92	-2.23	-2.03	-1.43
2-3	-1.34	-.78	-1.30	.40	1.00	-.51	-.30	-.97	-.40	-.16	-1.02	-.16	-.26
3-4	.70	.55	.26	1.79	3.01	1.79	1.09	1.85	1.65	1.61	.66	1.22	.87
4-5	3.43	2.43	2.73	2.21	-	3.05	3.28	1.86	3.58	2.83	2.59	3.28	2.44
거리 평균	2.07	1.54	1.47	1.28	1.72	1.62	1.46	1.53	1.94	1.30	1.61	1.40	1.02

다음으로 수험자에 대한 관찰 점수와 조정 점수의 차이가 작을수록 척도별 외적합도(Outfit Mnsq)는 1에 근접하는 값이 나타나는데, 12명의 수험자 중 관찰 점수-조정 점수를 계산하였을 때 E03은 .08점의 가장 큰 차이가 나타났다. 반대로 차이가 가장 작은 수험자는 E06이었으며 .01점의 차이가 나타났다. 이와 관련하여 채점자들의 채점 척도 사용의 외적합도를 살펴보면, R06은 점수 척도별 외적합도 평균이 0.57로 가장 낮았으며(과적합), 반대로 R02는 1.48로 가장 높게 나타났다(부적합)(<표 Ⅲ-19> 참조).

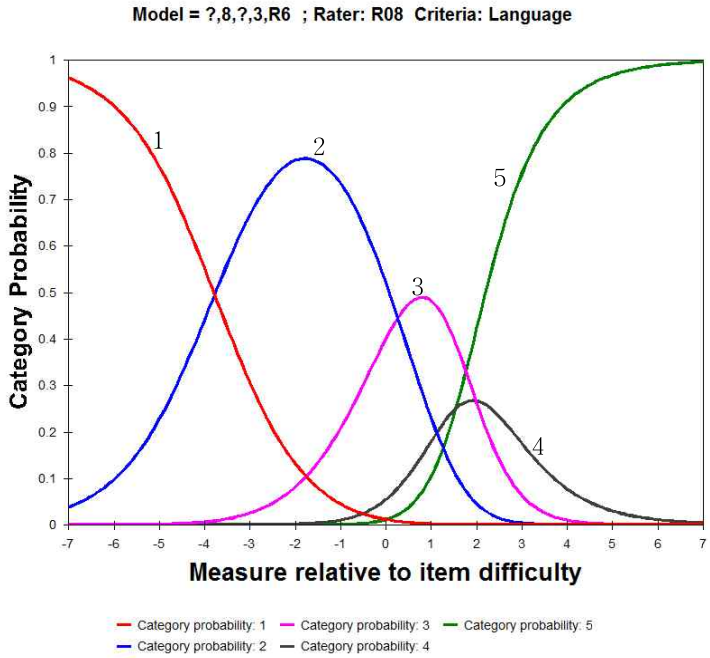
<표 Ⅲ-19> 채점자의 척도별 외적합도

척도	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R13
0	-	-	-	.7	1.1	1.0	1.1	-	1.4	1.0	-	1.0	1.3
1	.8	1.4	.5	.5	.8	.5	.6	.7	1.3	.4	.5	.5	1.9
2	.8	.8	.8	.5	.7	.6	.9	1.7	1.2	1.0	1.0	.6	1.5
3	.9	1.9	.8	.7	.7	.5	1.1	1.7	1.3	.9	.7	1.7	1.3
4	.9	1.5	1.1	1.0	.6	..6	.8	1.2	1.0	.7	1.3	1.0	1.0
5	1.0	1.8	1.1	.9	-	.8	1.2	.9	1.0	1.2	.9	1.0	1.6
평균	0.88	1.48	0.86	0.72	0.78	0.57	0.95	1.24	1.20	0.87	0.88	0.97	1.43

마지막으로 집중 경향성과 관련하여 MFRM 분석에서 확인할 수 있는 정보는 '분석 모형의 예측에서 벗어난 평가 결과(unexpected responses)'이다. R13의 경우를 제외하고 모두 예상 값보다 점수를 낮게 준 것으로 나타난 가운데, 표준화 잔차의 크기에 따라서 R08이 E08의 1번 문항에서 어휘·문법 사용 능력을 채점한 결과의 표준화 잔차가 가장 큰 것으로 나타났다. 이와 관련하여 R08의 어휘·문법 사용 영역에 대한 척도 특성 곡선을 확인한 결과, 2점을 받을 확률에 대한 집중 경향성이 나타나는 가운데, 3~5점 척도가 서로 중첩하여 있어 변별력이 떨어지는 것으로 나타났다(<표 Ⅲ-20>, [그림 Ⅲ-4] 참조). 이러한 탈모형적 채점 사례들을 살펴보면, 문항 중에서는 1번 경험말하기가 75%, 채점 영역에서는 어휘·문법 사용 능력이 50%로 빈도가 가장 높았다.

<표 Ⅲ-20> MFRM 분석 결과에 나타난 채점자 편향

문항	채점영역	수험자	조정 점수	관찰 점수	잔차	표준화 잔차[36]	채점자
1	어휘·문법	E08	4.8	3	-1.8	-4.7	R08
1	어휘·문법	E04	4.8	3	-1.8	-4	R12
1	담화	E04	4.8	3	-1.8	-4	R12
1	어휘·문법	E08	4.8	3	-1.8	-4	R02
1	전반적	E04	4.8	3	-1.8	-3.8	R12
3	전반적	E03	4.4	2	-2.4	-3.6	R10
3	발음	E05	2.1	5	2.9	3.6	R13
2	전반적	E12	2.7	0	-2.7	-3.4	R09
1	어휘·문법	E02	4.2	2	-2.2	-3.3	R08
1	어휘·문법	E10	4.7	3	-1.7	-3.3	R02
1	발음	E10	4.7	3	-1.7	-3.3	R07
1	어휘·문법	E10	4.6	3	-1.6	-3.1	R07



[그림 Ⅲ-4] R08의 어휘·문법 사용 능력의 척도별 확률 곡선

36) 표준화 잔차(Standardized Residual)는 관찰 점수가 기대 모형으로부터 얼마나 떨어져 있는지를 나타낸 값이다.

④ 주관적 채점 척도 사용의 영향

다음으로 분석 결과에서 채점자 영향 중에 척도 사용의 주관성으로 인한 무작위성과 후광성이 나타났다. 무작위성과 후광성은 채점에서 공통적으로 적용해야 하는 과제나 채점 기준을 따르지 않고, 임의적이거나 또는 특성 차이를 고려하지 않는 주관적인 채점의 결과라고 볼 수 있다.

ⓐ 무작위성 영향

채점자 영향으로서 무작위성은 평가 준거와 척도를 고려하는 경향에 일관성이 없으며, 따라서 수험자의 능력 수준을 신뢰할 만한 수준으로 측정하지 못하였다는 증거로 볼 수 있다. 무작위성은 수험자 국면의 분석 결과 중 고정된 카이제곱값과 분리비, 층위, 분리비의 신뢰도 확인을 통하여 이루어졌다.

전체 Facets 분석 결과에서 고정된 카이제곱값은 2989.1(d.f.=11)이였으며, 유의한 차이가 있는 것으로 나타났으며($p<.01$), 이는 무작위성 영향에 대한 증거가 수험자 분포 상에 나타나지 않았음을 말한다. 다음으로 모집단의 분리비(16.34)와 층위(22.12), 분리 신뢰도(1.00)를 확인하였을 때, 채점자들은 수험자 간의 능력 차이를 충분한 수준에서 구별하고 있는 것으로 나타났으며, 따라서 무작위성의 영향을 확인할 수 없었다.

개별 채점자 수준에서의 무작위성의 영향을 알아보기 위하여, 채점자의 적합도 지수를 확인하였을 때, 본 연구에 참여한 채점자들 중 적합도 지수가 1을 넘는 채점자는 모두 5명이었으며, 그 중 가장 적합도 지수가 큰 것으로 나타난 채점자는 R13이었다(Infit MnSq: 1.82, Outfit MnSq: 1.48).[37] 적합도 상에서 부적합으로 나타난 것과 관련하여 R13과 다른 채점자들 간의 신뢰도(Correlation PtMea-PtExp)를 확인하였을 때 관찰 점수에서나 보정 점수에서 모두 가장 낮은 신뢰도 수준을 나타내었다(PtMea: .66, PtExp: .76). R13은 적합도 지수와 채점자 신뢰도 분석에서 특이점을 나타났기 때문에, 점수 결정 과정에서도 다른

37) 무작위성을 의심할 수 있는 경우는 기준값인 1보다 큰 적합도 지수를 나타낸 경우이다. 이런 경우에는 채점자가 수험자의 수행으로부터 능력 특성 수준의 차이를 구별하지 않았기 때문에 발생한다.

채점자들과의 차이가 나타날 것으로 예상이 되었다.

ⓑ 후광성 영향

다음으로 후광성에 대한 집단 수준의 채점자 영향을 확인하기 위하여, 등급 척도 모형에서 각 채점 영역별 점수 분포의 고정된 카이제곱값과 분리비, 층위, 분리신뢰도를 확인하였다. 분석 결과에서 4개 평가 준거의 곤란도 사이에 통계적으로 유의한 차이가 나타났다($p<.05$). 이는 평가 결과에서 채점자들이 네 가지 준거를 분리해서 채점하고 있음을 의미하며 따라서, 후광성은 확인할 수 없었다. 다음으로 집단비(1.33), 층위(2.11), 분리신뢰도(.64)를 확인하였을 때는 수험자들의 말하기 능력 특성을 약 2개 수준으로밖에 구분하지 못하였고, 분리 신뢰도 측면에서도 0~1 사이의 중앙에 근접하여 위치하고 있어 특성 수준을 잘 구별하지 못하였음을 알 수 있었다.

개별 채점자들의 수준에서 살펴보면, 무작위성 분석과 마찬가지로 채점자별 누적 신뢰도와 적합도 지수의 측면에서 살펴보았을 때, R13은 부적합 수준이었으며, 채점자 간 신뢰도도 상대적으로 낮은 수준이어서 후광성의 영향을 받은 것으로 추정되었다(<표 Ⅲ-16> 참조). 평가 준거별 척도 사용이 서로 영향을 받는 후광성을 확인하기 위하여 대조적으로 채점자 간 신뢰도가 가장 우수한 수준이었던 R06과의 점수열(score string)을 비교하였을 때, R13의 평가 준거별 점수의 유사성이 높아 후광성 영향이 있었음을 확인할 수 있었다(<표 Ⅲ-21> 참조).

<표 Ⅲ-21> R06와 R13의 점수열 비교

	E02(전반적 / 발음 / 어휘·문법 / 담화)			
채점자	문항 1	문항 2	문항 3	문항 4
R06	3 / 4 / 4 / 3	2 / 4 / 3 / 3	3 / 3 / 3 / 4	3 / 3 / 3 / 3
R13	5 / 5 / 5 / 5	5 / 5 / 5 / 4	5 / 5 / 5 / 5	3 / 5 / 4 / 4

ⓒ 평가 요소와의 상호작용적 영향

채점 과정에서 채점 척도나 준거를 구별하지 못하는 현상과 관련하여

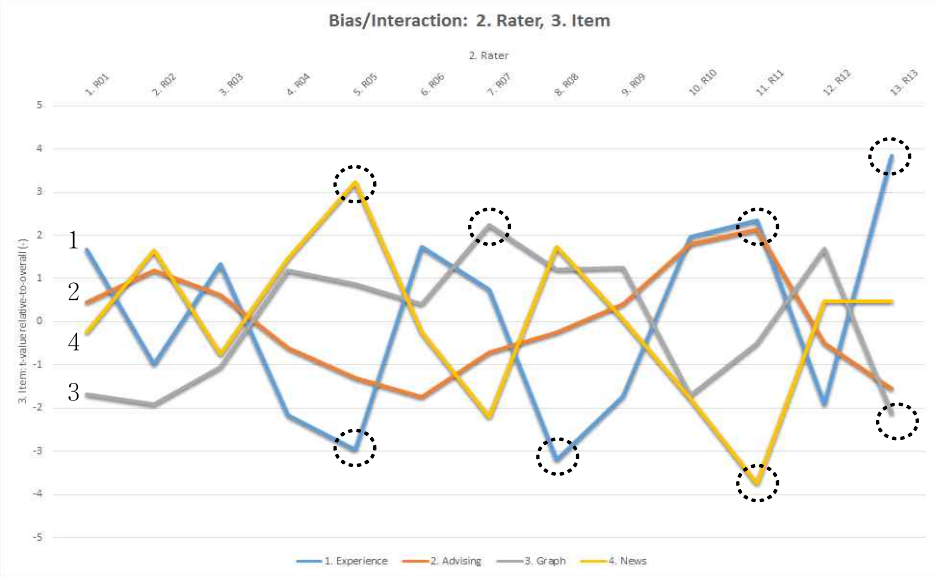
MFRM을 통해 접근할 수 있는 방법 중의 하나는 '채점자 편향(differential rater functioning; DRF, Eckes, 2011) 분석'을 활용하는 것이다. DRF 분석은 채점자가 수험자, 과제, 특성 등의 평가 국면과의 상호작용에서 모형의 예측에서 벗어나 엄격하거나 관대한 경향을 확인하는 방법(Eckes, 2011:90)으로, 본 연구에서는 채점자 영향을 중심으로 분석을 실시하였다(<표 Ⅲ-22> 참조).

<표 Ⅲ-22> 평가 국면별 상호작용 편향 분석

통계량	채점자와 평가 국면별 상호작용 유형					
	수험자×채점자	과제×채점자	준거×채점자	수험자×채점자×과제	수험자×채점자×준거	수험자×과제×준거×채점자
상호작용수	156	52	52	624	624	2496
±2 척도 초과 t(%)	33.33	21.15	3.85	18.4	5.93	0.00
유의확률.05미만(%)	28.85	21.15	3.85	3.04	0.48	0.00
최소t(df)	-4.22	-3.75	-2.37	-5.3	-3.93	-3.25
최대t(df)	5.28	3.85	2.46	4.61	2.6	2.57
평균	-0.04	0.20	0.00	-0.05	-0.03	-0.14
표준편차	1.9	0.02	0.88	1.53	1.12	0.85

채점자 국면을 기준으로 수험자, 과제, 특성과의 상호작용을 분석한 결과, 평가 과정과 관련된 과제와 말하기 특성을 고려하는 상호작용 중 모형 예상치를 벗어나는 사례가 가장 많았던 것은 '과제와 채점자의 상호작용(21.15%)'이었으며, 다음으로는 '수험자와 채점자, 과제가 함께 상호작용'한 경우(18.4%)였고, 모두 '과제'가 상호작용에 개입하였다는 공통점이 있었다.[38] 이러한 편향 차이를 정교화하기 위하여 유의확률 수준을 검토한 결과, '과제와 채점자의 상호작용'이 21.15%로 다른 상호작용에 비하여 월등하게 높게 나타났다. '과제'와의 상호작용으로 인한 채점자 편향 사례(p<.05)를 자세히 살펴보면, '경험 말하기' 문항이 5회로 가장 높은 빈도를 나타내었고, 그 다음은 '뉴스 읽고 의견 말하기' 문항이 3회, '그래프를 보고 설명하기' 문항이 2회, '조언하기' 문항이 1회로 나타났다.

38) 가장 큰 상호작용 편향 비율이 나타난 요소는 수험자(33.33%)였다. 이러한 상호작용에 의한 편향은 평가 상황과 관련하여 발생한 것으로 추정된다. 상호작용으로 인한 편향 연구를 위해서는 수험자와 채점자가 모두 익숙한 상황을 바탕으로 하는 연구 설계가 필요할 것으로 판단된다.



[그림 Ⅲ-5] 문항 × 채점자 상호작용 분석 도표

다음으로 평가 문항과 채점자의 상호작용 분석의 결과를 도표를 바탕으로 확인하였다. [그림 Ⅲ-5]에 제시한 평가 문항과 채점자 상호작용 차이를 비교하는 도표에서 t검정 결과가 ±2를 벗어나 특정 문항에 대해서 더 엄격한 채점자(R05, R07, R08, R11, R13)와 더 관대한 채점자(R05, R07, R11, R13)를 확인할 수 있다. R05는 1번 문항에서는 상대적으로 더 관대한 경향을, 그리고 4번 문항에서는 상대적으로 더 엄격한 경향을 나타내는 양상을 보였다. 문항별로 보았을 때는 2, 3번 문항의 편차에 비하여 1, 4번 문항의 편차가 더 큰 것으로 나타났다.

[그림 Ⅲ-6] 평가 준거 × 채점자 상호작용 도표

다음으로 4개 평가 준거와 채점자 간의 상호작용 결과에 나타난 편향을 확인하였다. [그림 Ⅲ-6]은 평가 준거와 채점자 상호작용에 따른 차이 비교를 위하여 t검정 결과를 그래프로 나타낸 것이다. 분석 결과에서 t 값이 ±2를 넘어서는 경우 중 모형보다 더 관대한 채점을 한 경우는 R04의 발음 채점이었으며, 더 엄격한 채점을 한 경우는 R08의 어휘·문법에 대한 채점이었다.

이상 Ⅲ장에서 살펴본 말하기 평가 결과에 나타난 채점자 영향의 분석 결과를 요약하면 다음과 같다. 첫째, '한국어 말하기 능력 시험'의 평가 결과에서 채점자의 전반적인 채점 경향은 관대한 것으로 나타났다. 이러한 현상이 어떻게 나타난 것인지 파악하기 위하여 채점자 가운데 가장 관대한 채점 경향을 보인 R13과 평가 국면의 상호작용 분석 결과에서 분석 모형의 예측치를 초월한 엄격성이나 관대성을 나타낸 채점자가 있음을 확인하였다. 이러한 현상의 원인을 파악하기 위해서는 해당 채점 사례의 점수를 결정하기까지 어떤 영향 요소가 개입한 것인지를 확인하기 위하여 같은 문항이나 준거에 대한 다른 채

점자의 채점 과정 비교가 이루어질 필요가 있다. 둘째, R05와 같이 채점자 영향 가운데 특정 점수를 집중적으로 부여하는 경향이 나타난 경우에는 3점을 가장 많이 사용한 것 외에도 한 번도 5점을 부여하지 않은 척도 사용의 특징이 나타났다. 제한적인 채점 척도의 사용과 관련하여 특정 점수 척도의 집중적인 사용 경향이 두드러진 채점 사례에서 다른 채점자들의 채점 과정을 비교하여 집중 경향성이 나타난 경로를 확인해야 한다. 셋째, 일부 채점자들은 무작위성과 같이 분석 모형의 예상을 벗어나는 부적합 또는 과적합한 채점 경향을 나타냈다. 채점 척도 사용의 차이의 원인을 파악하기 위하여 해당 채점 경향을 나타낸 채점자의 채점 과정의 차이를 알아보아야 한다. 넷째, 분석적 채점에서 개별적으로 채점해야 하는 평가 준거 사이의 구별을 하지 않는 후광성 영향이 나타난 경우가 있었다. 이와 관련하여 채점자 영향별로 채점자의 채점 과정을 비교하여 원인을 파악할 필요가 있다. 다섯째, 말하기 평가의 채점 결과를 변화시킬 수 있는 채점자 영향이 안정적인 채점자 중에도 일부 문항과 평가 준거와 상호작용하는 가운데 모형의 예상을 벗어나 편향된 채점을 한 경우가 나타났다. 편향적 채점 사례와 같은 문항 및 평가 준거의 점수 결정을 하는 채점 과정의 특징을 파악할 필요가 있다.

Ⅳ. 말하기 평가의 채점 과정 분석

Ⅳ장에서는 Ⅲ장에서 제시한 '한국어 말하기 능력 시험'의 문항과 MFRM 분석의 결과를 바탕으로 말하기 평가의 실제적인 채점 과정을 분석하였다. 채점 과정은 채점자들은 언어 프로토콜 보고법(Ericsson & Simon, 1993)을 바탕으로 실제적인 말하기 성취도 평가의 채점 상황에 맞추어 수정한 채점 과정 보고법을 따라 자신의 채점 과정을 보고하였다[39]. 채점 과정 보고 자료에 대한 분석은 평가 타당화에 관한 논거 기반 접근법(Kane, 1992, 2006, 2013; Mislevy et al., 2003; Chapelle et al., 2008)을 바탕으로 논증 요소를 중심으로 이루어졌다. 이는 채점 과정이 점수 결정에 이르는 채점자의 논증 과정이라고 보는 것인데, 본 연구에서는 이를 위하여 채점 과정에 대한 보고 발화를 논증 요소로 코딩하여 분석 단위별로 채점 과정의 구성과 발화 내용에 나타난 특징을 파악하고자 하였다.

1. 채점 과정 보고 자료의 수집과 분석 설계

채점 과정에 관한 자료의 수집은 채점자가 수험자의 말하기 평가 응답을 청취하면서부터 점수를 결정하기까지 떠오르는 생각을 보고하는 '채점 과정 보고'를 통하여 이루어졌다. 말하기 평가의 채점 과정은 채점 수행의 전개에 따라 응답 청취 과정과 점수 결정 과정으로 나눌 수 있는데, 채점자는 동시적 보고와 회상적 보고를 수행하면서 채점을 수행하였다. 채점 과정에 관한 채점자의 보고는 채점 수행을 통해 채점자가 주장하려는 평가 결과와 그에 관한 근거를 중심으로 구성되는데, 채점자의 인지적 특성과 보고 방법 등에 따라 보고

39) 언어 평가 연구에서 인지적인 처리 과정에 관한 자료 수집을 위하여 사용하는 대표적인 접근 방법은 언어 프로토콜 분석법(verbal protocol analysis; VPA)이다. VPA는 사고 구술법(think-aloud)을 기초로 보고의 절차와 형식을 강조하여 체계적인 자료 수집을 추구한다는 특징이 있다. VPA를 적용한 언어 평가 연구는 평가 맥락 요소 및 평가 요소의 관계를 파악하기 위하여 이루어졌으며, 주로 평가자의 평가 과정이나 수험자의 과제 수행 과정에 대한 동시적 절차 보고와 회상적 절차 보고로 이루어져왔다. 이러한 접근은 평가 상황에서는 나타나지 않는 내성적인 측면에 대한 정보를 제공한다는 점에서 평가 타당화에 관한 정성적 증거를 제공한다.

수준의 차이가 나타날 수 있으므로 자료 수집 과정에서 이에 대한 대비가 이루어져야 한다고 보았다. 이에 본 연구에서는 연구에 참여한 채점자들에 대한 채점 과정 보고 교육과 채점 과정 보고 연습, 피드백 제공을 통하여 채점자의 채점 과정 보고 자료의 타당성을 확보하고자 한다.

1) 채점 과정 보고 자료 수집의 설계

말하기 평가에서 채점자가 결정한 점수에 대한 통계 분석을 통해 확인한 채점자 영향은 채점 과정으로부터 발생한 것이다. 어떤 채점 과정으로부터 채점자 영향이 나타난 것인지를 파악하는 것은 말하기 평가의 채점 과정에 대한 실제적인 이해를 제공하며, 채점 과정과 평가 결과의 인과적 관계에 대한 근거가 된다. 말하기 평가의 채점 과정에 대한 자료의 수집은 언어적 접근을 통해 이루어질 수 있으며, 시선 추적이나 뇌파 검사와 같은 물리적 접근과 달리 인지적인 처리에 관한 직접적인 정보를 얻을 수 있다는 장점이 있다(Ericsson & Simon, 1985: 259). 채점 과정 보고는 채점자의 인지적인 채점 과정을 확인할 수 있는 질적 담화 자료로서 음성 자료라는 성격을 고려하여 분석을 위해 녹음한 후에 전사 및 의미 단위에 따라 분절하는 형식화 처리를 실시하였다.

(1) 채점 과정 보고 자료의 수집 방법

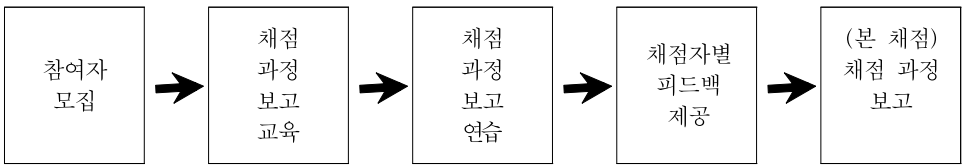
말하기 평가의 채점 과정을 파악하기 위한 자료의 수집은 채점 과정에 관한 접근법을 설정하는 것으로 시작한다. 말하기 평가 채점 과정은 채점자가 수험자의 과제 수행 응답을 청취하는 것과 그에 대한 수준 판정, 그리고 채점 척도에 기반하여 최종 점수를 결정하는 인지적 처리로 이루어진다. 각각의 인지적 처리 과정에서 채점자가 고려하는 평가 맥락적 요소와 개인적 요소에 차이가 있으며, 이는 채점 과정에서의 인지적 처리를 더욱 복잡하게 한다.

이러한 복잡한 채점 과정을 확인하기 위한 방법으로 언어 평가 연구에서 주목한 것은 채점자가 자신의 채점 과정이 어떻게 이루어졌는지를 설명하는 '언어 프로토콜 보고법(verbal protocol report)'이다. 언어 프로토콜(Ericsson & Simon, 1985, 1993), 또는 사고 구술 프로토콜은 어떤 활동을 수행하면서 머릿

속에서 주목한 모든 정보를 설명하는 활동을 통해 언어화한 자료를 바탕으로 해당 사고 과정의 특징을 파악하는 질적 자료 수집 방법이다. 언어 프로토콜은 연구 참여자가 절차적으로 보고한 내용을 바탕으로 특정한 인지 처리 현상에 대한 이해를 갖는 것을 목표로 하는데, 보고된 언어를 분석 자료로 삼기 때문에 자료 수집과 분석의 타당성 확보를 위하여 연구자의 주관성 개입을 주의해야 한다(Chi, 1997). 본 연구에서는 이와 관련하여 말하기 평가 상황에서 채점자가 채점 과정에 대한 언어 프로토콜 수집과 관련하여 개별 채점자에 대한 교육을 실시하고, 실제적으로 보고할 수 있는지를 알아보기 위하여 예비 검사를 실시하였다. 또한 채점 과정 보고의 형식적 요소인 채점 척도와 채점 규칙, 채점 보고 절차를 문서로 제공하여 체계적인 보고 수집이 이루어질 수 있도록 하였다.

(2) 채점 과정 보고 자료의 수집 과정

채점 과정 보고 자료의 수집 과정은 연구 참여자의 모집과 연구 참여 방법 안내, 그리고 채점 과정 보고 연습, 피드백 제공, 본 채점에서의 채점 과정 보고로 이루어졌다([그림 Ⅳ-1] 참조).



[그림 Ⅳ-1] 채점 과정 보고 자료의 수집 절차

채점 과정에 대한 자료 수집의 타당성은 절차 기반 타당도(process-based validity)와 관련이 있는데, 이는 체계적으로 자료를 수집할 수 있도록 보고 방법과 절차의 명세화와 교육 제공, 그리고 타당한 분석 절차를 수반하는 것으로 이루어 진다(Downing, 2003). 본 연구에서는 체계적인 자료 수집과 채점 과정 보고에 대한 채점자의 적응을 위하여 채점 과정 보고 방법에 관한 안내 교육,

채점 과정 보고 연습, 그리고 연습 결과에 대한 피드백을 제공하였다. 채점 과정 보고는 채점자가 수험자 응답을 청취하는 과정에서는 동시에 이루어질 수 있기 때문에 컴퓨터 헤드셋을 사용하여 청취 및 채점 과정 보고의 녹음이 동시에 가능하도록 하였다.

연구에 참여한 13명의 교사는 12명의 수험자들이 '한국어 말하기 능력 시험'의 4개 문항에서 산출한 응답을 채점하는 과정에서 채점 수행 발화를 녹음하는 것에 동의하였다. 채점자들은 연구자로부터 약 1시간동안 채점 과정 보고 방법에 대한 안내를 받았다[40]. 교육 과정에서는 채점자들에게 연구의 목적과 수집하고자 하는 자료의 성격과 참여 절차를 안내하였으며, 이해를 돕기 위해 예비 시행(pilot test)[41]을 통해 수집한 채점자 보고 사례를 제시하였다. 또한 채점 과정에서 컴퓨터로 채점 과정 보고를 녹음하면서 응답 청취와 채점을 해야 하는 물리적·인지적 부담 수준을 통제하기 위하여 작업 과정을 절차화[42]하여 일관되게 안내하였다. 채점자들이 채점 과정 보고 연습을 통하여 제출한 녹음 자료에 대해서는 자연스러운 사고 과정 보고와 구체적인 사고 과정 보고를 위하여 피드백을 제공하였다. 언어 보고의 타당도를 확보하기 위한 조건으로는 첫째, 보고 산출을 위한 적절한 교육을 실시하여 사고를 설명하는 일에 대한 자신감을 갖도록 해야 하며, 또한 연구자의 개입을 최소화하여야 하고, 과제 완성과 언어적 보고 사이의 시간 지연이 일어나지 않도록 해야 한다는 점이다(Green, 1998: 10). 본 연구에서는 채점 과정 보고의 연습 수행에 관한 피드백으로 녹음 자료에 나타나는 음향적인 문제(예: 음성 크기, 소음 관련)와 보고 절차 관련 사항(예: 청취 후 보고에서 지속적으로 발화할 것), 보고 내용 관련 사항(예: 참고 자료(채점 척도)를 볼 때도 이야기할 것)을 제공하였다. 채점자들은 채점 과정 보고 연습 및 피드백을 제공 받은 후에 12명의 수험자들의 말하기 평가 응답에 대한 채점을 하면서 채점 과정을 보고하였다.

2) 채점 과정 보고 자료 분석의 설계

---

40) 채점자의 채점 과정 보고 방법에 관한 안내는 <부록 3>의 '채점 과정 보고 수행을 위한 사전 교육 자료'를 바탕으로 이루어졌다.

41) 말하기 평가의 채점 과정 보고 타당성 확인을 위한 예비 시행은 2015년 9월에 이루어졌으며, 초임 한국어교사 2명과 경력교사 1명이 참여하였다.

42) 채점 과정 보고 수행의 절차는 <부록 4>에 제시하였다.

(1) 채점 과정 보고 자료의 형식화

채점자의 채점 과정 보고 발화는 컴퓨터 녹음 프로그램을 사용하여 헤드셋을 통해 수집하였다. 채점 과정 보고 자료의 분석은 수집한 자료의 성격이 구어 담화 자료라는 점을 고려하여 전사한 후에 보고 방법에 따라 분할하고, 다시 코딩을 위하여 보고의 의미 단위로 분절하는 형식화(formatting) 과정을 거쳤으며, 이는 Chi(1997)에서 제시한 언어 보고 자료의 질적 분석 절차를 바탕으로 하였다.

본 연구에서는 채점 과정에 대한 자료 수집을 위하여 연구에 참여한 13명의 채점자에게 12명의 수험자의 음성 답안을 들으면서 채점하는 과정에서 떠오르는 모든 생각을 보고할 것을 요청하였으며, 그 결과, 총 2,456분 분량의 채점 과정을 보고한 음성 자료를 수집하였다.

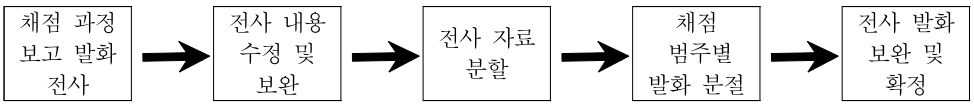
수집한 보고 자료의 전사는 내용의 신뢰도를 확보하기 위하여 2차에 걸쳐 이루어졌다. 1차는 인공지능 기반 음성 인식 기술(Google cloud console speech-to-text api)을 활용하여 전사하였으며, 2차에서는 3인의 보조원[43]이 해당 채점자의 채점 과정 보고를 청취하면서 1차 전사 자료를 보완하는 방식으로 이루어졌다. 2차 전사 과정에서는 전사 자료의 보완과 함께 채점자별 보고 내용을 크게 수험자와 문항에 따라 구분하고, 다시 보고 방법에 따라 내용을 구분할 수 있도록 기호를 부여하였으며, 보고 내용별로 분절(segmenting)하는 작업이 이루어졌다([그림 Ⅳ-2] 참조).[44][45]

---

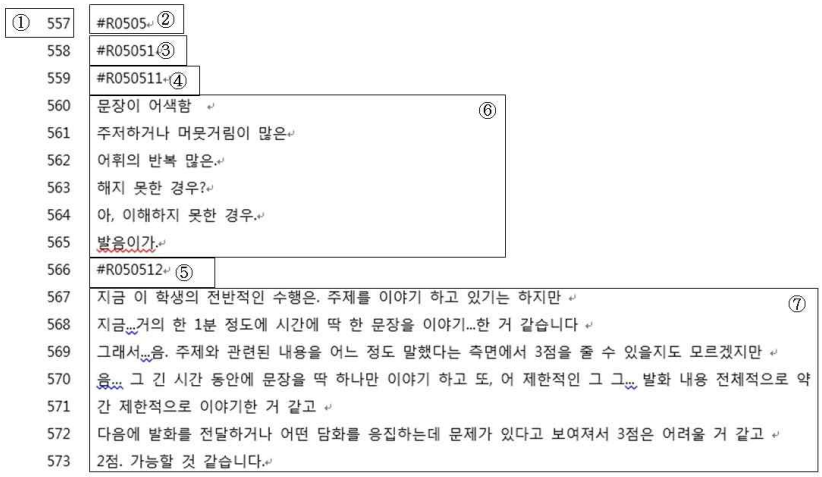
43) 채점 과정 보고의 전사 내용 보완 및 분절 작업을 담당한 3인은 대학원에서 한국어교육을 전공하는 이들이었으며, 연구자로부터 자료 정리 방법을 안내 받은 후에 작업을 수행하였다.

44) 채점 과정 보고 분석의 체계적 접근을 위하여 보고 내용을 세부적으로 분할 및 분절한 것은 Bejar(2012:4)에서 채점자들이 구성형 응답 문항을 채점할 때 채점 척도를 바탕으로 각각에 해당하는 증거들을 찾아서 점수 또는 등급에 따라 분류한다는 의견을 고려하여 코딩의 수월성을 확보하기 위하여 이루어졌다.

45) 채점자들의 채점 과정 보고에 대한 전사 자료는 MS Word(v. 2013)를 이용하여 기록하였다. 전사 자료는 채점 과정 보고에서 문자화가 가능한 모든 내용을 대상으로 하였으며, 불분명한 발화는 음성 자료의 음향적 보정을 통해 이해 가능한 것은 다시 기록하였으며, 알아들을 수 없는 일부 발화는 들리는 대로 표기하였다



[그림 Ⅳ-2] 채점 과정 보고 자료의 형식화 절차



[그림 Ⅳ-3] 채점 과정 보고 전사 자료의 형식화 사례

[그림 Ⅳ-3]은 채점자 R05가 수험자 E05의 1번 문제 응답에 대한 채점 과정 보고를 전사한 후 형식화한 것이다. 그림에서 ①은 채점 과정 보고를 전사한 문서의 줄 번호이다. ② ~ ⑤는 전체 발화를 사용 목적에 맞게 구분하기 위하여 입력한 보고 단위별 분류 코드인데, # 다음에 오는 숫자는 채점자, 수험자, 문항, 보고 성격을 나타낸 것이다. 예를 들어 ④의 '#R050511'에서 'R05'는 채점자, 그 다음의 '-05'는 수험자를 표시하며 그 뒤의 '1'은 문항 번호, 마지막 숫자는 보고 성격에 따라 동시적 보고는 '1', 회상적 보고는 '2'로 표시하여 구별하였다. 채점 과정 보고 가운데 ⑥은 수험자 응답 청취 중에 보고한 것이며, ⑦은 수험자 응답 청취를 마친 후에 점수를 결정하는 과정에서 이루어진 회상적 보고 내용이다.

(2) 채점 과정 보고 자료의 논증 요소 분석

말하기 평가의 채점 과정은 채점자가 청취한 응답을 바탕으로 수험자의 능력 수준을 추론하는 과정으로 볼 수 있으며, 이는 채점 과정이 채점자가 부여하는 점수를 결정하는 논증이라는 관점과 맥을 같이 한다(이성준, 2018). 이와 관련하여 본 연구에서는 형식화한 말하기 평가 채점 과정 보고 자료를 논증 요소로 코딩하여 채점 과정의 양상과 특징을 파악하고자 하였다.

언어 평가 연구에서 채점 과정은 '학습자 수행에 대한 채점자의 해석'(McNamara, 1997)으로 보거나, '과제 특성에 따른 변인들의 영향'(Bachman, 2002), 또는 '채점 척도에 따른 조작적인 점수의 사용'(Fulcher, 2003)으로 여겨져 왔으며, 담화 자료에 대한 질적 분석을 통한 연구에서는 채점 스타일(Vaughan, 1991), 채점자의 수행 자료 접근법(Cumming et al., 2002), 채점자의 인지적인 전략(Suto & Greatorex, 2008) 등으로 여겨져 왔다. 이러한 채점 과정에 관한 접근의 기본적인 관점은 채점자가 주어진 자원을 적절하게 활용하여 합리적으로 결론을 도출할 수 있다는 점이다. 채점 과정을 분석한 기존의 연구에서는 채점자의 점수 결정 방법(스타일)이나 판단을 내리는 채점자의 언어에 주목하였는데, 이러한 접근은 채점 과정을 구성하는 거시적·미시적 양상에 관한 정보를 제공한다는 의의가 있다. 그러나 자료 분석의 타당성 확보가 이루어지지 않은 상황에서 거시적인 접근의 경우에는 여러 가지 평가 요소가 중첩되어 있어 해석이 모호하며, 미시적인 접근의 경우에는 채점자의 발화 의도는 파악할 수 있으나, 전반적인 채점 과정을 논리적으로 설명하지 못한다는 한계가 있다.

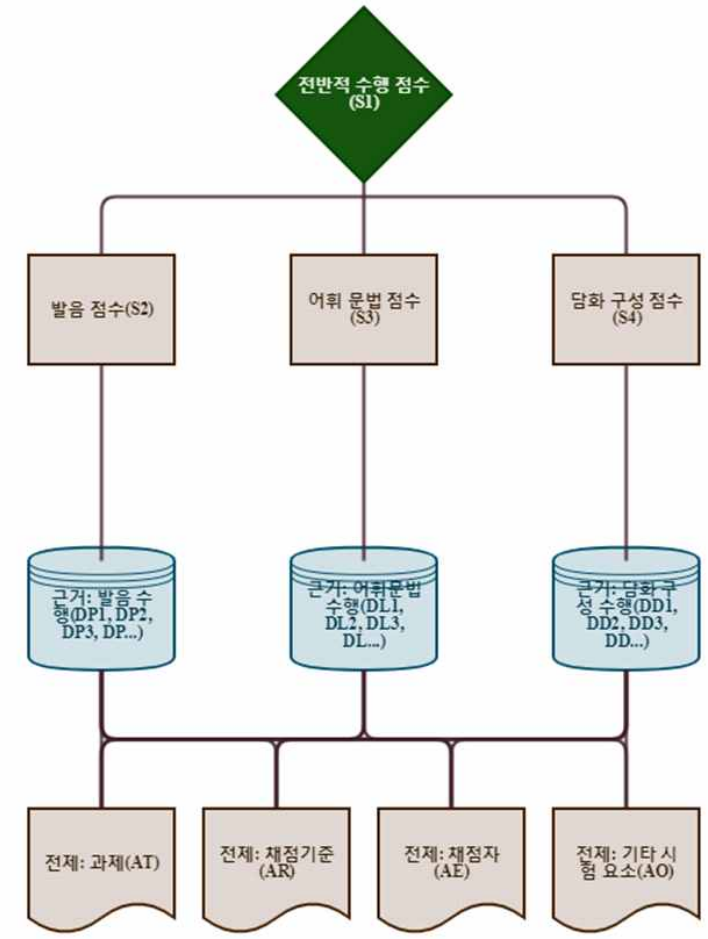
따라서 본 연구에서는 채점 과정 보고 자료를 분석하는 방법으로 채점자의 '추론 구조'를 파악하기 위한 논증 요소 분석을 실시하였다. 논증 요소 분석은 옳고 그름을 이유를 들어 살핀다는 논증의 정의를 바탕으로 하며, 채점자가 채점 과정에서 점수를 결정한 것에 대한 정당화가 채점 과정 보고에 나타난다는 점에서 채점 과정 분석을 위한 방법으로 선정하였다. 논증 요소 분석은 채점자의 주장인 최종 점수 결정과 채점 과정 보고에 나타난 점수 결정의 가정과 근거의 관계를 확인하는 방법으로 이루어졌다. 채점 과정을 논증으로 보는 관점은 평가 타당화를 위한 '논거 기반 접근법'(Kane, 1992, 2006, 2013; Mislevy et

al., 2003; 2006; Chapelle et al., 2008)에서 확인할 수 있는데, 이는 평가 전반이 평가 결과 및 활용을 위한 체계적인 논증이라는 점을 바탕으로 한다.

'논증'이란 합리성을 추구하는 담화 실천 행위인데, 논증과 관련하여 대표적인 툴민(Toulmin, 1958, 2004)의 논증 도식은 논증의 구성과 유형을 분류할 수 있는 틀로서 논증의 건전성을 평가할 수 있는 방법론을 제공하였다(민병곤, 2004: 185). 툴민의 논증 도식은 비형식 논리학을 바탕으로 하며, 자료(Datum)로부터 결론(Claim)에 이르는 과정에서 보장(warrant)과 이를 뒷받침하는 보증(Backing), 예외 조건(Rebuttal), 한정(Qualifer)을 고려하는 모형으로 이루어져있다. 툴민의 논증 도식은 일상 논증의 이론화에 기여하였으나, 비형식적 논증의 요소 가운데 근거로 작용할 수 있는 보장, 보증, 자료의 구별이 어려울 수 있고(Van Eemeren & Grrotendorst, 1992), 예외 조건에 대한 고려는 새로운 자료와 결론을 유도한다는 점에서 새로운 논증으로 봐야 하기 때문에 해석의 어려움이 있다. 툴민의 논증 도식을 적용할 수 있는 논증은 일상 논증 가운데에서도 형식성이 높은 상황에 적용이 가능하기 때문에 다양한 요소가 자유롭게 개입하는 구어 담화에는 적용이 어렵다고 볼 수 있다. 따라서 본 연구에서는 말하기 평가의 채점 과정 보고에 관한 분석의 틀로서 툴민의 논증 모형이 갖고 있는 한계를 고려하여 일상 담화 상황에 적용할 수 있는 논증 도식으로 키엔 포인트너(1992)에서 제시한 '보장(warrant), 논거(argument), 결론(conclusion)'을 바탕으로 채점 과정 보고를 분석하고자 하였다.

채점 과정에서 채점자들이 점수 결정을 위하여 주목한 수험자 응답의 특징과 그에 대한 인식은 점수 결정을 위한 보장하는 자료(data)라고 볼 수 있다. 근거는 채점자가 점수를 결정는 평가 준거에 따라서 나눌 수 있으며, 본 연구에서는 채점 척도에서 수험자 발화에 대한 분석적 평가 준거인 발음(Data for Pronunciation; DP), 어휘와 문법(Data for Language; DL), 담화(Data for Discourse; DD)를 따라 근거를 분류하였다. 채점자가 갖고 있는 평가에 관한 지식이나 관련된 경험을 통해 얻은 지식은 점수 결정을 가정(assumption)하는 평가 논거로서 특정한 결론을 내포하는 성격을 띤다. 이에 본 연구에서는 채점 과정에서 채점자가 고려하는 가정으로 채점자의 경험(Assumption by Experience; AE)이나 평가 과제의 요건(Assumption by Task; AT), 채점 척도의 내용(Assumption by Rating scale; AR), 그리고 그 밖의 평가 상황에 대한 고려

(Assumption by Others; AO)가 나타날 수 있다고 보았다. 채점자가 채점 과정에서 확보한 근거와 논거를 바탕으로 결정한 점수(score)는 채점자가 채점 과정을 통해 수험자의 응답에 관한 주장으로 볼 수 있으며, 본 연구에서 사용한 채점 척도를 따라 총체적 평가 준거인 전반적 수행에 대한 점수(S1)와 발음 점수(S2), 어휘와 문법 점수(S3), 담화 점수(S4)로 부호화는 체계를 구안하였다([그림 Ⅳ-4] 참조).



[그림 Ⅳ-4] 채점 과정 보고 자료 분석을 위한 논증 요소 코드의 체계

[그림 Ⅳ-4]는 채점 과정 보고의 분석을 위한 논증 요소 코드의 체계를 나타

낸 도식이다. 점수 결정을 위한 논증으로서의 채점 과정은 수험자 응답으로부터 형성한 점수 결정의 근거와 결정한 점수를 바탕으로 하는 주장으로 이루어져 있다. [그림 Ⅳ-4]에서 수험자 과제 수행으로부터 형성한 근거와 가정은 채점자가 주장하고자 하는 점수 결정의 원인으로서 작용하며 채점 과정을 구성한다. 채점 과정 분석에서는 이러한 논증 도식을 구성하는 요소들을 적용하여 채점 과정 보고 발화를 코딩하고, 이를 바탕으로 채점 과정 구성에 나타난 특징을 파악하고자 한다.

채점자 발화에 대한 논증 요소 코딩의 절차는 문항별로 최종 점수(S1~S4)를 먼저 코딩하고, 그에 관한 근거(DP, DL, DD)와 가정(AT, AR, AE, AO)을 코딩하는 순서로 진행하였다.[46] 평가 준거 중 '전반적 수행'은 다른 준거들의 종합적인 채점 기준으로서 나머지 분석적 채점 영역인 발음과 어휘·문법, 담화에 관한 판단을 포함한다는 점에서 도식에서 가장 상위에 배치하였으며, 그 하위에는 분석적 평가 준거에 대한 주장을 배치하였다. 그리고 각 분석적 채점 영역에 대한 청취 중 및 청취 후 보고에 나타난 근거와 가정을 배치하였다. 채점자들이 보고한 내용에서 발음, 어휘·문법, 담화에 관하여 언급한 근거와 가정이 반복해서 나타날 경우에는 빈도를 표시하도록 하였다. 구체적인 분석 코드 및 채점 과정 보고 사례는 <표 26>으로 제시하였다.

46) 채점 과정 보고에 대한 코딩은 채점자가 보고한 모든 발화 가운데 논증으로서의 성격을 고려하여 최종 점수를 통하여 주장하고자 하는 바와 그에 관한 근거와 가정이 나타난 발화를 대상으로 이루어졌다. 코딩을 하지 않은 채점 과정 보고로는 채점자가 같은 말을 반복한 발화(예:"다 1점이네요", "그래도 2점...까지는 줄 수 있을 거 같습니다."), 의미가 불분명한 발화(예:"뭐랄까", "어 뭐라고 해야 하지"), 점수 결정을 고민하는 발화(예:"3점? 4점?", "발음은 1점을 줘야 할지 2점을 줘야 할지 어 망설여지는데")가 있었는데, 이 중에서 점수 결정을 고민하는 발화는 점수 결정에 대한 채점자의 심리적인 갈등 상황을 나타낸다는 점에서 그 원인과 배경에 대한 탐구가 필요한 발화라고 볼 수 있다. 이러한 고민하는 내용을 포함하는 발화는 채점자의 최종적인 점수 결정의 원인이나 가정이 아니며, 고민을 유발하는 평가 요소에 관한 논증이라는 점에서 또 다른 층위의 분석을 요구한다. 이와 관련하여 말하기 평가의 채점 과정에 관한 후속 연구를 통해 점수 결정 고민 발화를 중심으로 하는 탐구가 이루어져야 할 것이다.

<표 Ⅳ-1> 분석 코드 및 채점 과정 보고 사례

코드 분류	코드	채점 과정 보고 사례
점수	S1	“전반적인 수행 점수도 2점 주겠다.” “이 친구 전반적 수행은 탁월 5점 줄 수 있고요.”
	S2	“발음을 5점” “발음에 문제가 없었으니깐 5점 주고요”
	S3	“어휘 문법 점수 3점 주겠다.” “네, 어휘 문법도 5점 줄 수 있을 것 같아요”
	S4	“담화 구성은 4점을 주자” “역시 담화 구성도 5점.”
근거	DP	“아 발음이 별로 안 좋다” “굉장히 유창하게 말을 이어가네.” “발음은 어 수정해야 될 심각한 문제들이 반복해서 나타나고 있고”
	DL	“비행기가 내려가? 이륙하다고 말해야 되고” “각각의 표현을 이해할 수는 있지만” “계속 조사가 틀리고 있어.”
	DD	“전반적으로 담화의 응집성이 떨어져서” “내용이 좀 부실하네” “자료에 없는 자기 생각 자기가 알고 있는 배경지식을 포함해서 이야기하기도 했고”
가정	AT	“어휘 문법 측면에서는 그래프 설명에 필요한 표현이 거의 나타나지 않았어요.” “문제를 발생한 문제를 먼저 얘기하고 그 문제를 해결하기 위해서 이런 방법들을 해야 한다 이렇게 넘어갔어야 되는데 ” “음, 일단. 뭐 하고 싶은 직업을 결정하지 못해 고민하는 친구에 대한 조언, 과제를 잘 다루고 있죠.”
	AR	“표를 보면.. 몇 가지 오류가 나타나지만.. 주제와 관련된 어휘.. ” “응답이 과제와 거의 관련이 없다고는 볼 수 없지 제한적으로 어느 정도 일부 다루고 있고” “발화가 덜 명확한 편이고 수정 주저하는 일 종종 나타났고 노력을 해야 될 부분이라고 보기 때문에”
	AE	“왜냐하면 문장에 오류가 나타난 건 아닌데 굉장히 2,3급 수준의 단순한 표현으로만 문장을 다 구성했어요.” “이 학생은 4급이 아니야.” “그 사소한 문제들 그 중국 화자 특유의 그런 발음의 어색한 부분들이 있는데”
	AO	“2번과 같이” “확실히 고급 어휘가 나오기 시작하면서 앞에 1,2,3번 문제보다는 틀리는 발음이 많이 나왔네요” “사실 뭐, 이 친구만 틀리는 것도 아니고”

본 연구에서는 채점 과정 보고에 대한 분석 방법과 논증 요소를 바탕으로 구성한 코드를 적용하여 ‘한국어 말하기 능력 시험’에서 채점자들의 채점 과정

보고 발화를 분석한다. 채점 과정 보고에 대한 분석은 말하기 평가 문항별 특징과, 평가 결과 분석에서 나타난 특징을 중심으로 이루어진다. 이러한 접근을 통하여 구성형 응답 문항으로 이루어진 말하기 평가의 채점 과정에서 채점자의 채점 수행 양상과 채점자 간 차이가 일어나는 원인을 확인할 수 있었다.

2. 채점 과정 보고 자료 분석

말하기 평가 채점 과정을 실증적으로 알아보기 위하여 논증 요소에 따라 채점 과정 보고를 부호화하고, 문항 및 채점자별 양상을 확인하는 분석을 실시하였다. 말하기 평가의 채점 과정은 평가 결과를 도출하기까지의 채점자의 연쇄적인 인지적 처리로 이루어진다. 이 과정에서 채점자의 주관적인 채점 경향의 영향 아래에서 채점이 이루어지며, 이는 채점자가 수험자의 과제 응답에서 어떤 정보에 주목하며, 평가 준거별로 어떤 근거와 가정을 구성하여 점수를 결정하는지의 양상을 통하여 나타난다. 따라서 평가 결과의 원인을 제공하는 채점 과정을 파악하는 것은 평가 결과의 타당성을 확보할 수 있는 정보를 얻는 방법이라고 할 수 있다.

채점 과정 보고의 분석은 분석의 체계성과 객관성을 확보하기 위하여 '한국어 말하기 능력 시험'의 체제와 양적 분석 결과를 바탕으로 이루어졌다. 채점 과정 분석의 결과는 평가 문항 및 채점자별 채점 과정 보고 사례들 중에 평균적 사례를 중심으로 상위 및 하위에 해당하는 사례의 채점 과정 보고에 나타난 논증 요소를 비교하는 방법으로 제시하였다.

1) 분석 대상 자료의 특성과 분석 과정

(1) 분석 대상 자료의 특성

채점 과정 보고를 분석하기 위하여 먼저 분석 대상 자료의 기본적인 특성을 확인하였다. 본 연구의 분석 대상 자료는 채점자 13인의 말하기 평가 채점 과정 보고 담화를 녹음한 자료이다. 전체 자료에서 채점자들은 평균 188.92분 동안 12명의 수험자가 4개의 구성형 문항에 응답한 발화를 채점하였다. 채점 과

정의 분석을 위해 녹음한 자료를 전사한 결과, 발화를 의미 단위로 분절하였을 때 채점자별 평균 보고량은 A4용지를 기준으로 평균 49.69매였다. 그리고 채점자별로 채점 과정 보고에 사용한 단어의 수는 평균 7,704.54개였다.[47]

<표 Ⅳ-2> 채점자별 채점 과정 보고 자료의 특성

채점자 ID	시간 (분)	분절자료 (쪽)	총 단어 (개)
R01	137	26	4071
R02	150	28	3908
R03	127	35	4919
R04	243	66	12916
R05	240	62	11919
R06	213	64	8652
R07	137	18	3090
R08	199	36	8363
R09	215	52	6077
R10	260	93	12209
R11	172	76	9231
R12	197	50	8269
R13	166	40	6535
평균	188.92	49.69	7,704.54

<표 Ⅳ-2>은 전체 채점자의 채점 과정 보고 자료의 특성을 정리한 표이다. 자료를 살펴보면, 채점자 중에 R01과 R02, R07은 보고의 양이 평균에 비해 상대적으로 적은 편(총 단어 수 평균=3689.67개)으로 나타났으며, 특히 R07의 경우에는 발화 중에 긴 휴지가 많이 나타나, 채점 과정 보고 방식에 잘 적응하지 못한 것으로 나타났다. 채점자 R04, R05, R10은 채점 과정 보고의 양이 평균에 비하여 매우 많은 것(총 단어 수 평균 12,348개)으로 나타났으며, 보고 형식에 잘 적응하였으나 경제성이 다소 부족한 채점자들인 것으로 나타났다.

(2) 코딩의 절차와 방법

채점 과정 보고의 분석은 형식화한 보고 자료를 논증 요소에 따라 코딩하는

47) 단어 수에는 채점자 구분 기호, 수험자별 구분 기호, 문항별 구분 기호, 청취 중 보고와 청취 후 보고의 구분 기호로 사용한 384개 표시가 포함되어 있다.

것으로 이루어졌다. 코딩은 Excel 시트에 채점 과정 보고별로 주장, 근거, 가정의 종류에 따라 코드를 부여하는 것으로 이루어졌다(<표 28>참조). 코딩은 채점자가 채점 과정에서 수험자의 응답 특성이나 채점자 판단을 근거로 삼아 채점 척도와 과제, 평가 관련 이전 신념, 그리고 채점을 수행하면서 갖게 된 새로운 신념 등의 가정을 바탕으로 0~5점의 준거별 점수 결정을 주장하는 과정을 확인하기 위한 목적으로 이루어졌다. 채점 과정 보고 자료의 코딩에는 3명의 한국어 교사[48]가 참여하였으며, 코드 목록 점검 및 코딩 일치도 확인을 위한 3차에 걸친 코딩 연습 및 회의를 한 후에 코더들의 코드에 대한 이해와 코딩의 일관성을 확인한 후에 본 코딩을 수행하였다.

<표 Ⅳ-3> 채점 과정 보고의 코딩 사례(R05-E12, 문항1)

순서	논증 요소	채점 과정 보고
1		#R05121
2		#R051211
3	DP1	유창하다
4	DL1	사용하는 어휘는... 약간 중급스럽다, 중급 이하의 표현을 쓰고…있다
5		#R051212
6	DD1	이 학생은 전반적으로 들었을 때는, 과제와, 과제와 관련해서 적절하게 과제를 어 수행하고 있어 보입니다
7	DL2	그런데 전반적인 표현 자체가 그렇게 고급 표현을 사용하는 게 아니라 중급 이하의 표현을 사용해서 유창하게 이야기를 하고 있는 것으로 봐서
8		4점까지 주기에는 조금 어려움이 있을 것 같고
9	DL3	어, 그 다음에 끝나는 부분도 갑자기 완결되는 느낌이 조금 있어서
10	S1:3	음 전반적 수행은 3점 정도를 주는 것이 적절해 보입니다.
11	DP2	발음은 상당히 자연스럽고 어 매끄럽습니다
12	DP3	반복도 별로 없었고 주저하는 일도 없고
13	S2:4	그래서 사실 발음 같은 경우에는 4점 정도를 줘도 큰 문제가 없을 것 같습니다
14	AT1	다음 어휘와 문법은 어 내용을 잘 들어보면 사실 주제와 관련한 이야기를 하고 있기는 하지만
15	DL4	그 주제와 관련된 이야기를 할 때 사용하는 어휘 수준이 그렇게 높지 않고
16	DL5	어 중급 이하의 표현들을 주로 사용하고 있습니다.
17	DL6	그렇지만 어떤 복잡한 문형 자체를 사용하고 있기 때문에
18		4급을 주는 것은 어려울 것 같고
19		이 정도 수준이면 3급을 주는 것이 주는 것이 적절하, 아니
20	S3:3	3점을 주는 것이 적절해 보입니다.

48) 코딩에 참여한 3명은 모두 한국어 교사로서 한국어교육학 전공 석사학위 소지자였다.

<표 Ⅳ-3>에서 가장 좌측에 있는 '순서'열은 분절된 채점 과정 보고의 순서를 나타낸다. '순서'열에서 1, 2, 5에는 채점 과정 보고를 형식적으로 구분하기 위한 기호가 들어가 있다. 다음으로 '논증 요소'열은 채점 과정 보고의 내용에 따라 어떤 논증 요소인지를 나타낸다. '논증 요소'열에서 순서열 18, 19는 해당하는 발화의 내용이 점수 결정을 고민하는 내용이기 때문에 채점의 근거로 보기에 어려움이 있으므로 논증 요소를 부여하지 않았다. 코더들은 각 수험자의 문항별 채점 과정 보고에 대해 논증 요소로 코딩하면서 한 문제에서 나타난 같은 종류의 논증 요소에 숫자를 붙여 논증 요소 고려의 양상을 파악할 수 있도록 하였다. 가장 우측에 있는 '채점 과정 보고'열은 분절된 채점 과정 보고 발화의 내용을 확인할 수 있다.

제시한 채점 과정 보고를 간단하게 나타낼 수 있는 방법은 아래와 같이 채점 과정을 나타내는 논증 요소 코드를 연결하여 채점 과정을 나타내는 것이다. 예를 들어 채점자 R05의 채점 과정 보고를 논증 요소에 따라 코딩하였을 때 DP1-DL1-DD1-DL2-DL3-S1:3-DP2-DP3-S2:4-AT1-DL4-DL5-S3:3의 순서로 이루어졌다면 채점 과정은 수험자의 응답에서 발음에 관한 어떤 특징 또는 판단에 관한 보고로 시작(DP1)하여 어휘·문법(DL1), 담화(DD1), 다시 어휘·문법(DL2, DL3)에 관하여 보고한 후에 '전반적 수행'(S1)에 관하여 3점으로 점수를 결정하였으며, 다시 발음에 관하여 두 가지 측면을 보고(DP2, DP3)한 후에 '발음'에(S2)에 관해 채점하고, 평가 과제에 대한 고려(AT1)와 어휘 문법에 대한 고려(DL4~6)를 한 후에 어휘·문법(S3)에 점수를 결정하였다는 의미인 것이다.

논증 요소 코드로서 채점 과정을 나타내는 것의 장점은 채점 과정을 비교할 때 직관적으로 그 차이를 알아 볼 수 있도록 한다는 점이다. 본 연구에서 수집한 전체 채점 사례는 총 624개(수험자 12명×4개 문항×채점자 13명)이며, 이 중에서 채점 과정 보고 분석의 대상으로 평가 문항 유형별 응답 수준에 따른 대표 사례와 채점자 영향이 나타난 사례를 중심으로 채점 과정 보고 발화에 대한 논증 요소 코딩을 실시하였다. 이러한 접근은 직관적으로 채점 과정을 파악하는 데 도움이 될 수 있을 것으로 판단하였다. 이후의 채점 과정 보고의 분석 결과는 채점 과정에 관한 직관적 이해와 상세한 이해를 도모하기 위하여 논증 요소로 코딩한 것과 실제 담화 자료를 함께 제시하였다.

(3) 자료 분석의 방법과 절차

논증 요소에 따라 코딩한 채점 과정 보고 자료의 분석은 평가의 4개 문항 유형별 채점 과정 양상을 확인하는 접근과 채점자 영향의 원인을 확인하기 위한 접근으로 이루어졌다. 먼저 평가 문항 유형별 채점 과정 보고의 분석은 채점자가 평가의 조건에 따라서 채점 과정에서 어떠한 인지적인 차이를 나타내었는지를 알아보기 위한 것이며, 평균 점수를 기준으로 4개 문항별로 최상위-중위-최하위 수준에 해당하는 수험자에 대한 전체 채점자의 채점 과정 보고 양상을 논증 요소를 바탕으로 확인하고, 채점 척도 사용을 기준으로 유형을 구분하여 채점 과정의 특징을 비교하는 것으로 이루어졌다(<표 Ⅳ-4> 참조).

<표 Ⅳ-4> 문항 유형별 채점 과정 보고의 분석 대상

수준	문항 유형			
	경험 말하기	조언하는 말하기	도표 보고 설명하기	기사 읽고 문제와 해결 방안 이야기하기
최상위	E03	E03	E04	E03
중위	E10	E02	E12	E01
최하위	E09	E05	E05	E09

<표 Ⅳ-4>에서 제시한 문항 유형별 채점 과정 보고 분석의 대상은 각 문항별로 평균 점수에 가장 가까운 점수를 받은 수험자를 중위, 가장 높은 평균 점수를 받은 수험자를 최상위, 반대로 가장 낮은 평균 점수를 받은 수험자를 최하위로 선정한 결과를 정리한 것이다.

다음으로 평가 결과의 오차를 발생시키는 원인인 채점자 영향이 각 유형에 어떠한 채점 과정의 양상으로 나타나는지를 알아보기 위하여 채점자 영향에 따른 채점 과정 분석을 실시한다. 채점 과정의 분석은 연구에 참여한 채점자로부터 나타난 채점자 영향 가운데 주관적인 채점 경향성(엄격성/관대성)과 특정 척도에 관한 집중 경향성, 채점 척도나 준거를 구별하지 않거나 영향 관계를 고려하지 않는 경향성(무작위성 및 후광성)을 기준으로 이루어졌다(<표 Ⅳ-5> 참조).

<표 Ⅳ-5> 채점자 영향별 채점 과정 보고의 분석 대상

채점자 영향			
주관적 채점 경향		특정 점수 선택 경향	평가 준거의 주관적 사용 경향
더 엄격한 경향	더 관대한 경향		
R05(1번 문항) R08(1번 문항)	R11(1번 문항) R13(1번 문항)	R05(3점)	R06(제한적 사용)
-	R11(2번 문항)	R13(5점)	R02(무선적 사용)
R13(3번 문항)	R07(3번 문항)	R01(척도별 간격 넓음)	R08(어휘·문법의 편향 채점)
R07(4번 문항) R11(4번 문항)	-	R13(척도별 간격 좁음)	R04(담화의 편향 채점)

<표 Ⅳ-5>에서 제시한 채점자 영향에 따른 분석 대상 사례들은 MFRM 분석을 통하여 평가 결과에서 채점자의 개입으로 인한 특정한 채점 경향이 나타난 경우에 해당한다. 이들에 대한 채점 과정 보고 자료의 분석은 채점자 영향이 나타난 채점 사례에서 다른 채점자가 채점한 평균적인 채점 사례와의 비교를 바탕으로 이루어졌다.

2) 분석 결과

본고에서는 말하기 평가의 채점 과정은 점수를 결정하기 위한 채점자의 논증이라는 관점에서 채점 과정을 보고한 발화에 대해 논증 요소로 코딩을 한 후에 그 양상을 분석하였다. 연구에 참여한 채점자는 수험자의 과제 수행 발화를 청취하면서부터 최종적인 점수를 부여하기까지의 모든 떠오르는 생각을 보고하였으며, 이는 채점 과정을 직접적으로 나타낸 것이다.

채점 과정 보고에 대한 분석은 채점 과정을 검토하여 얻을 수 있는 두 가지 효용을 중심으로 하였다. 먼저 구성형 응답 문항으로 이루어진 말하기 평가에서 평가 문항의 유형에 따라 채점 양상을 파악하고 그 특징을 확인하였다. 다음으로는 말하기 평가 결과 분석에서 확인한 채점자 영향이 나타난 사례와 관련하여 어떤 채점 과정을 통해 그러한 영향이 나타난 것인지를 살폈다. 이를 위해 같은 수험자에 대한 다른 채점자의 채점 과정에 나타난 논증을 비교하여 측정의 오류를 유발하는 채점 과정의 요인을 규명하였다.

(1) 문항 유형에 따른 채점 과정 보고의 차이

말하기 평가의 문항 유형에 따른 채점 과정 보고에 대한 분석은 '한국어 말하기 능력 시험'의 4개 문항을 기준으로 먼저 중위 수준의 수험자에 대한 채점자들의 채점 과정 보고에 나타난 특징을 살펴본 후에 최상위 수준의 수험자와 최하위 수준의 수험자에 대한 채점 과정 보고를 살펴보는 것으로 이루어졌다. 중위 수준의 사례는 평균적인 수준, 즉 해당 문항 유형의 대표적인 사례로서 먼저 검토가 이루어졌다. 최상위·최하위 수준은 해당 문항 유형에서 나타날 수 있는 양 극단에 해당하는 사례라는 점에서 중위 수준보다는 응답의 특성이 명료할 수 있으므로 덜 복잡한 과정으로 채점이 이루어졌을 것으로 예상할 수 있다. 이러한 비교를 통하여 채점 과정 보고 양상의 수준별 차이를 가늠할 수 있을 것으로 판단하여 중위 수준 다음으로 최상위와 최하위 사례에 대한 채점 과정 보고 분석 결과를 제시하였다.

① '경험 말하기' 문항의 채점 과정 보고

'한국어 말하기 능력 시험'의 첫 번째 문항은 한국어를 학습한 경험에 대해 말하는 것이었으며, 수험자들은 자신의 한국어 학습 과정에서 만난 문제에 대한 경험을 바탕으로 응답을 구성하였다. 이 문항에서 중위, 즉 평균적인 수준으로 평가 받은 수험자는 E10이었으며, E03은 최상위 수준, E09는 최하위 수준으로 나타났다(<표 Ⅳ-6> 참조).

<표 Ⅳ-6> '경험 말하기' 문항에서 최상위-중위-최하위 수험자의 평가 준거별 평균 점수

수험자		평가 준거			
		전반적 수행	발음	어휘와 문법	담화
최상위	E03	4.85	4.92	4.92	4.62
중위	E10	4.23	4.15	4.08	4.38
최하위	E09	2.62	2.54	2.54	2.69

● 중위 수준

먼저 평가 결과에서 중위 수준으로 나타난 E10에 대한 채점자 13인의 채점 과정 보고에서는 채점 척도에 제시된 순서를 따라 총체적 평가 준거를 채점한 후에 분석적 평가 준거를 채점하는 순차형 채점자(R02, R05~R08, R10)와 분석적 평가 준거를 먼저 채점한 후에 종합적 준거를 채점하는 종합형 채점자(R01, R04, R09, R11~R13)가 나타났다.

<표 Ⅳ-7> '경험 말하기' 문항에서 중위 수준 수험자(E10)에 대한 채점 과정

채점자	채점 과정																						유형
R01	DL 1	DL 2	DP 1	DP 2	DL 3	DP 3	S2: 4	DL 4	S3: 4	S4: 5	AO 1	S1: 4											종합형
R02	DP 1	DL 1	DL 2	AR 1	DL 2	S1: 4	DP 2	S2: 4	AE 1	DL 3	S3: 3	DD 1	S4: 4										순차형
R03	DL 1	DD 1	S1: 5	AE 1	DL 2	S3: 5	DD 2	S4: 5	DP 1	AE 2	S2: 4												무작위형
R04	DP 1	DP 2	DL 1	DP 3	DP 4	S2: 4	DP 5	DL 2	DL 3	AE 1	DD 1	S3: 3	DD 2	DD 3	AR 1	S4: 4	DL 3	AT 1	AR 2	AR 3	S1: 3		종합형
R05	DL 1	DL 2	DP 1	DP 2	AT 1	DL 3	DD 1	DD 2	S1: 4	AE 1	AR 1	S2: 4	DL 3	DL 4	AE 2	DL 5	S3: 4	DD 3	DD 4	AR 2	S4: 4		순차형
R06	DL 1	DL 2	DL 3	DL 4	DD 1	DP 1	S1: 5	DP 2	DP 3	S2: 4	AT	DL 5	DL 6	S3: 5	DD 2	DD 3	S4: 5						순차형
R07	DD 1	DD 2	DL 1	S1: 4	DP 1	S2: 3	DL 2	S3: 3	DD 3	AR 1	S4: 4												순차형
R08	DL 1	DL 2	DL 3	DP 1	AT 1	DD 1	S1: 5	DP 2	DP 3	S2: 5	S3: 5	DL 4	DD 2	DD 3	AR 1	S4: 5							순차형
R09	DL 1	DL 2	DD 1	DP 1	DP 2	DP 3	S2: 4	DL 3	DD 2	S3: 3	DL 4	AR 1	DD 3	DD 4	AR 2	S4: 3	DD 5	DD 6	S1: 3				종합형
R10	DL 1	DL 2	DL 3	DL 4	DD 1-7	S4: 4	DL 5	DL 6	AE 1	S3: 4	DP 1	DP 2	DP 3	AE 2	DP 4	S2: 5	DL 7	DP 4	DD 8	S1: 4			순차형*[49]
R11	DL 1	DL 2	DL 3	DL 4	DL 5	DD 1	DL 6	DD 2	DP 1	DP 2	DP 3	DP 4	S2: 4	AR 1	DL 7	S3: 4	AR 2	S4: 4	AR 3	S1: 4			종합형
R12	DP 1	DP 2	DP 3	DP 4	DP 5	S2: 4	DL 1	DL 2	DL 3	S3: 5	DD 1	DD 2	AR 1	S4: 5	DP 6	AR 2	DP 7	S1: 5					종합형
R13	DP 1	DL 1	DP 2	S2: 5	DL 2	DL 3	S3: 5	DD 1	S4: 5	DD 2	DP 3	S1: 5											종합형

49) 채점 과정 유형 분류에서 '순차형*'은 채점을 하는 순서가 '순차형'의 반대, 즉 채점 척도에서 가장 마지막에 있는 준거부터 채점을 하는 것을 가리킨다. 이는 가장 처음에 제시된 준거부터 채점을 하는 '순차형'과 방향은 반대이지만, 척도 순서를 따른다는 점에서 '순차형'으로 분류하였다.

<표 Ⅳ-7>은 수험자 E10의 1번 문항 응답에 대한 전체 채점자들의 채점 과정 보고를 논증 요소로 코딩한 것을 정리한 표이다. 이 중에서 대표적인 사례로서 척도 사용상의 차이가 나타난 R05와 R11의 채점 과정 보고를 비교하였다. 순차형 채점을 한 R05와 종합형 채점을 한 R11은 모든 평가 준거에 대해 동일하게 모두 4점(우수)을 부여하였다. 이와 관련하여 두 채점자의 채점 과정 보고 상의 차이를 확인하여 보았다(#R05101, #R11101).

순차형 채점을 한 R05와 종합형 채점을 한 R11의 채점 과정 보고(#R05101, #R10101)에서는 보고의 양과 판단을 위한 근거의 종류의 측면에서 차이를 확인할 수 있다. 두 채점자의 분절된 채점 과정 보고의 수는 23개로 동일하였지만, 수험자 응답 청취 중 보고의 횟수를 비교하였을 때 R05는 3회, R11은 7회로 차이가 있었다. 평가 준거별 점수 결정의 근거를 비교해 보면, 전반적 수행에 대한 채점 과정에서 R05는 여러 가지 응답의 특징을 고려하는 가운데 담화적 특징을 중심으로 채점을 한 것과 달리 R11은 채점 척도에 기술된 기준을 참고하여 점수를 결정하였다. 다음으로 발음에 대해서 R05는 자신의 교육 경험(AE)과 채점 척도(AR)를 바탕으로 점수를 결정하였다면, R11은 응답에서 나타난 특징을 바탕으로 점수를 부여하였다. 어휘·문법에 대한 채점에서 R05는 R11에 비해 다양한 발화의 특징과 교육 경험을 고려하는 경향을 보였으며, 담화를 채점할 때에는 R05와 R11 모두 응답의 특징과 채점 척도를 바탕으로 점수를 부여하는 것으로 나타났다.

	#R05101
	#R051011
DL1	말하기꽈
DL2	조사 오류
DP1	어 발화 속도가 엄청 빠르네
	#R051012
DP2	이 학생은 전반적으로 발화 속도가 상당히 빠르고 유창하게 이야기를 아... 하고 있습니다
AT1	음 그래서 어 이 1번 과제에 대한 전반적인 그 수행 자체는… 과제를 적절하게 이해하고 있고
DL3	다음에 물론 약간의 실수가 있지만
DD1	구어적, 문어적, 이런 표현들을 어 본인의 말로 쉽게 풀어 설명하고 있지만

DD2	내용이 명료하고 유창하게 응집성 있는 담화를 구성하…고 있다고 보여져서
S1:4	전반적 수행은 4점 정도를 줘도 괜찮을 것 같습니다
AE1	발음도 어 약간의 그 중국인들이 일으키는 모어의 어 특징이 나타나기는 하지만 약간 있지만
AR1	이해를 위해서 그렇게 큰 노력을 하지 않아도 되는 어 발음을 어 발음으로 이야기를 했습니다
S2:4	어… 하지만 5점을 주기에는 음. 약간의 어 아쉬움이 있고 그렇기 때문에 발음도..한 우수? 이 정도를 줘도 4점! 4점 정도를 줘도 괜찮을 거 같습니다. 5점은 조금 무리인 거 같아요.
DL3	다음에 어휘와 문법은 음, 아까 뭐 그, 말하기과라고 얘기한다거나 그.. 사용 면에서 조금 오류가 나타났습니다
DL4	그 다음에 어 좀 더 이제 구어적일 때 사용한다, 문어적일 때 사용한다라는 표현이 있었지만 그것을 책에 있을 때, 말을 할 때, 뭐 요런 정도로 수준을…
AE2	약간 좀 고급 어휘를 구사하는 것은 아니었고 다음에 그래서 난이도가 아주 탁월한 정도는 아니지만
DL5	그래도 자신이 하고자 하는 내용을 어 적절한 주제와 관련된 내용으로 그 대안 어휘를 사용해서 말하고 있다는 점에서는 나쁘지 않았고 그 내용을 적절하게 표현하고 있어서어
S3:4	여기도 4점을 줘도 큰 무리는 없을 거 같아 보입니다
DD3	그 다음에 담화 구성도 음…전개 과정은 우선 한국어를 배울 때 뭐가 어려웠다, 그 어려운 내용들을 이야기하고 그에 대한 구체적인 예도 이야기하면서 이야기를 하고 있기 때문에
DD4	어 100%, 100% 완결된 구조로 잘 갖추어서 말하고 있다고 보기에는 조금 어렵지만
AR2	그래도 대체적으로 완결된 구조와 문장으로 음… 이야기하고 있다고 보여져서
S4:4	담화 구성도 한 4점 정도 줘도 나쁘지 않을 것 같습니다.

	#R11101
	#R111011
DL1	책에서 쓰는 말, 구어..
DL2	한자어
DL3	쓸 때
DL4	구어적 표현, 문어적 표현
DL5	문법
DD1	같은 얘기를 지금 첫 번째 두 번째로 나눠서 얘기하고 있네
DL6	아 어휘가 구어적 표현
	#R111012

DD2	어휘 구어적 표현이랑 문법 구어적 표현을 얘기하고 싶었던 건가?
DP1	일단 발화가 명확하긴 했지
DP2	반복과 수정하는 일 거의 없었고
DP3	발음 문제 좀 있었지만
DP4	이해 불가능한 수준은 아니었으니깐
S2:4	발음은 4점 발음은 4점을 주고
AR1	그 다음에 어휘 문법을 보면 대체로 주제 관련 복잡한 문형 사용을 적절하게 표현했고 몇 가지 오류가 나타났지만 의미 이해..
DL7	뭐 구어 같은 건 생각 안 나니까 말할 때 쓰는 문법 이렇게 자기만의 회피 전략도 괜찮았던 거 같고
S3:4	4점을 주고
AR2	그 다음에 담화 구성 구성을 보면 대체적으로 논리적임 완결한 구조… 문장의 연결에 짜임새가 있음
S4:4	이것도 4점
AR3	그 다음에 전반적 수행을 보면 응답에서 과제 관련해서 적절한 내용을 다루고 있었고 문법에 실수 있지만 유창했고
S1:4	4점

● 최상위 수준

다음으로 ‘경험 말하기’ 문항에서의 최상위 수준에 해당하는 수험자 E03에 대한 채점자들의 채점 과정 보고에 대해 논증 요소로 코딩한 것을 살펴보았다. E03의 채점 과정 점수를 부여한 척도의 순서를 보면 다수의 채점자들은 순차형 채점(R01~R03, R05~R08, R10)을 한 것으로 나타났으며, 그 밖에도 종합형 채점(R04, R09, R11, R12)과 무작위형 채점(R13)을 한 경우도 나타났다(<표 Ⅳ-8> 참조).

<표 Ⅳ-8> '경험 말하기' 문항에서 최상위 수준 수험자(E03)에 대한 채점 과정

채점자	채점 과정																						유형
R01	DL 1	DP 1	DL 2	DD 1	S1: 4	S2: 5	S3: 5	S4: 5															순차형
R02	AT 1	AE 1	S1: 5	DP 1	AR 1	S2: 5	DL 1	DL 2	S3: 5	DD 1	DD 2	DD 3	AR 2	DD 4	S4: 4								순차형
R03	AE 1	AE 2	DP 1	S1: 5	DP 2	S2: 5	DL 1	S3: 5	DD 1	AO 1	DD 2	DD 3	S4: 4										순차형
R04	DP 1	DL 1	DP 2	DP 3	DP 4	AR 1	S2: 5	DL 2	DL 3	DL 4	S3: 5	AR 2	DD 1	DD 2	S4: 5	DD 3	AR 3	S1: 5					종합형
R05	DP 1	DP 2	DD 1	AT 1	DD 2	DD 3	S1: 4	DP 3	DP 4	AE 1	S2: 4	AR 1	AE 2	S3: 4	DD 4	DD 5	DD 6	S4: 4					순차형
R06	AO 1	AT 1	DD 1	DD 2	S1: 5	DP 1	S2: 5	DL 1	DL 3	S3: 5	AR 1	S4: 5											순차형
R07	DP 1	DL 1	DD 1	DL 2	S1: 5	DP 2	S2: 5	AR 1	DL 3	S3: 5	AR 2	S4: 5											순차형
R08	AO 1	DL 3	DP 2	AT 1	DL 4	S1: 5	DP 2	AO 2	DP 3-5	S2: 5	AE 2	AE 3	S3: 5	DL 5	AR 1	AR 2	AT 2	DD 1	DD 2	S4: 5			순차형
R09	DP 1	AE 1	DP 2	AR 1	S2: 5	AE 2	DL 1	DL 2	DL 3	S3: 5	DD 1	DD 2	DD 3	S4: 4	DD 4	DD 5	AR 2	DD 6	DD 7	DD 8	S1: 5		종합형
R10	DL 1-3	DP 1	DD 1	DL 4	DP 2	AE 1	DP 3	AO 1	DD 2	DD 3	AR 1	S4: 5	DL 4	AE 2	DL 5	S3: 5	DP 3	DP 4	AE 3	S2: 5	S1: 5		순차형*
R11	DL 1-5	AE 1	DL 6	DD 1	DP 1	AR 1	S2: 5	AR 2	S3: 5	DD 2	DD 3	AR 3	DD 4	DD 5	DD 6	S4: 5	AR 4	DL 7	AR 5	S1: 5			종합형
R12	DP 1	DL 1	DP 2	DP 3	AR 1	DP 4	S2: 5	DL 2	DL 3	DL 4	S3: 5	AR 2	DD 1	S4: 4	DL 5	DP 6	DD 2	S1: 5					종합형
R13	DP 1	DL 1	DL 2	DL 3	DL 4	DP 2	DP 3	S3: 5	DD 1	DL 5	DD 2	S4: 5	DP 4	S2: 5	S1: 5								무작위형

<표 33>은 수험자 E03의 '경험 말하기' 문항에 대한 전체 채점자의 채점 과정 보고를 논증 요소로 코딩한 것을 정리한 표이다. 채점 사례 중에서 채점자 R06과 R11의 채점 과정 보고를 비교하여 보았다(#R06031, #R11031). R06과 R11은 모든 채점 척도에서 5점을 부여하였는데, R06은 순차적 채점을 하였고, R11은 종합형 채점을 한 것으로 나타났다.

E03의 '경험 말하기' 문항 응답에 대한 채점자 R06과 R11의 채점 과정 보고에서 순차형으로 채점한 R06은 응답 청취 중에는 보고를 하지 않았으며, 총 보고 발화량에서도 분절된 발화의 수를 기준으로 보았을 때 R11에 비해 적은 것으로 나타났다(R06: 12개, R11: 23개). 순차형으로 채점을 한 R06은 종합형으로 채점을 한 R11에 비해 채점 척도를 가정(AR)한 점수 결정 횟수가 적었으며, R06의 점수 결정의 근거에서 '특별히 ~한 것은 없다'와 같이 인상적인 접근이 나타난 것과 달리 R11에서는 구체적인 정보에 대해 채점 기준을 바탕으로 점검하는 양상이 나타났다.

	#R06031
	#R060311
	#R060312
AO1	이 학생은 특별하게 잘못된 부분을 얘기할 부분은 없는 거 같습니다
AT1	일단 과제에도 다 충족하고
DD1	세부 내용도 잘 표현을 했고
DD2	담화의 응집성도 되게 좋았다고 생각해서
S1:5	전반적인 수행을 5점을 주고요
DP1	발음도 명확하고 자연스럽고 특별하게 거슬리는 발음의 문제가 없었습니다
S2:5	그래서 발음도 5점
DL1	비슷비슷하다 뭐니 뭐니 해도 이런 고급에서 사용하는 고급 어휘들을
DL3	그 상황에 맞게 잘 사용을 했었던 것 같아서
S3:5	어휘와 문법도 5점을 주고
AR1	전체적인 내용도 적절하고 그 다음에 흐름도 문장 구조도 정확하고 문장 연결도 잘 짜임새가 있었기 때문에
S4:5	담화 구성도 5점을 주도록 하겠습니다
	#R11031
	#R110311

DL1	느낀 점은
DL2	먼저 문법..
DL3	그다지 그리 어려운 점? 중 고급?
DL4	비슷비슷한 문법? 어휘가 좋고
DL5	있으나..
AE1	잘하네… 고급 학생인가?
DL6	차이점을 이해하기 어려웠습니다.. 뭐니 뭐니 해도..
	#R110312
DD1	문법…문법과 표현 두 가지에 대해서 나눠서 얘기했고
DP1	그러면 일단 보면은 발음이 명확한 편이었고
AR1	반복과 수정이 거의 없지만 머뭇거리거나 주저하는 일이 거의 없었고 자연스럽게 이어졌지
S2:5	발음은.. 5점? 발음은.. 5점
AR2	그리고 어휘 문법도 어휘 문법이 복잡한 문형을 사용해 효과적으로 표현했고 적은 오류, 의미 방해한 적 없고
S3:5	그러면 이것도 5점
DD2	그 다음에 담화를 보면 적절한 정보 포함돼 있었고
AR3	흐름이 일관됐으며 완결한 구조, 문장 연결에 짜임새가 있었음..
DD3	짜임새 있었지
DD4	담화도.. 담화 담화 표지도 적절하게 얘기했고.. 먼저, 마지막으로 이렇게
DD5	그래서 자기가 전체적으로 연결에 짜임새가 있게 문단을 만들었고
S4:5	
AR4	그 다음에 전반적으로 보면 사소한 실수.. 사소한 실수.. 사소한 실수가 딱히 없었던 거 같은데?
AR5	과제 수행...설명 잘했고 실수 거의 없고 담화 응집성이 있었음
S1:5	5점

● 최하위 수준

‘경험 말하기’ 문항에서 최하위로 나타난 수험자 E09에 대한 채점자들의 채점 과정 보고에서는 응답 청취 중에 이루어진 동시적 보고에서 발음 문제와 담화 구성의 문제에 대한 집중적인 지각이 이루어졌음이 나타났으며, 종합형 채점을 한 채점자(R01, R04, R05, R09~R13)가 8명으로 가장 많았고, 순차형 채점을 한 채점자(R02, R06~R08)와 무작위형 채점을 한 채점자(R03)가 나타났다(<표 Ⅳ-9> 참조).

수험자 E09의 ‘경험 말하기’ 문항 응답에 대한 평가 결과에서 채점자 R01과 R07은 똑같이 모든 척도에서 동일하게 3점을 부여하였으며, R01은 종합형 채점을 하였고, R07은 순차형 채점을 하였다는 차이가 있었다. 이와 관련하여 위에 제시한 #R01091과 #R07091의 채점 과정 보고를 살펴보면, 발화량에서는 큰 차이가 나타나지 않았지만, 평가 구인 중에 전반적 수행과 발음, 어휘·문법에 대한 점수 결정의 근거에 차이가 있음을 알 수 있다. R07은 점수 결정 과정에서 ‘...뭐 적절하게...’, ‘...하진 않은 편’, ‘...다고 그래야 되나’와 같이 주관적인 인상을 바탕으로 채점을 하고 있음을 나타내는 담화 표지를 사용하고 있었다.

<표 Ⅳ-9> '경험 말하기' 문항에서 최하위 수준 수험자(E09)에 대한 채점 과정

채점자	채점 과정																												유형
R01	DP 1	DP 2	DD 1	DL 1	DD 2	DD 3	DD 4	DP 3	S2: 3	DL 2	S3: 3	DD 5	S4: 3	S1: 3															종합형
R02	DP 1	DP 2	DP 3	DP 4	S1: 3	DP 5	S2: 2	AE 1	AR 1	DL 1	AR 2	S3: 3	DD 1	DD 2	DD 3	S4: 4													순차형
R03	DL 1	DL 2	DD 1	AO 1	DL 3	S3: 2	DD 2	S4: 2	DD 3	DD 4	DP 1	S2: 2	S1: 2																무작위형
R04	DP 1	DP 2	DL 1	DD 1	DP 3	AE 1	DP 4	DP 5	DD 2	S2: 2	AR 1	AE 2	DL 2	DL 3	AE 3	AR 2	AE 4	S3: 2	DD 3	DD 4	AR 3	S4: 2	AT 1	S1: 2					종합형
R05	DP 1-3	DL 1	AT 1	DL 2	DD 1	DD 2	S1: 2	DP 4-6	S2: 2	AE 1	DL 3	AE 2	AR 1	DL 4	S3: 2	AT 2	DD 3	DD 4	S4: 2	AT 3	AR 2	S1: 2							종합형
R06	DP 1	DP 2	DL 1	DP 3	DL 2	DL 3	DD 1	AT 1	DD 2	DP 4	DP 5	DD 3	S1: 2	DP 6	DP 7	S2: 2	DL 4	DL 5	DL 6	DL 7	S3: 2	DD 4	S4: 2						순차형
R07	AE 1	DP 1	DL 1	DD 1	DD 2	DD 3	DD 4	S1: 3	DP 2	DP 3	S2: 3	DL 2	DL 3	S3: 3	DD 3	S4: 3													순차형
R08	DL 1	DL 2	DL 3	DL 4	DP 1	AT 1	S1: 4	DP 2	AR 1	S2: 3	DP 3	AR 2	DL 5	S3: 3	AR 3	S4: 4													순차형
R09	DP 1	DD 1	DP 2	DP 3	S2: 2	DL 1	DL 2	DL 3	S3: 1	DD 2	S4: 1	DL 4	DD 3	S1: 1															종합형
R10	DP 1	DL 1-5	DD 1	DD 2	DP 2	AR 1	S2: 3	DL 6-8	AR 2	S3: 3	DD 3	DL 9	DD 4	DD 5	AR 3	DD 6	S4: 3	DD 7-9	DL 10	DP 3	DL 11	S1: 3							종합형
R11	DL 1	DP 1	AE 1	DP 2	DP 3	DD 1	DP 3	DL 2	DP 4	AR 1	AR 2	DP 4	S2: 2	AR 3	DL 3	S3: 2	AR 4	DD 2	DD 3	AR 5	S4: 2	AR 6	S1: 2						종합형
R12	DP 1	DD 1	DP 2	DP 3	S2: 3	DL 1	AE 1	DL 2	S3: 3	AE 2	DD 2	AR 1	S4: 3	AR 2	S1: 3														종합형
R13	DL 1	DL 2	DP 1	DL 3	DL 4	DD 1	DP 2	S2: 4	DL 4	DL 5	S3: 4	AR 1	DD 2	S4: 4	DP 3	S1: 4													종합형

	#R01091
	#R010911
DP1	어휘
DP2	오히려
DD1	어휘력이 없어서 어휘를 외우지 않았다고?
DL1	좋아하게 됐습니다
DD2	현재네 과거에 대한 이야기를 하는 것이 아니고
DD3	그런데 그런데 어휘를 본인이 외우지 않는 것이 한국어 공부가 어려운 이유라고 말할 수는 없지 않나?
	#R010912
DD4	내용이 이 정도면 좀 부족한 거 같은데
DP3	그리고 어휘 어휘 어휘 이렇게 좀 너무 주저하고 반복하고 다시 이야기하는부분이 있었으니까
S2:3	발음은 3점
DL2	어휘 문법은 특별하게 뭐 이야기한 부분은 없었기 때문에
S3:3	이것도 3점
DD5	담화 구성도 어휘를 뭐 내가 외우지 않은 부분... 조금 어색한 것들이 있기 때문에
S4:3	3점
S1:3	그래서 전반적인 수행 3점
	#R07091
	#R070911
AE1	몽골 학생인가?
DP1	얼 어러운?
DL1	한국어를 좋아졌어? 좋아졌어
	음
DD1	어려운 점 어휘인데
DD2	학생 경험 내용이 별로 없네
DD3	마지막에 다짐은 음
	#R070912
DD4	응답 뭐 적절하게 응답은 했는데 응집성이 별로 없었어
S1:3	3점
DP2	발음 어러운 같은 것도 좀 안됐고
DP3	발화가 좀 명확하지 않은 편이니까
S2:3	3점
DL2	어휘 문법 단어 오류가 많진 않았는데
DL3	좀 제한적인 사용이 있었고
S3:3	3점
DD3	담화 구성은 내용이 대체적으로 완결은 돼있는데 내용이 별로 없다고 그래야 되나
	음3점 4점?
S4:3	3점

채점자들은 ‘경험 말하기’ 문항의 중위, 최상위, 최하위 수험자에 대한 채점 과정 보고에서 중위에서는 순차형과 종합형 채점을 한 비율이 같았으나, 상위에서는 순차형 채점이 많았고, 반대로 최하위에서는 종합형 채점을 한 경우가 더 많은 것으로 나타났다. 각 유형별로 같은 점수를 받은 사례로서 순차형과 종합형의 대표 사례를 선정하여 살펴보았을 때는 중위 수험자에서는 채점 유형 간에 채점 과정 보고량과 점수 결정 근거의 차이가 나타났으며, 상위 수험자에 대한 채점 과정에서는 채점 척도에 대한 가정과 주관적인 판단의 개입 여부에서 차이가 있었다. 하위 수험자에 대한 채점 과정에서는 평가 준거에 따라 점수 결정 근거와 주관적 판단의 개입 여부에서 차이가 나타났다. 끝으로 각 유형에서 표준 편차가 가장 컸던 채점 사례들(E10-R09, E03-R05, E09-E13)에서는 다른 채점자들과 달리 점수 결정 과정에서 상대적으로 평가적인 가정의 논증 요소를 고려하지 않거나, 채점자의 경험(AE)만을 고려하는 양상이 나타났다.

② ‘조언하는 말하기’ 문항의 채점 과정

다음으로 두 번째 문항인 ‘조언하는 말하기’에서 채점자들이 중위, 최상위, 최하위로 평가한 수험자의 응답을 어떻게 채점하였는가를 채점 과정 보고를 통해 살펴보았다. ‘조언하는 말하기’ 문항에서 최상위로 평가 받은 수험자는 ‘경험 말하기’ 문항과 같은 E03이었다. 그리고 중위에 해당하는 수험자는 E02였고, 최하위는 E05였다.(<표 Ⅳ-10> 참조).

<표 Ⅳ-10> ‘조언하는 말하기’ 문항에서 최상위-중위-최하위 수험자의 평가 준거별 평균 점수

수험자		평가 준거			
		전반적 수행	발음	어휘와 문법	담화
최상위	E03	4.62	4.54	4.77	4.46
중위	E02	3.15	3.46	3.31	3.15
최하위	E05	0.46	0.69	0.54	0.38

● 중위 수준

먼저 중위에 해당하는 수험자 E02에 대한 채점자 13인의 채점 과정 보고를 살펴보았다. 채점 척도를 기준으로 채점 과정 보고에 나타난 양상을 확인한 결과 순차형 채점을 한 채점자가 7명(R01~R03, R05~R08)으로 가장 많았으며, 종합형 채점이 4명(R04, R09, R11, R12), 그리고 무작위 채점을 한 경우가 2명(R10, R13)으로 나타났다(<표 Ⅳ-11> 참조).

수험자 E02의 '조언하는 말하기' 문항 응답에 대한 채점에서 각 평가 준거별로 평균값에 근접한 채점을 한 사례를 바탕으로 해당 수준의 채점 과정에 나타나는 특징을 확인하기 위하여 채점자 R04와 R08의 사례를 확인하였다(#R04022, #R08022). R04는 종합형 채점을 하였으며, R08은 순차형 채점을 하였는데, Ⅲ장에서 실시한 MFRM 분석에서 R04와 R08 모두 다소 관대한 채점을 한 것으로 나타난 가운데, R08(-2.14 logit)이 R04(-0.72 logit)에 비해 더 관대한 채점을 한 것으로 나타난 바 있다.

R04와 R08은 평가 결과로 보았을 때는 비슷한 채점을 한 것으로 보였지만(R04 평균: 2.5, R08 평균: 2/75), 채점 과정 보고에서는 점수 결정 근거의 차이가 나타났다. 수험자 E02에 대한 채점 과정 보고에서 평가 준거 중 '전반적 수행'에 관하여 R04는 과제를 적절하게 수행하였는가를 바탕으로 발음 구사를 고려하여 점수를 결정하였다면, R08은 과제 수행의 적절성만으로 점수를 결정하였다. 또한 발음에 관한 준거를 채점할 때도 근거의 차이가 나타났는데, R04의 경우에는 발음의 명료도를 기준으로 삼았다면("...발화 자체가 말끔하게, 깔끔하게 들리지 않아요..."), R08은 발음의 정확성을 중심으로 점수를 결정하고 있었다("여기서 어, 약간 쌌다를 샀다라고 했다든지 그래서 그런 부분에 있어서는 좀 노력을 해야 될 것 같네..."). 이러한 수험자 응답으로부터 나타난 지각 정보의 차이는 R04와 R08이 모두 2점을 부여한 어휘·문법에 관한 점수 결정 과정에서도 확인할 수 있는데, R04는 사용한 어휘·문법의 수준과 오류, 문형 사용의 문제를 바탕으로 점수를 결정하였지만, R08은 단순한 어휘·문법의 사용을 근거로 점수를 결정하고 있었다.

<표 Ⅳ-11> '조언하는 말하기' 문항에서 중위 수준 수험자(E02)에 대한 채점자별 채점 과정

채점자	채점 과정																																		유형
R01	DD 1	S1: 4	DP 1	S2: 3	DL 1	DL 2	S3: 4	DD 2	S4: 4																										순차형
R02	DL 1	DL 2	AT 1	AR 1	S1: 2	DP 1	DP 2	AR 2	DP 3	S2: 3	DL 2	AR 3	S3: 2	AT 2	DD 1	AR 4	S4: 2																		순차형
R03	DP 1	S1: 3	S2: 3	DL 1	S3: 3	AT 1	DD 1	DD 2	S4: 3																										순차형
R04	DD 1	DL 1	AO 1	DP 1-6	S2: 3	S3: 2	AE 1	DL 2	DL 3	DL 4	S3: 2	DD 2	S4: 2	AT 1	DD 3	AT 2	AR 1	S1: 3																	종합형
R05	DL 1	DL 2	DL 3	DD 1	AT 1	DD 2	DD 3	AR 1	DL 1	S1: 2	DP 1	AR 2	S2: 2	AE 1	AR 3	DL 2	S3: 2	DD 3	DD 4	S4: 2															순차형
R06	AT 1	DD 1	AT 2	S1: 2	DP 1	DP 2	S2: 4	AE 1	DL 1	S3: 3	DD 2	AT 3	AT 4	S4: 3																					순차형
R07	DD 1	AE 1	AR 1	S1: 4	DP 1	AR 1	S2: 5	DL 1	S3: 5	DD 2	DD 3	S4: 4																							순차형
R08	DD 1	DL 1	DL 2	DD 2	DD 3	AT 1	S1: 3	DP 1	DP 2	AR 1	S2: 3	DL 3	S3: 2	DD 3	AR 1	DD 4	S4: 3																		순차형
R09	AO 1	DP 1	S2: 4	DL 1	DL 2	DL 3	DL 4	DL 5	S3: 3	DD 1	DD 2	DD 3	AR 1	DD 4	DD 5	S4: 4	AT 1	DP 2	DL 6	S1: 3															종합형
R10	DP 1	DP 2	DL 1-4	AE 1	DD 1	DD 2	AE 2	AR 1	S3: 5	DD 3	AR 2	DD 4-7	S4: 5	DP 3	AO 1	DP 4	DP 5	S2: 4	AT 1	DP 6	DD 8	S1: 4													무작위형
R11	DL 1-3	DP 1	DD 1	DL 4-9	DP 2	DP 3	AR 1	DP 4	S2: 2	AR 2	DL 10	S3: 3	AR 3	AT 1	DD 2	AR 4	DD 3	S4: 3	AT 2	AR 5	DD 4-6	AR 6	S1: 3												종합형
R12	DP 1	DP 2	DP 3	DP 4	DP 5	DP 6	DP 7	S2: 3	DL 1	AR 1	S3: 2	DD 1	DD 2	DD 3	S4: 2	AT 1	DP 8	DD 4	S1: 2																종합형
R13	DL 1	DL 2	DD 1	DD 2	DL 3	S3: 5	DL 4	DL 5	DD 1	DD 2	S4: 4	DL 6	S2: 5	DP 1	AR 1	AT 1	S1: 5																		무작위형

#R04022

#R040221

DD1 자기 경험을 얘기했네

DL1 뭐야 여기도 그렇게 그렇게 하는 것이 좋겠습니다

AO1 비슷하네 1번에서 나왔던 발화 형태하고

#R040222

DP1 발음부터 보면 발음에 딱히 내가 못 알아 들을 정도의 그런 큰 문제나 오류가 나타나지는 않았어요

DP2 근데 역시나 1번에서 마찬가지로 발화 자체가 말끔하게, 깔끔하게 들리지 않아요 명확하지가 않죠

DP3 머뭇머뭇 거리는 것도 많고,

DP6 뭐 예를 들면 하고 싶은 게 뭔지 이렇게 말을 해야 된다면 하고 싶은진 뭔지 뭐 이런 식으로 정확하게 말하지 않고 얼버무리면서 넘어가는 경우들도 좀 보였고

S2:3 그래서 아까 마찬가지로 3점 정도 주면 될 거 같아요

S3:2 그리고 어휘 문법은 이 표현에서, 이 부분에서 2점..

AE1 이번 문제에서도 역시 한 2, 3급 정도 수준의 어휘와 표현 밖에 들리지 않았어요 제가 들었을 때는

DL2 오류가 뭐 크게 이해가 안 될 정도의 큰 오류가 나타나는 건 아니었는데‘머릿 속에를 정리하고’라든지 이렇게 약간 조사 오류도 중간 중간 나타났고

DL3 또 아까 앞에서 말했던 것처럼 그 활용, 활용을 명확하게 표현하지 않는경우가 많이나타났고

S3:2 역시 2점

담화 구성

DD2 1번과 마찬가지로 여기도 역시 전체적으로 자기가 말할 내용을 정리가 다 된 상태에서 말하는 그런 말하기가 아니었어요 내 생각나는 대로 말하는 것 같은? 그 정도 수준이었고

S4:2 다..그래서 여기도 2점 아까랑 마찬가지로

그 다음에 전반적인 수행 능력을보면

AT1 이게 문제가 하고 싶은 직업을 결정하지 못해 고민하는 친구에게 어떤 조언을 해줄지 그리고 그 이유를 말하는 건데

DD3 일단 뭐 친구한테 하고 싶은 것이 뭔지 물어보고 싶..물어본다고 했고, 그 다음에 조언으로서 경험을 쌓아보라고 말해 주고 싶다라고 한다고 했고, 그 뒤에 왜냐하면이라고 이유를 말하려고 하기는 했어요

AT2 그래서 그 과제에서 요구하고 있는 사항들을 다루고는 있었는데

AR1 유창성이 좀 부족했고 이해 가능 했지만... 이해가 가능한 편임.. 발화 전달과 응집

	성에 문제가 있긴 했는데 미흡에서 말한 것처럼 전반적으로까지는 볼 수가 없을 것 같고요
S1:3	그래서 아까보다 조금 높은 3점을 줄 수 있을 것 같아요 여기서는
	#R08022
	#R080221
DD1	음. 하고 싶은 게 뭔지
DL1	머릿 속에를 준비하고
DL2	음…하고에 대한 반복이 좀 많구나
	#R080222
DD2	일단 음..전반적으로 결국 내가 하고 싶은 일을 찾아서 경험을 좀 쌓고 그렇게 해야 된다 이런 얘기를 했는데 음..
	뭐라고 해야 될까 역시나 이 친구도 이해는 가능하지. 그것이 어떤 문, 응집성이나 이런 거에 문제를 가지고 있지는 않아.
DD3	그렇다고 또 완벽하게 잘했다고 볼 수는 없고
AT1	그래도 과제와 관련된 내용을 좀 말하려고 변했으니까
S1:3	3점 주고요.
DP1	발음, 일단 이 친구 발음은 확실히 좀 문제가 있네
DP2	여기서어. 약간 쌌다를 샀다라고 했다든지 그래서 그런 부분에 있어서는 좀 노력을 해야 될 것 같네 본인의 노력이 좀 필요할 것 같고
AR1	음…그렇지만 심각한 문제 까진 아니지
S2:3	그래서 이거도 한 보통? 줄 수 있을 것 같고.
DL3	어휘나 문법에서 계속 같은 표현을 좀 반복한다든지, 단순하고 좀 어휘가좀 단순하고 사용하는 문형도 역시나 그렇고.
S3:2	그래서 이거는 2점 주고요.
DD3	어 내용이 있어서 관련이 없는 내용은 확실히 말하지는 않아
AR1	그 내용이 뭘 말해야 되는지 아는데 일관성이나 이런 게 좀 부족한 편이고 논리도 그렇고.
DD4	문장을 연결하는 짜임새도 이 친구는 좀 부족하다고 볼 수 있지.
S4:3	그래서 3점 줘야겠네.

● 최상위 수준

다음으로 ‘조언하는 말하기’에서 최상위 수준으로 평가 받은 수험자 E03에 대한 채점자들의 채점 과정을 확인하였다. 채점 척도 사용을 기준으로 보았을

때 순차형 채점을 한 채점자가 8명(R01~R03, R05~R08, R10)으로 가장 많았고, 종합형 채점을 한 채점자가 3명(R04, R09, R12) 무작위 채점을 한 채점자가 2명(R11, R13)으로 나타났다(<표 Ⅳ-12> 참조). 최상위 수준인 E03의 '조언하는 말하기' 채점 과정에서 채점자들은 수험자 응답에 나타난 특정 어휘나 표현, 발음의 오류 같은 정보보다, 채점 척도나 채점자 경험, 평가 상의 고려 등을 중심으로 접근하는 양상을 확인할 수 있었다.

수험자 E03의 '조언하는 말하기' 문항의 채점 과정 보고 사례에서 채점 과정 유형에 따른 차이를 알아보기 위하여 준거별로 같은 점수를 부여한 R07과 R10의 채점 과정 보고를 비교하여 보았다(#R07032, #R10032). R07은 채점 과정 보고량이 가장 적은 채점자였으며, R10은 전체 채점자 중에 두 번째로 보고량이 많은 채점자였다. 순차형 채점을 한 R07의 채점 과정 보고에서는 전체 발화량이 적은 만큼 점수 결정의 근거가 구체적으로 드러나지 않고 있었다. '경험 말하기'의 채점 사례에서도 나타났듯이, R07은 응답에 대한 인상을 중심으로 채점을 하고 있음을 알 수 있다("응답이 잘 되었고", "발음도 좋았고", "대체로 좋다라고 해서"). R10의 경우에도 구체적으로 수험자의 응답 내용을 언급하는 부분은 나타나지 않았지만, 각 평가 준거별로 발화한 내용에 대한 인상 및 채점 척도, 채점자 경험 등을 종합하여 점수를 결정하고 있었다.

<표 Ⅳ-12> '조언하는 말하기' 문항에서 최상위 수준 수험자(E03)에 대한 채점자별 채점 과정

채점자	채점 과정																							유형
R01	AT 1	AO 1	DD 1	DL 1	DL 2	DL 3	S4: 4	S3: 4	AO 2	S2: 4	S1: 4													순차형*
R02	DD 1-4	AR 1	S1: 4	DP 1	S2: 4	AE 1	DL 1	S3: 5	DD 5	AT 1	S4: 5													순차형
R03	DD 1	DD 2	S1: 4	DD 3	DP 1	S2: 4	DL 1	S3: 4	DD 4	S4: 4														순차형
R04	DL 1	DL 2	DL 3	DL 4	DL 5	DL 6	DD 1	DP 1	S2: 5	S3: 5	AE 1	DL7 -10	DD 2	S4: 5	AT 1	AO 1	S1: 5							종합형
R05	DD 1	AT 1	DD 2	AR 1	S1: 4	DP 1	AR 2	S2: 4	AR 3	S3: 4	DD 3	AR 4	S4: 4											순차형
R06	DP 1	AT 1	S1: 5	AO 1	DP 2	S2: 4	DL 1	S3: 5	DD 1	AR 1	S4: 5													순차형
R07	AE 1	DD 1	S1: 5	DP 1	S2: 5	AE 2	DL 1	S3: 5	DD 2	S4: 4														순차형
R08	DL 1	DP 1	DL 2	DL 3	DD 1	AT 1	AT 2	DD 2	DL 4	AR 1	S1: 5	DP 2-4	AR 2	S2: 5	DL 5-8	AR 3	S3: 5	DD 3	DD 4	S4: 5				순차형
R09	DP 1-3	S2: 4	DL 1-4	S3: 5	AT 1	DD 1	DD 2	S4: 4	S1: 4	DD 3	AT 2													종합형
R10	DD 1	DP 1	DD 2	DL 1	AE 1	DL 2	DD 3	AR 1	DP 1	DD 4	AO 1	S4: 4	DL 2	DL 3	S3: 5	DP 2-3	AE 1	S2: 5	DP 3	AR 2	DD 5-6	S1: 5		순차형*
R11	DL 1	DP 1	DL 2	DP 3	DL 3	DL 4	AT 1	DD 1	DP 4	AR 1	S3: 5	DP 5	AR 2	S2: 5	DD 2	AR 3	S4: 5	AR 4	S1: 5					무작위형
R12	DL 1	DD 1	DP 1	DP 2	S2: 5	DL 2	DL 3	DL 3	S3: 5	DL 4	DD 2	AR 1	S4: 4	AR 2	S1: 5									종합형
R13	DL 1	AT 1	DL 2	DD 1	S4: 5	DD 2	DD 3	DP 2	DP 3	DD 4	S1: 5	AE 1	DP 4	DP 5	AR 1	S2: 5	DL 3	AE 2	S3: 5					무작위형

	#R07032
	#R070321
AE1	중국 학생인가, 발음이?
	아 이어폰으로 들으면 확실히 좀 너무 섬세하게 듣게 된단 말이야
	응
	응…?
	#R070322
DD1	전반적으로 어 응답이 잘 되었고
S1:5	
DP1	발음도 좋았고
S2:5	
AE2	딱히 고급 단어라든지 어휘 문법 없었지만,
DL1	음 오류가 적고 내용에 적절하게 맞았으니까
S3:5	5점
DD2	구성 자체가 이유 조언해 줬는데 이유가 좀 적절치 못해 적절하다? 음 충분하다라고 하기 좀 어려우니까 대체로 좋다라고 해서
S4:4	4점

	#R10032
	#R100321
DD1	첫 번째 문장이 약간 주제에 어긋난 거 같은데
DP1	무슨 말인지 모르겠어
	#R100322
DD2	문장 수는 몇 개 안되는데 아주 핵심을 잘 정리해서
DL1	오류가 거의 없이 말을 하는구나
AE1	이 정도면 뭐 대학원 가서 충분하게 토론하고 발표할 수 있는 정도의 수준인 거 같은데
DL2	첫 번째 문장이 약간 약간 뭔가 발화를 수정하면서 어 이게 무슨 말이지 불명확했는데
DD3	두 번째 문장부터 그걸 명료하게 얘기를 해서 전체적으로는 이해할 수 있었어
AR1	주제에 관한 적절한 정보를 충분하게 포함하고 있고
DP1	첫 번째 문장에서 내가 음 한 번 했기 때문에 일관성에서 약간 마이너스
DD4	그렇지만 대체로 논리적으로 잘 전개해서 구조가 괜찮았어
	5점을 줘야 될 거 같아
	만약에 이 학생이 약간 미흡한 부분이 그래도 있었으니까 그 부족한 점 때문에 4점을 주자 해 버리면
AO1	앞에서 응답한 그 학생보다는 훨씬 잘 했는데 담화 구성은 좀 비슷했구나
S4:4	그러면 담화 구성은 처음에 생각했던 것처럼 일단 4점을 주고

DL2 어휘 문법에서 오류가 거의 발생하지 않았기 때문에
DL3 그리고 복잡한 문형을 아주 잘 구사하기 때문에
S3:5 5점을 주는 것으로
담화 구성은 4점
어휘 문법은 5점
DP2 발음도 뭐 머뭇거리거나 주저 하는 일은 거의 없었어 자연스럽게 이어지고
DP3 약간의 뭐 수정하는 부분이 있었지만 그 정도는 사소하다고 봐야지
AE1 한국 사람들이 나도 말할 때 수정을 하는 편인데 그 정도 수준?
S2:5 그래서 발음도 5점?
응 5점으로 줘야 될 거 같은데
전반적 수행 면에서는 탁월하다고 봐야 될 거 같아
DP3 실수가 아예 없었던 건 아니지만 수정하면서 문장을 수정하면서 약간 어색한 부분들이 한 번 두 번 정도 나왔으니까
AR2 그런데 지금 평가 기준표에 보면 응답의 사소한 실수가 있지만 과제를 충족하며 적절한 세부 설명 포함했다
DD5 응 맞지 담화의 응집성이 좋았고 괜찮았고
DD6 내가 잘 이해할 수 있었던 것들이라서
S1:5 과제 수행은 5점 줄 수 있겠다

● 최하위 수준

‘조언하는 말하기’에서 최하위 수준에 해당하는 수험자 E05에 대한 채점자들의 채점 과정 보고에서는 채점 척도 사용에 따른 유형에서 순차적 채점을 한 채점자가 9명(R01~R08, R10)으로 다수를 차지하였으며, 종합형 채점을 한 채점자가 3명(R11~R13), 무작위 채점을 한 경우가 1명(R09)으로 나타났다. 최하위 수험자에 대한 채점 과정에서는 수험자 응답의 전반적인 문제들을 지적한 후에 평가 준거별로 같은 점수를 연속해서 부여한 사례들(R03~R06, R10)이 나타났으며, 이 중에서 R03을 제외하고는 모두 0점을 부여하고 있었다.

최하위 수준 응답에 대한 채점 과정 보고의 특징을 파악하기 위하여 수험자 E05의 ‘조언하는 말하기’ 문항 응답에 대한 채점 과정 보고 사례 중에 순차적 채점을 한 R01과 종합적 채점을 한 R11을 비교하였다(#R01052, #R11052). R01(-1.36 logit)과 R11(-1.03 logit)은 MFRM 분석에서 채점 경향이 관대한 것으로 나타났었으며, R01은 모형 적합도 분석(Zstd)에서 제한적인 채점 경향이

나타났었다. 두 채점자는 모든 평가 준거에 대해 1점을 부여하였다(<표 Ⅳ-13> 참조).

<표 Ⅳ-13>에서 제시한 채점자 R01과 R11의 수험자 E05에 대한 '조언하는 말하기' 문항의 채점 과정 보고에서 두 채점자는 평가 준거 중 '전반적 수행'에 대한 채점에서 미세한 인식의 차이를 나타내고 있었다. '전반적 수행'에 대해 1점을 부여할 수 있는 근거로 R01은 "그래도 하고 싶은 것을 생각해서라는 걸 말했으니까"라는 점에 주목하였다면, R11은 "응답이 과제와 관련이 없다고 보기는 힘들지만..."이라고 언급하였다. 이는 모두 '조언하는 말하기' 문항의 과제 수행에 관한 측면과 관련이 있는 근거들이지만, R01은 응답에서 과제의 요건을 충족하는 정도를, R11은 응답과 과제의 관련성을 고려하고 있었다. R01은 수험자가 응답한 내용 중에 과제의 요구에 부합하는 것이 있으므로 1점을 부여한 것이라면 R11은 응답이 과제의 요건 면에서는 부족하지만 일부 관련성이 있다고 판단하여 1점을 부여한 것이라는 점에서 R01은 수험자 수행에 나타난 미시적인 준거 요소의 유무를 따라 채점을 한 것이라면 R11은 거시적으로 의미를 부여하면서 채점한 것으로 볼 수 있다.

<표 Ⅳ-13> '조언하는 말하기' 문항에서 최하위 수준 수험자(E05)에 대한 채점자별 채점 과정

채점자	채점 과정																				유형
R01	DP 1	AT 1	DP 2	AT 2	DD 1	DD 2	DD 3	S1: 1	DP 3	DD 4	S2: 1	S3: 1	S4: 1								순차형
R02	DP1	S1: 1	DD 1	DP 2	S2: 1	DL 1	S3: 2	DD 2	S4: 1												순차형
R03	DL 1	DD 1	AR 1	S1: 1	S2: 1	S3: 1	S4: 1														순차형
R04	DD 1	DD 2	DL 1	DL 2	AO 1	DD 3	DL 3	DD 4	S1: 0	S2: 0	S3: 0	S4: 0									순차형
R05	DP 1	AT 1	DP 2	DD 1	AT 2	DD 2	DD 3	S1: 0	S2: 0	S3: 0	S4: 0	AT 3	DD 4	AT 4							순차형
R06	DP 1	DP 2	AT 1	AT 2	DD 1	S1: 0	S2: 0	S3: 0	S4: 0												순차형
R07	DD 1	DD 2	S1: 0	S2: 0	S3: 0	S4: 0															순차형
R08	DP 1	DP 2	AT 1	DD 1-3	AT 2	DL 1	DL 2	DD 4	DP 3	S1: 2	DP 4	S2: 1	DL 3	DL 4	DL 5	S3: 2	DD 5-7	AR 1	S4: 1		순차형
R09	S1: 0	S4: 0	S3: 0	DP 1	DD 1	DD 2	S2: 1														무작위형
R10	DP 1	DD 1	AO 1	DD 2	DD 3	S1: 0	S2: 0	S3: 0	S4: 0	DD 4	AR 1										순차형
R11	DP1 -4	DL 1	DP 5	DP 6	DL 2	DD 1	DD 2	AR 1	DP 7	S2: 1	AR 2	DL 3	S3: 1	AR 3	DD 3	S4: 1	AT 1	DD 4	S1: 1		종합형
R12	DP1	DD 1	AT 1	DP2 -7	S2: 2	DD 2	S3: 0	AT 2	DD 3	S4: 0	DD 4	S1: 0									종합형
R13	DP 1	DL 2	DP 2	DP 3	S2: 1	AR 1	S3: 0	S4: 0	S1: 0												종합형

	#R01052
	#R010521
DP1	긴장하는 건가?
AT1	이 문제를 뭐 이 문제에 대해서 생각해 본 적이 없어서 그런 건가?
DP2	상당히 머뭇거리네
AT2	응? 별로 어렵지 않은 문제 인데?
	#R010522
	아 이건 어떻게 해야 되지
DD1	하고 싶은 것을 생각해서 여기까지 말했으니까
	1점은 줘야 되는 건가 아닌가
DD2	그래도 한 문장도 끝까지 마무리 하지 않았으니까
	0점을 줘야하나
DD3	음 그래도… 하고 싶은 것을 생각해서라는 걸 말했으니까
S1:1	1점
DP3	이렇게 머뭇거리고
DD4	내용을 말 못 했으니
	전체1점
S2:1	전체1점
S3:1	전체1점
S4:1	전체1점

	#R11052
	#R110521
DP1	주저하고..
DP2	주저하고..
DP3	주저하고..
DP4	20초 가량..
DL1	사회..
DP5	계속 주저하고
DP6	주저하고 주저하고..
DL2	이거에 대한 어휘가 부족한 거 같진 않은데 전혀 생각해 보지 않은 문제구나
DD1	내용적으로 부족한.. 자기가 할 수 없는 부족한 일이구나
	포기? 거의 포기에 가깝지
	포기했구나
	#R110522
DD2	이거는 거의 대답을 보기가 어렵구나
	그러면 처음에 발음 같은 경우에도 한 2부터 볼까
AR1	발화의 반복과 수정, 주저하는 일이 자주 나타남

DP7 긴 휴지지
1번..
S2:1 발음은 1점
그 다음에 어휘는
AR2 주제와 관련해 사용한 어휘가 매우 제한적이며
DL3 얘기한 게 거의 없으니까
S3:1 1점
그 다음에 담화도
AR3 주제와 관련이 없어 일관성이 없고 논리가 매우 부족함 구조가 드러나지 않았음
DD3 얘기한 게 없으니까
S4:1 1점
그 다음에 전반적으로 봤을 때
AT1 응답이 과제와 관련이 없다고 보기는 힘들지만
DD4 얘기한 게 없는데 기억에 남는 게 없는데
S1:1 1점
그래서 이 학생의 1번 문제는 다 1점

'조언하는 말하기' 문항에 대한 수험자의 응답에 대한 채점 과정 보고에서는 모든 수준에서 순차형 채점을 한 채점자가 더 많았으며, 그 중에서도 최하위 수준에서 9명으로 가장 많았다. 평가 결과에서 평가 준거별로 부여한 점수가 같거나 비슷하면서 채점 유형에서 차이가 있는 채점자들의 채점 과정 보고를 분석한 결과, 평가 준거에 따라 점수 결정을 위해 고려하는 근거의 차이가 나타났으며, 특히 채점자의 청취를 통한 지각 정보의 차이로 인한 것임을 확인할 수 있었다(중위 수준). 최상위 수준의 수험자에 대한 채점에서는 구체적인 응답 정보를 근거로 고려하기 보다는 전체적인 인상이나 채점 척도, 채점자 경험 등을 바탕으로 점수를 결정하는 양상이 나타났다. 최하위 수준의 채점 과정 보고에서 순차적 채점 유형은 미시적 응답 내용에 대한 고려를 통한 채점 양상이 나타났으며, 종합형 채점에서는 거시적인 의미 부여를 통한 점수 부여의 양상이 나타났다.

③ '도표 보고 설명하기' 문항의 채점 과정

다음으로 ‘도표 보고 설명하기’ 문항에서 채점자들의 채점 과정 양상과 특징을 확인하기 위하여 해당 문항의 평가 결과를 바탕으로 중위, 최상위, 최하위에 해당하는 수험자에 대한 채점자의 채점 과정 보고를 확인하였다. 이 문항에서 중위로 나타난 수험자는 E12였으며, 최상위는 E04, 최하위는 E05였다(<표 Ⅳ-14> 참조).

<표 Ⅳ-14> ‘도표 보고 설명하기’ 문항에서 최상위-중위-최하위 수험자의 평가 준거별 평균 점수

수험자		평가 준거			
		전반적 수행	발음	어휘와 문법	담화
최상위	E04	4.00	4.38	4.31	4.00
중위	E12	2.77	3.38	3.08	2.85
최하위	E05	0.85	1.31	0.92	0.92

● 중위 수준

먼저 ‘도표 보고 설명하기’ 문항에서 중위 수준으로 분류한 수험자 E12에 대한 채점자들의 채점 척도 사용에 따른 채점 과정 보고 유형을 확인하였다. 이 문항에서 수험자들은 한국의 시간당 최저 임금의 변화에 대한 도표를 보고, 그 양상을 양과 비율 변화에 따라 설명해야 했는데, 도표를 바탕으로 담화를 구성하는 것이 과제였기 때문에 채점 과정에서 고려한 점수 결정 근거 중에 담화(DD)가 다수를 차지하고 있는 것으로 나타났다.

전체적인 채점 과정 보고 양상을 살펴보면 채점자들 중에 순차형으로 채점을 한 경우는 7명(R02, R03, R05~R08, R10)이었으며, 종합형으로 채점을 한 경우는 5명(R01, R04, R09, R11, R12)이었고, 무작위형은 1명(R13)이었다(<표 Ⅳ-15> 참조). 이 중에서 중위 수준의 평균값에 근접한 채점자 R07과 R12의 채점 과정 보고 사례를 비교하여 보았다(#R07123, #R12123). 앞서 살펴보았듯이 채점자 R07은 채점 과정 보고량이 가장 적었으며, 관대한 채점을 한 경향(-1.67 logit)을 나타났다. 채점자 R12는 채점 과정 보고량은 평균적인 수준에 가까웠으며, MFRM 분석에서는 R12와 채점 경향과 적합도 지수가 유사한 경향이 나타났다.

수험자 E12의 ‘도표 보고 설명하기’ 문항의 응답에 대한 채점자 R07의 채점 과정 보고에서는 수험자 응답의 마지막 부분에 나타난 문제에 대한 언급이 여러 번 나타나고 있었다(“아 또 하다 말아”, “마무리가 좋지 않네”, “마지막에 주저한 게 있어서”, “마지막 마무리 못했으니까”). 채점자 R12는 채점 과정에서 어휘·문법에 관한 구체적인 응답 내용을 언급하였으며, 채점 척도의 내용을 바탕으로 채점을 수행한 것으로 나타났다. R12는 평가 준거 측정에서 양적인 접근을 취하고 있었는데, “...문장들은 거의 없을 만큼...”, “...할 수 있는 말들이 적어졌고...”, “...굉장히 적은 발화량...”, “좀 더 다양한 정보를 다 담아내지는 못했다고 생각되어서...”와 같은 표지들은 계량적인 접근을 고려하고 있다는 점이 드러난 것으로 볼 수 있다.

<표 Ⅳ-15> '도표 보고 설명하기' 문항에서 중위 수준 수험자(E12)에 대한 채점자별 채점 과정

채점자	채점 과정																								유형
R01	DL 1	DP 1	DL 2	DD 1	DL 3	DD 2	DD 3	DP 2	S2: 3	DL 4	S3: 3	S4: 3	S1: 3												종합형
R02	DL 1	DL 2	DL 3	DD 1	AR 1	DD 2	S1: 2	AO 1	DP 1	S2: 2	AE 1	DL 4	S3: 2	DD 3	DD 4	AR 2	S4: 2								순차형
R03	DL 1	DL 2	S1: 2	DD 1	DD 2	AO 1	AT 1	DL 3	AT 2	DD 3	S2: 2	S3: 2	S4: 2	DD 4	DL 4										순차형
R04	DL 1	DD 1	DL 2	DP 1	DL 3	DD 2	DP 2-3	S2: 4	AT 1	DL 4	DL 5	AT 2	AE 1	S3: 3	DD 3-6	AR 1	S4: 3	DD 7	DP 4	AR 2	S1: 2				종합형
R05	DL 1	DP 1	DL 2-4	DP 2	DD 1	DP 3	DD 2	DL 5	DD 3	AR 1	DL 6	S1: 2	DP 3	DP 4	S2: 2	DL 7	DL 8	S3: 2	DD 4-6	AR 1	S4: 2				순차형
R06	DD 1	DP 1	DL 1	DL 2	AT 1	AT 2	DD 2	DD 3	AR 1	S1: 2	DP 2	AR 2	S2: 3	DL 3	DL 4	S3: 3	DD 4	AR 3	S4: 2						순차형
R07	DD 1-3	AT 1	AO 1	S1: 3	DP 1	DP 2	S2: 3	DL 1	S3: 3	DD 3	DD 4	S4: 3													순차형
R08	DL 1	DL 2	DP 1	AT 1	DD 1	AR 1	DP 2	S1: 4	DP 3	DP 4	S2: 5	DL 3	S3: 4	DD 2	S4: 5										순차형
R09	DL 1	DL 2	DP 1	DP 2	DP 3	DP 4	DP 5	DP 6	S2: 3	DL 3	AR 1	AT 1	S3: 2	DD 1	DD 2	DD 3	AR 2	S4: 1	DD 4	DD 5	DD 6	S1: 2			종합형
R10	DP 1	DD 1-7	AR 1	DD 8	S4: 3	DL 1	DL 2	DL 3	DD 9	DL 4	DL 5	S3: 4	DP 1	DP 2	DP 3	S2: 3	DD 10-12	S1: 3							순차형*
R11	DD 1-6	DL 1	DL 2	DP 1	DD 7	DP 2	DP 3	DP 4	S2: 4	DL 3	AR 1	S3: 4	DD 8-14	AR 2	AR 3	S4: 3	DP 4	S1: 3							종합형
R12	DD 1	DL 1	DD 2	DD 3	DP 1	DP 2	S2: 4	DD 4	DL 2	DD 5	DL 3	DD 5	S3: 3	DD 6	AR 1	DD 7	S4: 3	AR 2	DD 8	S1: 3					종합형
R13	DL 1	DD 1	DD 2	DL 2-4	DP 1	S2: 5	DD 3	DD 4	S4: 4	DL 5	DL 6	DD 5	S3: 4	AR 1	DD 5	DD 6	S1: 4								무작위형

	#R071231
DD1	대체로 좀 발화가 짧네
DD2	아 또 하다 말아
	#R071232
DD3	마무리가 좋지 않네
AT1	이게 어려운가 보다
AO1	전반적으로 1번에 너무 기대했나
S1:3	3점
DP1	발음 발음은 괜찮았어
DP2	좀 머뭇거리고 마지막에 주저한 게 있어서
S2:3	3점
DL1	어휘 특별한 게 없고 좀 제한적이었으니까
S3:3	3점
DD3	담화는 구조가 잘 안 드러나고 일관성이 부족하고
DD4	마지막 마무리 못했으니까
S4:3	3점
	#R12123
	#R121231
DD1	전체적으로 그래프는 잘 짚어내고 있다
DL1	인상률이 급증했다는 표현을 썼네?
DD2	끝난건가?
DD3	내용이 너무 없는데?
	#R121232
	먼저 발화 발음은
DP1	발화가 명확한 편이었고
DP2	어, 일단 들으면 이게 무슨 말이지?라고 그니깐 이해가 안되는 문장들은 거의 없을 만큼 모음 자음 받침 다 전체적으로 발음이 명확했기 때문에
S2:4	발음은 4점
	그 다음에 어휘 문법은
DD4	어, 도표 자료를 읽어내는 초반 앞부분에서는 어, 유창하다고 느껴질 만큼 도표 자료를 잘 설명을 했고
DL2	뭐 인상률 최저 임금 등등 사용해야 단어들이나 문형들도 적절하게 잘 사용을 했는데
DD5	뒤로 갈수록 어, 할 수 있는 말들이 적어졌고
DL3	그것은 어휘력이 부족한 건가 싶은 생각이 들었고
DD5	결국에는 굉장히 적은 발화량으로 마무리를 했기 때문에
S3:3	어휘 문법 3점
DD6	다음으로 담화 구성도 어, 주제와 관련된 정보들을 포함을 하고 있기는 하지만

AR1 어..구조가 잘 드러나지 않았고 문장 연결도 좀 부족했고
DD7 일단은 충분히 도표를 다 읽어내지는 못했기 때문에
S4:3 3점
마지막으로 전반적인 수행 능력도
AR2 어, 과제와 관련한 응답을 어느 정도 적절하게 다루고 있기는 했지만
DD8 좀 다양한 정보를 다 담아내지는 못했다고 생각되어서
S1:3 3점 주겠다

● 최상위 수준

다음으로 '도표 보고 설명하기' 문항에서 최상위 수준에 해당하는 수험자 E04의 응답에 대한 채점자들의 채점 과정 보고 양상을 살펴보았다(<표 Ⅳ-16> 참조). 채점 척도 사용에 따라 채점 과정 보고 유형을 분류해 보면 순차형 채점을 한 채점자가 7명(R01, R02, R03, R05, R06, R07, R08), 종합형 채점이 4명(R04, R09, R11, R12), 무작위형 채점이 2명(R10, R13)으로 나타났다. 이 중에서 평균값을 기준으로 가장 근접한 채점을 한 채점자 R01과 R05의 채점 과정 보고를 비교하여 살펴보았다(#R01043, #R05043).. 채점자 R01과 R05는 채점 척도 사용을 기준으로 모두 순차형 채점을 하였지만, R01은 방향이 반대인 역순차형 채점을 하였으며, 두 채점자 모두 모든 준거에 대해 4점을 부여하였다.

'도표 보고 설명하기' 문항의 최상위 수준으로 나타난 E04에 대한 R01과 R05의 채점 과정 보고를 살펴보면, R01의 경우에는 담화적인 특징에 관해서는 상세한 근거들을 제시하면서 채점을 하고 있었지만, 어휘·문법이나 발음, 전반적 수행에 대한 채점에서는 구체적인 근거를 제시하지 않으면서 점수를 결정한 양상이 나타났다. R05의 채점 과정 보고에서는 R01에 비해 E04의 응답에 나타난 구체적인 사실들을 바탕으로 점수를 결정하고 있음이 나타났는데, R05의 채점 과정 보고에서는 응답의 전체적인 양상에서 채점자가 받은 인상을 고려하는 담화 표지들("아쉽다", "좀 아쉬움이 있어서")을 반복해서 사용하는 것이 특징적이었다.

<표 Ⅳ-16> '도표 보고 설명하기' 문항에서 최상위 수준 수험자(E04)에 대한 채점자별 채점 과정

채점자	채점 과정																									유형
R01	DP 1-3	AT 1	DD 1	DD 2	AE 1	DD 3	DD 4	S4: 3	DD 5	DL 1	S3: 4	DP 4	S2: 4	S1: 4												순차형*
R02	DP 1	DL 1-3	DD 1-3	S1: 4	DP 2	S2: 5	DL 4	AE 1	DL 5	S3: 5	DD 4	DD 5	S4: 5													순차형
R03	DL 1	DL 2	DD 1	DD 2	DD 3	AT 1	DL 3	AE 1	AT 2	S4: 3	DL 4	DL 5	S3: 3	DP 1	S2: 3	AT 3	S1: 3									순차형*
R04	DL 1	DL 2	DL 3	DL 4	DL 5	DP 1	AE 1	S2: 5	AT 1	S3: 5	DD 1	DD 2	S4: 5	AT 2	DD 3	AO 1	DD 4	S1: 4								종합형
R05	AT 1	DL 1	DL 2	DL 3	DP 1	DL 2	DL 3	DD 1	DD 2	DD 3	S1: 4	DP 2	S2: 4	DL 4	DP 3	AR 1	S3: 4	DD 4	DD 5	DD 6	DD 7	S4: 4					순차형
R06	DL 1	DL 2	DD 1	DD 2	DL 3	DD 3	DD 4	DD 5	AT 1	S1: 4	AR 1	S2: 5	DL 4	DL 5	S3: 4	AR 2	S4: 4									순차형
R07	DL 1	DL 2	DL 3	DD 1	DD 2	DD 3	DD 4	S1: 4	DP 1	S2: 4	DL 4	S3: 4	AR 1	AR 2	S4: 4											순차형
R08	DL 1	DL 2	DL 3	DD 1	DP 1	DD 2	DD 3	AT 1	DD 4	AR 1	DD 5	S1: 5	DP 2	S2: 5	DL 4	DL 5	S3: 5	S4: 5	DD 5	DL 6	DD 6					순차형
R09	DD 1	DP 1	DP 2	AO 1	S2: 3	DL 1	DL 2	AT 1	S3: 4	DD 2	AT 2	DD 3	DD 4	S4: 3	DL 3	DD 5	DD 6	DD 7	AR 1	S1: 4						종합형
R10	DP 1	DL 1-4	DD 1	DD 2	DL 5-7	DP 2	AE 1	S2: 4	DD 3-5	AR 1	DD 6	S4: 3	DP 3-5	AE 2	DP 6	DL 7	AR 2	S3: 3	AR 3	DD 7	AR 4	S1: 3				무 작 위 형
R11	DL 1-2	DP 1	DD 1	DP 2	DL 3	DD 2-5	DL 4-5	DD 6	DP 3	AT 1	DP 3	S2: 5	DL 6-7	AO 1	S3: 5	AR 1	DD 7	AR 2	DD 8-10	S4: 4	AR 3	DD 11	S1: 4			종합형
R12	DP 1	DD 1-2	DL 1	DP 2-6	S2: 5	AE 1	DL 2	DL 3	AE 2	DL 4	S3: 5	DD 3	AR 1	DD 4	AT 1	S4: 4	DD 5	AR 2	S1: 4							종합형
R13	DP 1	DP 2	DP 3	DP 4	DL 1	DD 1	DL 2	DL 3	DL 4	AR 1	DL 5	AR 2	DL 6	S3: 5	DD 2	DD 3	S4: 5	DD 4	DP 5	S2: 5	S1: 5	DD 5				무 작 위 형

#R01043
#R010431
DP1 매년
DP2 이르렀습니다
DP3 매우?
#R010432
AT1 이유를 물어보지 않아도 다 얘기를 하는구나
DD1 그래프만 설명하기엔 시간이 남는다고 생각하나?
DD2 그래프 내용이 많아서 사실 하나 하나 다 설명을 하면 시간은 충분할 텐데
AE1 보통 이렇게 습관적으로 근거라든지 현상이라든지 결과 방법 해결 방법 이런 것들을 설명하려고 하는 학생들이 많기도 하고 그리고 그게 아니더라도 시간이 남는다고 생각을 해서 더 그 주제에 대해서 더 많이 말을 하려고 하는 거 같네
그럼 이런 경우에는 담화 점수를 어떻게 해야지? 그래도 담화 점수를 다 줄 순 없는데
DD3 왜냐면 그래프에 대해서 더 세밀하게 설명을 할 수도 있는 거니까 충분히
DD4 너무 짧게 그래프 설명을 했으니까 그러면
S4:3 3점
DD5 사실 제대로 설명을 다 설명 안 한 것 같고
그래도 어휘와 문법같은 경우에는
DL1 그래프 설명하는데 필요한 부분을 어느 정도 사용을 했다고 보여주기 때문에
S3:4 4점
DP4 발음도 괜찮은 편이어서
S2:4 4점
S1:4 전반적인 것은 4점
#R05043
#R050431
AT1 그래프는 제대로 이해
DL1 보여주는 것을 제대로 잘 읽고 있고 수치 틀리지 않았음
DL2 꺾은선 그래프 얘기하고 있고
DL3 그래프를 잘 읽어냄
DP1 머뭇거림이 거의 없고 자연스럽게 발화를 하고 있네
DL2 법을 정확하게 시켜야 합니다?
DL3 마지막에 약간 이 법을 시켜야 합니다? 지켜야 합니다.
DD1 아 맨 마지막 부분에 조금 아쉽다
#R050432
DD2 어, 이 학생이 전반적 수행은
우선 그래프 꺾은선 그래프와 막대 그래프에 있는 수치와 내용을 정확하게 이해하고 있고
DD3 여기에 자신의 생각까지 적절하게 넣어서 이야기를 하고 있다는 점에서 4점.

	어 5점은 좀 아쉬움이 있어서
S1:4	4점. 4점을 줄 수 있을 거 같고.
DP2	발음은 어…상당히 자연스럽고 머뭇거리거나 주저하는 일이 맨 마지막에 약간 있고 거의 없기 때문에
S2:4	여기서도 4점 .주는 것이 적절하다고 보여집니다.
DL4	어휘나 문법 부분에서도 맨 마지막에 지불하는 을지부하는, 어 이 법을지켜야 한다를 시켜야 한다 등으로 약간의 오류가 있지만
DP3	그 이전에는 거의 어...자연스럽게 발화를 했기 때문에
AR1	음.그래도 크게 의미를 이해하는데 방해가 되지 않아서 5점까지는 무리고
S3:4	4점을 줘도 무방하다고 보여집니다.
DD4	담화 구성도 맨 처음에 그 그래프의 어떤 연도와 수치를 읽어내고 그 다음에 인상률에 대한 꺾은선 그래프...에 대한 수치도 정확했고
DD5	다음에 자신의 생각을 넣어서 어 .정부가 어떤 최저 임금에 대해서 관심을 많이 가졌기 때문에 이렇. 어 그렇게 급격한? 임금 인상률을 보인 거 같다 등으로 대체적으로 약간 논리적으로 전개가 잘 되고 있다가
DD6	맨 마지막 부분에 좀 자신감을 상실한 듯 보이고 어 연결…마무리가 있기는 하지만 조금. 근데 그렇다고 해서 3점을 주기에는 너무 낮게 준 거 같고 5점을 주기에는
DD7	마지막 부분에 구조에 조금 아쉬움이 있어서
S4:4	여기도 4점을 주는 것이 적절하다고 보여집니다.

● 최하위 수준

다음으로 ‘도표 보고 설명하기’ 문항에서 최하위 수준이었던 수험자 E05에 대한 채점자들의 채점 과정 보고를 살펴보았다. E05의 채점 과정 보고에서 9명의 채점자가 순차형(R01~R08, R10)으로 채점을 하여 다수를 차지하고 있었고, 나머지 4명의 채점자는 종합형(R09, R11~13) 채점을 한 것으로 나타났다(<표 Ⅳ-17> 참조).

이 중에서 평균값을 기준으로 순차적 채점을 한 R06과 종합적 채점을 한 R11의 채점 과정 보고를 통해 해당 문항에서 최하위 수준의 채점 과정 보고에 나타난 특징을 살펴보았다(#R06053, #R11053). 채점자 R06(-1.08 logit)과 R11(-1.03 logit)은 MFRM 분석에서 비슷한 채점 경향을 갖고 있는 것으로 나타났으며, 그 중에서 R06은 모형 적합도의 표준점수(Zstd)값이 -3.5로 나타나

채점 척도를 제한적으로 사용하고 있음이 나타난 바 있다. 수험자 E05의 '도표 보고 설명하기' 문항 응답에 대한 채점 과정 보고에서 채점자 R06과 R11은 매우 유사한 근거를 바탕으로 평가 준거에 대한 점수 결정을 하고 있음이 나타났으며, 점수 결정을 위해 고려하는 가정의 측면에서 차이가 있었는데, R06은 평가 과제의 측면을 고려한 반면에 R11은 채점 척도를 중심으로 판단을 내리고 점수를 부여하고 있음을 알 수 있다.

<표 Ⅳ-17> '도표 보고 설명하기' 문항에서 최하위 수준 수험자(E05)에 대한 채점자별 채점 과정

채점자	채점 과정																				유형
R01	DP 1	D L 1	D D 1	A T 1	A O 1	A O 2	S1: 1	S2: 1	S3: 1	S4: 1	D D 2	D D 3									순차형
R02	D L 1	A R 1	S1: 1	S2: 1	S3: 1	S4: 1															순차형
R03	DP 1	D D 1	A T 1	S1: 1	S2: 1	S3: 1	S4: 1														순차형
R04	D D 3	S1: 0	S2: 0	S3: 0	S4: 0	D D 4	D D 5	DP 2													순차형
R05	DP 1	D D 1-3	S1: 1	A R 1	S2: 1	D L 1	S3: 1	A R 2	A T 1	S4: 1											순차형
R06	D L 1	D L 2	D L 3	D D 1	DP 1	A T 1	D D 2	D D 3	S1: 1	DP2	S2: 1	S3: 1	D D 4	S4: 1							순차형
R07	D D 1	D D 2	A T 1	S1: 1	S2: 1	S3: 1	S4: 1														순차형
R08	D L 1	DP 1	D D 1	A T 1	D D 2	A R 1	D D 3	S1: 2	DP 2	DP 3	S2: 1	D L 2	D L 3	S3: 1	D D 4-6	S4: 1					순차형
R09	DP 1	DP 2	A T 1	S2: 1	D L 1	D L 2	D L 3	S3: 1	D D 1	S4: 0	D D 2	D D 3	S1: 0								종합형
R10	DP 1	D D 1	D D 2-4	A R 1	S4: 1	A T 1	D L 1	S3: 1	A R 2	DP 2	S2: 1	A O 1	S1: 0	D D 5							순차형*
R11	D L 1	DP 1	D L 2	D D 1	D D 2	DP 2	S2: 1	A R 1	D L 3	A R 2	S3: 1	D D 3	D D 4	A R 3	D D 5	S4: 1	A R 4	D D 6	S1: 1		종합형
R12	D D 1	DP 1	D D 2	D D 3	DP 2	D D 4	S2: 2	D D 5	D D 6	S3: 1	DD7 -10	S4: 1	D D 11	S1: 0							종합형
R13	D D 1	D D 2	D L 1	DP1	S2: 5	D L 2	D L 3	D L 4	S3: 1	D D 3	S4: 2	A R 1	S1: 2								종합형

#R06053

#R060531

DL1 올라가서

DL2 증가를 했으면 좋을 거 같고

DL3 상승했으면 좋을 거 같고

DD1 최저임금이 어떻게 변하는지를 이야기하고 있고

DP1 중간에 민망한지 자꾸 웃고

AT1 그 다음에는 이 학생은 그래프 읽는 과제 수행을 제대로 하지 못한 거 같습니다

#R060532

DD2 어 일단 시간당 최저 임금 2017년에 얼마에서 2018년에 얼마 변화를 했는지를 보고 있는데 그리고 나서 최저 임금 인상률이 어떻게 되는지를 이야기하고 있지 않습니다

DD3 그래서 응답을 하기는 했지만 그 내용이 매우 부족하다라고 생각을 해서

S1:1 전반적인 수행을 1점을 주려고 하고요

DP2 그 다음에 발화의 긴 휴지가 있고 발화를 거의 하지 않았기 때문에

S2:1 발음도 1점

S3:1 어휘 문법도 1점

담화 구성도 1점을 주도록 하겠습니다

0점을 줘도 될 거 같긴 하지만

DD 그래도 그래프를 하나 읽으려고 했다는 부분에서

S4:1 1점씩 주도록 하겠습니다

#R11053

#R110531

목소리부터 자신이 없네

DL1 숫자 읽는 데 어려움이 있고

DP1 휴지가 엄청기네

DL2 숫자 읽는 게 어렵고

DD1 거의 이 것도 발화를 포기했구나

#R110532

DD2 얘기한 게 2007년하고 2018년밖에 없고

DP1 그러면 보면 발음도 거의 긴 휴지가 주저하는 일이 있었으니깐

S2:1 발음도 1점

AR1 주제와 관련하여 사용한 어휘가 단순하고 사용한 어휘가 제한적임, 심각한 오류, 의미 이해의 어려움

단순한 어휘가 매우 제한적이며 심각한 오류, 의미 이해..

DL3	의미를 이해할 수는 있었지만 말한 게 별로 없었지
AR2	주제와 관련해서 사용한 어휘가 단순하고.. 단순하고.. 문형이 제한적임 매우 제한적임.. 제한적임.. 매우 제한적임.. 매우 제한적임..
S3:1	
DD3	말하는 게 별로 없으니까
DD4	대부분 주제에 관련이 없는..은 아니고
AR3	일관성 부족하며 주제 전개 논리가 빈약함 구조가 드러나지 않으며 문장 연결에 짜임새가 거의 없음
DD5	이것도 얘기한 게 거의 없으니까는
S4:1	1점
	그리고 전반적으로
AR4	응답이 과제와 거의 관련이 없음..
DD6	관련이 없다고 보기는 어렵지만 얘기한 게 2007년에 3480원, 2018년에7530원 이 얘기밖에 한 게 없으니까
S1:1	1점

‘도표 보고 설명하기’ 문항에 대한 채점 과정 보고 분석의 결과, 채점 척도 사용에 따른 채점자들의 채점 유형에서는 중위, 최상위, 최하위 수준에서 모두 순차형 채점이 더 많이 나타났다. 그 중에서 평가 결과에서 중위 수준으로 나타난 경우의 채점 과정 보고에서는 점수 결정의 근거로서 주로 담화적인 특징(DD)을 고려하는 양상이 나타났다. 그리고 채점 유형에 따라서 순차적 채점을 한 경우에는 전체적인 인상에 근거한 판단을 내리고 있는 반면에 종합적 채점을 한 경우에는 수집한 정보에 대해 계량적인 접근을 하는 것을 알 수 있었다(중위 수준). 최상위 수준의 채점 과정 보고 중에 역순차형으로 채점을 한 사례에서는 가장 먼저 채점한 담화에 대해서는 구체적인 근거 고려가 나타났으나, 나머지 평가 준거에 대해서는 인상적인 접근을 취하고 있었다. 최하위 수준의 수험자에 대한 채점 과정 보고의 사례에서는 채점 유형에 관계없이 유사한 응답의 특성을 바탕으로 점수를 부여하고 있었으며, 순차형은 과제의 특성을 가정으로 고려하였다면, 종합형에서는 채점 척도의 내용을 가정으로 삼아 점수를 결정하고 있었다.

④ ‘기사 읽고 문제와 해결 방안 이야기하기’ 문항의 채점 과정

‘한국어 말하기 능력 시험’의 마지막 평가 문항은 ‘기사 읽고 문제와 해결 방안 이야기하기’였다. 이 문항은 수험자들에게 신문 기사를 제시하고, 기사 속에서 문제를 파악한 후에, 그에 대한 해결 방안과 함께 응답하는 것을 과제로 제시하고 있었다. 앞선 ‘도표 보고 설명하기’ 문항과는 자료를 바탕으로 담화를 구성하여 응답해야 한다는 유사점이 있지만, 자료의 유형과 요구하는 답안의 구성 방식에 차이가 있다. 이 문항에 대한 채점자들의 채점 과정 보고 분석을 위하여 수험자 중에 중위, 최상위, 최하위 수준을 선정하고, 이들에 대한 채점 과정 보고에 나타난 특징을 바탕으로 채점 과정의 양상을 파악하고자 하였다. 평가 결과, 해당 문항에서 중위 수준이었던 수험자는 E01이었으며, 최상위는 E03, 최하위는 E09였다(<표 Ⅳ-18> 참조).

<표 Ⅳ-18> ‘기사 읽고 문제와 해결 방안 이야기하기’ 문항에서 최상위-중위-최하위 수험자의 평가 준거별 평균 점수

수험자		평가 준거			
		전반적 수행	발음	어휘와 문법	담화
최상위	E03	4.62	4.23	4.69	4.62
중위	E01	2.46	2.85	2.62	2.62
최하위	E09	1.77	1.38	1.69	1.85

● 중위 수준

이 문항 유형에서 수험자 E01은 평가 결과에서 중위 수준으로 평가되었는데, 이와 관련하여 전체 채점자들의 채점 과정 보고를 분석하여 보았다(<표 Ⅳ-19> 참조). E01에 대한 채점자들의 채점 과정 보고에서 채점 척도 사용에 따른 채점 과정 유형으로는 순차형 채점이 8명(R01~R03, R05~R08, R13)으로 가장 많았으며, 종합형 채점이 3명(R04, R09, R12), 무작위형이 2명(R10, R11)으로 나타났다. 이 중에서 해당 문항 유형의 채점 과정의 특징을 파악하기 위하여 채점자 R07와 R08의 채점 과정 보고를 비교하여 보았다(#R07014, #R08014). 이 문항 유형에 대한 E01의 응답에 대해 순차형 채점을 한 R07과

종합형 채점을 한 R08의 평균은 동일하게 2.5점이었으나, 평가 준거 중에 전반적 수행과 발음의 점수에서 1점씩 차이가 있었다.

‘기사 읽고 문제와 해결 방안 이야기하기’ 문항의 수험자 E01에 대한 R07과 R08의 채점 과정 보고에서 R07은 다른 문항의 채점 사례에서와 마찬가지로 전반적으로 청취한 정보에 대한 구체적인 회상 보다는 인상 중심의 전반적인 접근을 위하고 있음이 나타났다. R07의 채점 과정 보고에서 나타난 “대충”, “좀 계속”, “전체적으로”, “너무” 등의 담화 표지는 직관으로부터 얻은 인상을 중심으로 채점을 하고 있음을 나타내고 있었다. R08의 채점 과정 보고에서는 응답 청취를 통해 R07에 비해 보다 다양하고 구체적인 특징에 대한 파악이 이루어졌음이 나타났으며, 점수 결정 과정에서는 과제의 요건과 채점 척도를 가정으로 활용하는 특징이 나타났다. 한편, R08은 발음을 채점하는 과정에서 과제 특성을 고려하는 모습(AT2)이 나타나기도 하였는데, 이러한 현상은 읽기 자료를 사용하여 응답을 하는 문항 유형의 특징이 채점 과정에 영향을 끼치고 있음을 나타낸 것으로 보인다.

<표 Ⅳ-19> '기사 읽고 문제와 해결 방안 이야기하기' 문항에서 중위 수준 수험자(E01)에 대한 채점자별 채점 과정

채점자	채점 과정	유형
R01	DI 1 DF 1 DL 2-4 DL 1 DL 2 DL 5 DL 3-7 S4 3 DL 6 DL 7 S3 2 DF 2 S2 2 S1 2	순차형*
R02	DF 1 DL 1-2 DF 2 DL 3-4 DL 1-2 AT 1 DL 3 AF 1 DL 4 S1 2 AC 1 DF 3 S2 3 DL 5 DL 6 AF 1 S3 3 DL 4 AF 2 DL 5 S4 2	순차형
R03	AT 1 DL 1 DL 1 DL 2 DL 2 AT 1 DF 1 DL 2 AT 3 S1 3 S2 4 S3 3 S4 3	순차형
R04	DL 1-2 DF 1-2 DL 3 DF 3-5 AF 1 S2 3 DL 4 AF 1 DL 5-6 AF 2 DL 7 AF 2 DL 8-9 AF 3 DL 10 AF 3 DL 11 S3 2 AT 1 DL 1-5 AF 4 S4 2 AT 2 AF 5 S1 2	종합형
R05	DL 1 DF 1 DL 2 DF 2 DL 3-7 DL 1 DL 8-9 DL 2-4 S1 3 DF 3-4 AF 1 DF 5 S2 3 DL 10-12 S3 3 DL 5 AF 2 DL 6-7 S4 2	순차형
R06	DL 1 DL 2 DL 3 DL 4 DL 5 AT 1 AT 2 DL 1 S1 3 DF 1 DF 2 DF 3 S2 3 AF 1 DL 2 DL 6 DL 7 S3 3 DL 3 DL 4 S4 3	순차형
R07	DF 1 DL 1 DL 2 DL 1 AT 1 S1 3 DF 2 DF 3 S2 2 DL 3 S3 2 DL 2 DL 3 S4 3	순차형
R08	DF 1 DL 1 DF 2-3 DL 2-3 DF 4 DL 1 DF 5-6 DL 4 DF 7 DL 5 DL 2-3 AT 1 DF 8 DL 4-5 S1 2 AF 1 AT 2 S2 3 DL 6-7 S3 2 DL 6-7 AF 2 DL 8 S4 3 DD 9-10	순차형
R09	DF 1 DL 1 DL 2 DL 1 DL 2-6 S2 2 DL 3-6 S3 2 S4 1 DL 2-3 DL 7 DL 8 DL 4 S4 1 DF 7 DL 5 S1 1	종합형
R10	DL 1-5 DL 1-2 DL 6-7 DL 3-4 DF 1 DL 5-6 S2 2 DL 7 AF 1 DD 9-10 S4 2 DL 8-9 AF 2 S3 2 AT 1 DF 2 DD 11-12 S1 2 DL 10 DL 13	무작위형
R11	DL 1-2 DL 1-2 DL 3-6 DL 3 DL 7 AF 1 DL 8-9 DL 1-3 DL 10 DL 4 DL 1 DL 1-14 DF 2 DL 5-17 DL 5 DL 8-21 DL 6 AF 1 DL 21-22 S3 3 DF 3-4 AF 2-3 DF 5-6 AF 4 S2 3 AF 5-6 DL 7 S4 3 AF 7 DD 8-10 S1 3	무작위형
R12	DF 1-5 S2 4 DL 1 DL 2 S3 3 DL 1 S4 4 DL 3 DL 2 DL 3 S1 3	종합형
R13	DF 1 DF 2 DL 1 DL 2 DL 1 AT 1 DL 2 DL 3 DL 3 AF 1 DL 4 S1 3 DF 3 S2 3 DL 5 DL 6 AF 2 S3 3 DL 4 DL 5 AF 3 S4 3	순차형

	#R070141
DP1	유료? 유료화?
DL1	문제를 생길 수도
	무슨 말이냐 집중이 안되네
DL2	과학자에 의하면
	#R070142
	아휴 피곤해
DD1	해결하기 위한 방법 자체에 너무 뭉뚱그린 답변이 있었고
AT1	전체적으로 문제는 대충 얘기를 했고
S1:3	음…전반적으로는 3점
DP2	주저 발화 반복 이런 것들은 좀 계속 일어나니까
	3점이고
DP3	음 발음도 좀 듣기 어려웠어
S2:2	음 2점
DL3	어휘문법은 제한적… 이것도 어휘문법도 전체적으로 내용 이해하기가 힘들었으니까
S3:2	2점
DD2	적절한 내용을 일부 포함하고 있고,
DD3	아 난 논리 자체가 너무 부족하기 때문에 이것도
S4:3	3점

	#R08014
	#R080141
	음?
DP1	건강? 권강?
DL1	과학자들 라면 과학자들에 의하면.
DP2	지구온난화, 지구온난화
DP3	도로, 더러, 이륙 발음이 정확하지 않은데?
DL2	물이 부족한, 부족하면
DL3	기운이 올라가다, 올랐다. 같은 오류구나.
	으음.
DP4	대기오염..
DD1	해결 방법은 오염을 줄이자.
	으음, 그런데
DP5	음. 기구, 지구.
DP6	변화 편화.
DL4	으음. 사용 방법을 바꾸자?
DP5	터런화?
	으음…

DL5 벌써 지구 온난화를 바꿨습니다?
무슨 얘길까?
#R080142

DD2 일단 이 친구가 말한 요지는..기사에 대해서는 제일 마지막에 얘기한 고온현상에 의한 얘기를 잠깐 했고 .그래서 해결 방법으로 얘기한 게..오염되는현상을 줄이자. 이런 얘기를 한 거 같은데.

DD3 아, 일단 내용이.. 뭔가 귀에 쏙 들어오는 그런 내용이 아니네?

AT1 전반적으로..어, 그래도 과제에 대해서 적절한 내용을 좀 하려고는 했는데

DD4 음.. 그래도 이해는 어느 정도 되는데..

DD5 내용 전체가 전반적으로 이해가 좀 다 되었다고 하기는 어려운데..

S1:2 그래도 전반적인 수행은 2점을 줘야겠다.

AR1 왜냐하면 과제를 다루고는 있어도, 표현을 이해할 수 있어도 그래도 발화 전달이나 담화 응집성이나 이런 거는 조금 문제가 있었던 거 같아,
3점을 주고 싶긴 한데.

AT2 그렇다고 해서 과제 관련해서 막..본인이 여러 가지 이야기를 하거나 가져오지는 않은 거 같고.

S2:3 그러니까 발음은 3점.

DL6 어휘나 문법은 처음부터 끝까지 친구는 뭔가 복잡한 문형을 사용하거나 계속 그러지는 않았고.

DL7 문장이 그렇다고 이 사용한 어휘를.. 주제와 관련된 어떤 복잡한 것? 이것도 아니기 때문에 약간 단순한 운영 단순한 어휘 이런 것들

S3:2 그래서 이것도 2점 어휘문법도 2점 주고

DD6 담화구성…굉장히 이 친구가 내용을 뭔가 설명할 때 그거를 뒷받침할 수 있는 이런 얘기들 보다는 본인의 생각이 약간 즉흥적으로 오염 방법을 오염되는 걸 줄이자 방법을 바꾸자 했지만

DD7 마지막 같은 경우는 사실 거의 끝맺음이 잘 되지 않았고

AR2 결국 보면 이 친구는 문장 연결의 짜임새가 좀 부족한 편이지. 흐름이 일관성도 부족하다고 할 수 있고.

DD8 그렇지만 주제와 관련이 없는 정보를 많이 얘기하거나 그러진 않았어.

S4:3 그래서 담화 구성에서는 그냥 3점을 주는 걸로.
그래도 자신감 있게 하는 거나 이런 거는 참 좋은데.

DD9 음. 조금 마무리를 잘하면 그래서 좀 어떤 짜임새 있는 그런 문장으로 완병, 완결을 좀 하다 보면 좋을 거 같은데

DD10 정작 무슨 얘기를 하는지 조금 알 수 없었던 게 좀 많이 아쉽네.

● 최상위 수준

다음으로 이 문항 유형에서 최상위 수준으로 평가 받은 수험자는 E03이었다. 이미 앞선 첫 번째, 두 번째 문항 유형에서도 최상위 수준으로 평가 받은 E03은 MFRM 분석 결과에서 전체 수험자 중에 가장 우수한 성적을 거둔 것으로 나타난 바 있다. E03에 대한 전체 채점자의 채점 과정 보고의 양상은 <표 Ⅳ-20>과 같이 나타났는데, 채점 척도를 기준으로 유형을 나누었을 때 8명의 채점자가 순차형 채점(R01~R03, R05~R07, R10, R13)을 하였으며, 4명의 채점자가 종합형 채점(R04, R09, R11, R12)을 하였고, 나머지 1명의 채점자(R08)는 무작위형 채점(을 한 것으로 나타났다. 이 중에서 E03의 응답에 대한 평가 결과의 평균값에 근접한 R02와 R12의 채점 과정 보고 사례에 대한 검토를 통하여 이 문항 유형에서 최상위 수험자에 대한 채점 과정 보고의 특징을 알아보았다(#R02034, #R12034).

E03에 대한 채점자 R02와 R12의 채점 과정 보고에서는 순차형 채점을 한 R02의 채점 과정에서는 종합형 채점을 한 R12와 달리 전반적 수행에 대한 점수를 과제 수행 요건을 중심으로 결정하였음이 나타난다(“기사를 살짝 읽은 느낌이 있으므로”). R12의 경우에는 평가 준거에 대한 종합적인 관점에서 전반적 수행에 대한 점수를 부여하였기 때문에 지역적인 문제(“그런데 사소한 실수들이 조금 있기는 해도”)에 대한 고려를 반영하여 최종 점수를 결정하였다. 그밖에 발음과 어휘·문법, 담화에 대한 채점에서 R02와 R12는 점수 결정을 위해 고려하는 근거의 양이 비슷했지만, 자신의 경험(AE)을 1회 고려한 R02와 달리 R12는 과제(2회)와 채점 척도(1회)를 가정으로 고려하고 있었다.

<표 Ⅳ-20> '기사 읽고 문제와 해결 방안 이야기하기' 문항에서 최상위 수준 수험자(E03)에 대한 채점자별 채점 과정

채점자	채점 과정																														유형
R1	DL 1	DL 2	DD 1	AO 1	DD 2	S4: 4	DL 3	S3: 4	DP 1	DP 2	S2: 3	AT 1	S1: 4																		순차형*
R02	DP 1	DL 1	DL 2	DP 2	DL 4	DD 1	DL 5	DL 6	DP 3	DD 2-5	S1: 4	DP 4	DP 5	S2: 4	AE 1	DL 7	S3: 5	DD 6	S4: 5												순차형
R03	DL 1-4	DD 1	AT 1-3	DD 2-4	S1: 5	DD 5	DL 5	S2: 5	S3: 5	S4: 5																					순차형
R04	DL 1	DL 2	DP 1	DD 1	DD 2	DL 3-5	DP 2	DL 6	DL 7	S2: 5	DL 8	S3: 5	DD 3-6	S4: 5	AT 1	S1: 5															종합형
R05	DP 1-3	DD 1	DP 4	DP 5	DD 2-6	S1: 4	DP 6-10	S2: 4	DL 1	DL 2	AE 1	S3: 4	DD 7	S4: 4																	순차형
R06	DL 1	DD 1	DP 1	DD 2	DP 2	DL 2	DD 3-5	AR 1	S1: 5	DP 3	AR 2	S2: 5	DL 3	DL 4	S3: 4	DD 6	DD 7	S4: 5													순차형
R07	DP 1	DP 2	DL 1	DD 1	DL 2	DD 2	DL 3	DL 4	DD 3	S1: 3	DP 3	DP 4	S2: 4	DL 5	S3: 5	DD 4	S4: 5														순차형
R08	DD 1	DP 1	DL 1	DP 2	DD 2	DL 2	DL 3	DP 3	DP 4	DD 3-8	AT 1	DP 5-9	S2: 5	AT 2	DD 9	S1: 5	S4: 5	AR 1	AR 2	S3: 5											무작위형
R09	DP 1-3	S2: 4	DL 1-5	DD 1	AR 1	DD 2	S4: 4	S3: 5	DL 5	S1: 4	AO 1																				종합형
R10	DL 1	DP 1-2	DL 2-3	DD 1	DL 4-5	DD 2	AO 1	DD 3-4	AR 1	S4: 5	DL 6	AE 1-2	DL 7	AT 1	S3: 5	DP 3-4	S2: 4	DD 5	DP 5	S1: 5	AE 2	DD 6									순차형*
R11	DL 1-3	DP 1	DL 4	DL 5	DP 2	DL 6	DP 3-4	DL 7-9	DD 1-3	DL 10-13	DP 5	DL 14	DP 6	DD 4-6	AR 1	S2: 3	DL 15-16	AR 2	S3: 4	DD 7	AR 3	DD 8-9	S4: 4	AR 4	S1: 4						종합형
R12	AT 1	DP 1	DD 1	DD 2	DP 2-5	S2: 4	DL 1-4	S3: 5	AR 1	DD 3-4	S4: 4	DL 5	DD 5	AT 1	DL 6	DD 6	S1: 5														종합형
R13	DP 1	DD 1	DD 2	DL 1	DD 3	DL 2	DL 3	DD 4	DP 2	S2: 5	DD 5-7	DL 4	S3: 5	AR 1	S4: 5	DD 8	S1: 5														종합형

	#R02034
	#R020341
DP1	홍수.. 호수 아닌가?
DL1	인식 가지기..
DL2	구체적 방법
DP2	범인.. 범인..
DL4	..도록 하나..
DD1	개인뿐만 아니라 기업과 나라
DL5	대책, 정책..
DL6	어? 어휘력이 상당히 좋은데
DP3	뭘 전환시켜?
DD2	작은 젓가락이라도.. 아하
	#R020342
DD3	기사를 읽지 않으려고 노력은 했지만, 뭐 어쩔 수 없을 거 같고
DD4	먼저 기사를 분석하고 자기가 생각하는 이유를 말했는데
DD5	기사를 살짝 읽은 느낌이 있으므로
S1:4	전반적 수행은 4점
DP4	발음, 좋았는데 하나 이해 안 되는 부분이 있었고
DP5	홍수, 범인.. 약간의 발음 오류
S2:4	4점
AE1	어휘, 굉장히 다양한 고급 문법과 어휘를 사용하려고 노력했음
DL7	오류도 거의 없었음
S3:5	5점
DD6	담화구성, 논리적으로 개인, 정부, 기업이 해야할 일을 잘 설명했음
	#R12034
	#R120341
AT1	일단 문제는 잘 이해한 거 같다
DP1	특히를 계속 특이…라고 읽는 것 같음, 같다.
DD1	환경에 대한 의식을 지적하고 있구나
DD2	여러 가지 방안을 생각했네?
	어, 뭐라 그랬지? 못 알아 들었는데
DP2	일본 학생 특유의 받침 오류가 있네?
	#R120342
DP3	음, 먼저 발음은 전반적으로 어.. 명확한 수준이기는 한데
DP4	정책이라고 발음해야 하는 걸 니은으로 전책으로 발음하는 오류가 지속적으로 보이고
DP5	또 중간중간에 무슨 발음인지 조금 알아듣기가 어려웠던 부분들이 있어서

	어.. 근데 뭐 이런 부분들은 조금만 노력하면 될 것 같은 부분이어서 우수하다고 판단이 된다
S2:4	발음 점수는 4점 주겠다
DL1	그 다음에 어휘 수준은 여러 가지 단어들을 잘 사용을 하고 있는데
DL2	폐기가스라든가 일회용품이라는 단어들
DL3	그 다음에 뭐 환경에 대한 의식 등등 좀 다양한 어휘를 사용을 하고 있고
DL4	또 그것을 사용을 할 때 문법적인 부분에 있어서도 어, 그렇게 많은 오류가 느껴지지는 않았기 때문에 어휘 점수도 높게 줄 수 있을 것 같다
	음… 전반적으로 어휘 수준이 높았다고 느껴지기 때문에
S3:5	어, 어휘 점수는 5점을 주겠다
AR1	그 다음에 담화 구성에 있어서는 흐름이 굉장히 일관되고
	어, 해결 이런 문제를 어떻게 하면 해결할 수 있을지
DD3	개인과 기업과 나라의 순서로 점점 확장되는 의미의 단어를 순서대로 잘 사용을 하고 있고
DD4	그것을 그냥 열거하는 수준이 아니라 논리적으로 짜임새 있게 전개하고 있기 때문에
	어.. 담화 구성도 아주 우수한 수준이라고 생각된다
S4:4	그래서 담화 점수는 4점 주겠다
	마지막으로 전반적인 수행은 뭐, 뭐..
DL5	표현의 실수가 조금 있기는 했지만
DD5	담화의 응집성이 있고
AT1	어쨌든 과제를 상당히 잘 다루고 있다고 느껴지기 때문에
	어.. 4점 아니면 5점.
DL6	음, 그런데 사소한 실수들이 조금 있기는 해도
DD6	전체적인 논리성이 그것을 덮을 수 있는 수준이었던 거 같다
S1:5	따라서 탁월하다고 판단되는 수준의 점수 5점 주도록 하겠다

● 최하위 수준

‘기사 읽고 문제와 해결 방안 이야기하기’ 문항에서 최하위 수준으로 나타난 수험자 E09에 대한 채점 과정 보고에서 채점자들은 점수 결정의 근거로서 담화적 특징보다는 주로 응답에 나타난 발음과 어휘·문법의 문제를 고려하고 있었다. 채점 척도 사용에 따른 채점 과정 유형에서는 순차형 채점을 한 채점자가 7명(R02, R03, R05~R08, R10)으로 가장 많았고, 다음으로 종합형 채점

(R11~R13)과 무작위형 채점(R01, R05, R09)이 각각 3명으로 나타났다. 이 문항 유형에서는 이전의 문항 유형들의 최하위 수준 채점 과정 보고에서 순차적으로 0점이나 1점을 부여하는 채점 경향이 나타나지 않았다(<표 Ⅳ-21> 참조).

이 문항 유형에서 순차적 채점을 한 R07과 역순차적 채점을 한 R10은 평가 결과가 전체 평균에 가장 근접한 채점자였다. 이들의 채점 과정 보고에 나타난 특징을 통해 해당 문항 유형의 채점 과정 양상을 알아보았다(#R07094, #R10094).

전체 채점 과정 보고에서 채점자 R07은 보고량이 가장 적은 채점자였으며, R10은 두 번째로 보고량이 많은 채점자였는데, E09에 대한 채점 과정 보고 사례에서 R07은 총 분절된 발화의 수가 12개, R10은 39개로 3배 이상 차이가 있었다. 채점 과정 보고의 내용을 분석한 결과, R07과 R10은 모든 평가 준거에 같은 점수를 부여하였으나, 점수를 부여하는 근거의 종류에 차이가 있었다. 전반적 수행에 대한 채점에서 R07은 과제 수행의 적절성에 대한 판단을 근거로 점수를 부여하였다면, R10은 담화의 응집성 수준에 대한 고려를 포함하여 점수를 결정하였다. 발음에 대한 채점에서 R07은 유창성과 관련하여 “너무 주저하는 일이 자주 나타나...”의 특징을 근거로 점수를 결정한 반면에 R10은 “발음이 잘못됐어 그래서 무슨 말 하는지 기사를 안 보면 모르겠네”라며 정확성에 대한 문제를 주목하였다는 차이가 있었다. 어휘·문법에 대한 채점에서 R07은 “가끔 중급 이상의 단어가 들리긴 한데...”라며 사용한 일부 어휘의 수준을 고려하여 점수를 결정하였으며, R10은 구체적인 어휘·문법 사용상의 문제를 지적하다가 응답에 나타난 특정 어휘(중급 이상 수준)와 채점 척도를 고려하여 점수를 결정하였다.

<표 Ⅳ-21> '기사 읽고 문제와 해결 방안 이야기하기' 문항에서 최하위 수준 수험자(E09)에 대한 채점자별 채점 과정

채점자	채점 과정																														유형
R01	DP 1	DL 1	DP 2-4	DL 2	DD 1	DP 5	DL 3	DP 6	DP 7	DL 4	DL 5	DD 2	DL 6	S3: 2	DP 8	S2: 2	DD 3	S4: 3	S1: 2												무작위형
R02	DL 1	DP 1	DP 2	DP 3	DL 2	AR 1	DD 1	S1: 2	S2: 1	DL 3	S3: 1	DD 2	DD 3	S4: 2																	순차형
R03	DP 1	AO 1	AO 2	AT 1	AT 2	AT 3	DP 1	DL 1	AT 4	AO 3	AO 4	S1: 1	S2: 1	S3: 1	S4: 1																순차형
R04	DL 1	DP 1	DL 2	DP 2	DL 3	DL 4	DL 5	DP 3	DD 1	DL 6	DD 2	S2: 1	DP 4	AE 1	DL 7	DL 8	S3: 1	DD 3	DD 4	S4: 1	S1: 1	AR 1									무작위형
R05	DL 1	AE 1	DP 1-3	DL 2	DP 4-6	DL 3	DP 7-8	DL 4-6	S1: 2	AR 1	DD 1	DP 9	DL 7-9	DP 10-11	AR 2	S2: 1	DL 10-11	AR 3	DL 12	AR 4	DD 2	S3: 2	DD 3-4	AR 5	S4: 2						순차형
R06	DL 1	DP 1	DL 2	DD 1	DL 3-4	DP 2	DD 2	DL 3	DP 3	DL 4-5	AT 1	DD 3-4	DP 4-5	S1: 1	S2: 1	DP 6	DL 6-7	S3: 2	DD 5-6	S4: 1											순차형
R07	DP 1	DL 1	DL 2	AT 1	AR 1	S1: 2	DP 2	S2: 1	AE 1	S3: 2	AR 2	S4: 2																			순차형
R08	DL 1	DP 1	DP 2	DD 1	AR 1	DD 2	DP 3	DP 4	S1: 3	S2: 2	DL 2	S3: 2	S4: 2																		순차형
R09	DL 1	DP 1	DL 2	DP 2	DL 3	DP 3	AR 1	S2: 1	DL 4	AR 2	DL 4	DL 5	S3: 1	DL 6	DD 1	DD 2	S4: 1	AR 3	S1: 1												무작위형
R10	DP 1-3	DL 1	DL 2	DP 4-5	DL 3	AE 1	DD 1	AT 1	DD 2-3	AR 1	DD 4-6	S4: 2	DL 4-8	AR 2	S3: 2	DP 5	AE 1	DP 6-8	AE 2	S2: 1	AR 3	DL 10	DD 7	S1: 2							순차형*
R11	DL 1-3	DP 1	DP 2	DL 4-6	DP 3-5	DL 7	AR 1	S2: 1	DL 8	AR 2	S3: 1	DD 1	AR 3	DD 2	S4: 1	AR 4	S1: 1														종합형
R12	DP 1	DL 1	DP 2-6	DD 1	DP 7	DP 8	S2: 2	AE 1	DL 2	DL 3	S3: 2	AR 1	DD 2	DL 4	S4: 2	DD 3	DL 5	DD 4	DD 5	S1: 2											종합형
R13	DL 1	DP 1	DP 2	DD 1	AT 1	DL 2	DL 3	DP 3-7	DL 4	DP 8	S2: 2	DL 5	DD 2	DL 6	S3: 3	AT 1	DD 3	DD 4	DP 9	DP 10	S4: 4	AR 1	S1: 3								종합형

	#R070941
DP1	대기
DL1	많이 쓰네?
DL2	재활용? 재용할?
AT1	아 문제를 이해하는 것 같기는 한데 전혀 문제점 뭐 해결방법 잘 안되네
	#R070942
AR1	전반적으로 음 제한적인 주제 다루고 있고
S1:2	2점
DP2	발음도 너무 주저하는 일이 자주 나타나 이해하기 좀 힘들고
S2:1	1점
AE1	어휘 문법은 가끔 중급 이상 단어 들리긴 한데, 사용
S3:2	2점이라도 주고
AR2	담화 구성은 짜임새가 거의 없고 주제는 관련 있으니까
S4:2	2점
	#R10094
	#R100941
DP1	요즘을
DP2	와 또 발음이 너무 안 좋네 환경오염을 오검이라고 하는데
DP3	슬글 많이 사용 한다고?
DL1	대기오염에 문제가 있습니다 이것도 좀 어색하지
DL2	태풍이 나와? 발생하지
DP4	발음이 잘못됐어 그래서 무슨 말하는지 기사를 안보면 모르겠네
DP5	재용활 참 재활용
DL3	재용활을 많이 사용하면 좋겠습니다 수정이 불가능한 하고 싶지 않은 문장 비문을 말하고 끝내 버렸어
	#R100942
AE1	이 학생은 4급이 아니야
DD1	일단 담화 구성 면에서는 좋은 점수를 줄 수 없는 것이
AT1	문제를 발생한 문제를 먼저 얘기하고 그 문제를 해결하기 위해서 이런 방법들을 해야 한다 이렇게 넘어갔어야 되는데
DD2	그 언급도 없이 갑자기 뭐 쓰레기 자동차 얘기하고 대기오염 갑자기 그것도 마무리 짓지 않고 다시 태풍 홍수
DD3	이것도 문제로 다시 돌아가서 얘기를 하지 않나
AR1	흐름의 일관성 매우 부족하고 그러다 보니 논리가 빈약할 수밖에 없지
DD4	무슨 말 하는지 잘 모르겠고 구조도 잘 드러나지 않는 미흡한 수준
DD5	그런데 주제와 관련이 없는 내용들은 아니기 때문에 흠 부족하다

DD6 아주 없다라고 보기에는 이 학생이 발화량이 많다 보니까 아주 적은 건 아니어서
S4:2 담화구성 2점을 주고
DL4 어휘나 문법을 잘못 사용했어
여기서 좋은 점수를 줄 수 없는 것은
DL5 어휘를 좀 잘못 외워서 막 이상한 말로 바꿔서 하고
DL6 수정하고 주저하고 이러면서 더한 오류를 만들어서
DL7 마지막 문장은 비문으로 끝날 정도로 심각한 오류들이 많이 나타났지
DL8 지구 온난화란 얘기도 했었고
DL9 단순 문형만을 사용했다라고 볼 순 없으니까
S3:2 어휘 문법도 2점
발음은
DP5 이 학생의 문제는 일단 발음에 있기 때문에
AO1 어휘를 잘못 발화하는 것까지 그게 연결이 되는 거 같아
DP6 그래서 사실 동반 감점을 유발시키는 요인은 어쨌든 발음에 있어
DP6 주저하는 일이 뭐 간혹 있었지만
DP8 긴 휴지는 아니었지만
AO2 이게 발음이 심각한 문제가 전체적으로 다른 요인도 더 안 좋은 요인들을 더 생기게 하기 때문에
S2:1 발음 면에서는 1점을 주고
AR2 응답이 과제와 거의 관련이 없다고는 볼 수 없지 제한적으로 어느 정도 일부 다루고 있고
DL10 아 표현을 이해할 수는 있지만
DD7 응집성에 뭐 큰문제들이 있기 때문에
S1:2 전반적 수행은 2점

‘기사 읽고 문제와 해결 방안 이야기하기’ 문항에서 수험자의 평가 결과에 나타난 수준에 따라 채점자들의 채점 과정에 나타난 특징을 살펴보았을 때, 모든 수준에서 순차형 채점을 한 사례가 종합형이나 무작위형보다 많은 것으로 나타났다. 중위 수준의 수험자에 대한 채점 과정 사례에 대한 분석에서는 직관과 인상 중심으로 평가 준거를 순차적으로 채점을 한 사례와 구체적인 응답 특징에 대한 고려 속에서 종합적으로 이루어진 채점을 통해 점수를 부여한 사례의 결과는 비슷했지만, 자료를 활용하는 문항의 특성이 잘 나타난 것은 후자의 사례였음을 알 수 있었다. 최상위 수준의 수험자에 대한 채점 과정 분석에서는 순차형 채점을 한 채점자가 인상 기반의 접근을 한 것을 알 수 있었고,

종합형 채점을 한 채점자는 최종 점수 결정에서 지역적인 문제를 고려하고 있음이 나타났다. 최하위 수준의 수험자에 대한 채점 과정 보고에서는 채점 유형에 따라 평가 준거별로 점수 결정에 고려하는 근거의 양상에 차이가 나타났는데, 순차형 채점을 한 사례에서는 인상 기반의 접근을 통해 점수를 결정하였으며, 종합형 채점을 한 사례에서는 구체적인 문제를 바탕으로 점수를 결정하고 있었다.

이상에서 말하기 평가의 문항 유형에 따른 채점 과정 구성의 특징을 알아보기 위하여 각 문항별로 중위, 최상위, 최하위 수험자를 선정하고, 해당 수험자에 대한 전체 채점자의 채점 과정 보고에 나타난 특징과 평균적인 점수를 받은 채점자들의 채점 과정의 양상을 알아보았다. 이러한 접근을 통하여 확인한 '한국어 말하기 능력 시험'에서 구성한 4개의 구성형 응답 문항별로 응답 수준에 따른 채점자의 채점 경향을 정리하면 다음과 같다.

첫째, 구성형 응답 문항의 채점 과정 분석에서 채점자들은 채점 척도에 제시된 순서를 따라 점수를 부여하는 순차형 채점을 가장 많이 하고 있는 것으로 나타났다. 이러한 경향은 본 연구에서 활용한 4개 문항 유형 중에서 '경험 말하기'의 중·하위 수준을 제외하고 나머지 모든 문항 유형과 수험자 수준의 채점 과정 보고에서 나타났다. '경험 말하기'는 문항 곤란도가 가장 낮은 문항으로서 다른 문항에 비해 중·하위 수준의 수험자들이 받은 점수가 더 높았으며, 이는 다른 문항에서 같은 수준의 응답에 비해 채점 과정에서 처리해야 할 여러 가지 정보가 나타났고, 따라서 채점 과정에서 체계적인 접근을 더 요구하였을 것으로 볼 수 있다.

전체 13명의 채점자들의 채점 과정을 살펴보면, 수험자의 수준이나 문항의 유형에 상관없이 항상 순차형 채점을 하는 채점자(R06), 항상 종합형 채점만을 하는 경우(R11, R12)도 나타났으며, 중위 수준에서는 종합형 채점을 하고, 상위나 하위 수준에서는 종종 순차형 채점을 하는 사례(R04) 등을 확인할 수 있었다. 순차형 채점을 한 채점자의 채점 과정에서는 수험자 응답에 대한 채점자의 인상 기반 접근을 하고 있음이 나타났으며, 종합형 채점 유형에서는 구체적인 응답 정보에 대한 판단을 포함하는 경향이 나타났다.

둘째, 채점자들마다 점수 결정을 위해 고려하는 근거와 가정의 차이가 발생할 수 있으며, 이는 채점자가 청취한 정보와 수집한 정보에 대한 판단, 그리고

점수 결정을 위한 접근법의 차이로 인하여 발생할 수 있다. 채점자들은 채점 과정에서 수험자 응답 및 개인적인 경험이나 지식 등으로부터 채점을 위한 근거나 가정을 형성할 수 있으며, 수험자의 응답과 관련하여서는 청취한 정보에 대한 선별적 지각의 차이가 이후의 회상적 정보 처리 과정에서 이루어지는 점수 결정에 영향을 끼칠 수 있다. 이러한 채점 자원들에 대한 채점자 인식의 차이는 채점 과정을 통해 부여하는 점수의 차이로 나타날 수도 있지만, 점수 결정을 위하여 고려하는 수준과 접근 방법에 따라 영향이 미미하거나 상쇄될 경우에는 확인이 어렵다. 이는 평가 결과만을 얻기 원하는 경우에는 문제가 되지 않겠지만, 평가를 통해 학습자와 교사가 교수·학습에 관하여 유용한 정보를 얻고자 하는 성취도 평가의 경우에는 평가 결과가 포함하는 구체적인 메시지를 알아 볼 수 있는 방법이 필요하다.

셋째, 구성형 평가 문항 유형 중에 자료를 사용하는 문항 유형에서는 채점 과정에서 담화적인 측면에 대한 고려와 과제 수행에 대한 고려가 점수 결정 전반에 영향을 끼칠 수 있음을 고려해야 한다. 본 연구에서 사용한 구성형 응답 문항 중에 '자료 보고 설명하기'와 '기사 읽고 문제와 해결 방안 이야기하기'의 채점 과정 사례에서는 점수 결정을 위한 근거로 담화적인 측면(DD)이 가장 많이 고려되고 있는 것으로 나타났으며, 또한 점수 결정을 위한 가정 중 과제(AT)를 고려하는 양상도 나타났다. 이러한 사실은 자료를 활용해야 하는 말하기 평가 문항에서는 준비한 자료가 수험자의 응답뿐만이 아니라 채점 과정에서 채점자의 인식에 영향을 끼쳐 전체적인 점수 결정에 관여할 수 있음을 보여주는 것이다.

넷째, 최종 점수 결정 과정에서 순차형의 채점을 한 경우에는 인상 기반의 접근을 하며, 종합형의 채점을 한 경우에는 사실 기반의 접근을 한다. 문항별 채점 과정 분석에서 채점자들은 최종적인 점수 결정 과정에서 일부 점수를 조정하려는 모습을 나타냈는데, 그 과정에서 순차형 채점을 한 경우에는 전체적인 인상 기반의 접근을 한 반면에 종합형 채점을 한 경우에는 특정 요소나 지역적인 사실에 근거하여 점수를 조정하고 있었다. 이러한 차이는 채점자가 점수 결정을 위한 자원이 부족하다고 느끼는 상황에서 어떤 방식으로 채점을 하고 있었는지에 따라 고려하는 근거의 차이로 인하여 나타난 것으로 볼 수 있다.

(2) 채점자 영향에 따른 채점 과정의 차이

말하기 평가의 채점 과정을 알아보기 위하여 '한국어 말하기 능력 시험'의 평가 결과 가운데 MFRM 분석을 통해 채점자 영향이 나타난 사례를 중심으로 채점 과정 분석을 실시하고, 채점 과정의 측면에서 채점자 영향의 양상과 그에 대한 원인을 탐색하였다.

① 근거 형성의 차이로 인한 엄격한 채점 경향

주관적인 채점자의 특성은 실제 수험자의 응답 수준보다 낮은 점수를 부여하는 엄격한 채점과 반대로 응답 수준보다 높은 점수를 부여하는 관대한 채점으로 나누어 볼 수 있다. 채점 과정 분석에서 채점자의 특정한 채점 경향(rating tendency)은 채점 과정에서 고려하는 준거의 영향과 점수 결정의 근거가 부족한 가운데 나타났다.

앞서 실시한 MFRM 분석에서 채점자 중 R08과 R05는 1번 문항에서 분석 모형의 예상을 벗어난 상대적으로 더 엄격한 채점을 한 것으로 나타났다. 반대로 R11과 R13은 같은 문항에서 더 관대한 채점을 한 것으로 나타났다. R05의 엄격한 채점 특성을 확인하기 위하여 전체 채점자의 평균과 편차가 1.27점으로 나타난 E12에 대한 채점 사례를 확인하였다(채점자 전체 평균=4.52, R05 평균=3.25). 해당 사례에서 R05의 채점 과정 특징을 파악하기 위하여 같은 사례에 대하여 평균 수준에 부합하는 채점을 한 경우(R03)와 비슷한 경향을 보인 경우(R09)의 채점 과정을 비교하여 보았다(<표 Ⅳ-22> 참조).

<표 Ⅳ-22> 엄격한 채점 특성이 나타난 채점 과정1(R05-E12, 1번 문항)

채점자		채점 과정																
엄격한	R05	DP1	DL1	DD1	DL2	DL3	S1:3	DP2	DP3	S2:4	AT1	DL4	DL5	DL6	S3:3	DD2	DD3	S4:3
평균	R03	DL1	DL2	DL3	DL4	DL5	AE1	S1:5	DD1	S4:4	DL6	DP1	S3:4	S2:5				
비교	R09	DP1	DP2	S2:4	DL1	S3:2	DD1	DD2	DD3	S4:3	AR1	DD4	DP2	DD5	S1:3			

엄격한 채점 특성이 나타난 채점 과정을 분석한 결과, 엄격한 채점 경향은 말하기 평가 과제의 '전반적 수행 능력'에 대한 채점이 먼저 엄격하게 이루어지면서 이후에 다른 준거들에 대한 채점의 기준으로 작용한 것으로 나타났다. 1번 문항에서 가장 엄격한 채점을 한 R05는 채점 과정에서 먼저 '전반적 수행'에 대해 3점을 부여하였으며, 이후에도 그와 유사한 점수대(3~4점)로 채점을 한 것으로 나타났다. 그에 반해 평균적인 채점을 한 R03은 채점 과정에서 먼저 '전반적 수행'에 대해 5점을 부여한 이후에, 다른 준거들에 대해서도 4~5점을 부여한 것으로 나타났다. 결국 먼저 결정한 전반적인 과제 수행에 관한 점수(S1)의 차이가 이후의 다른 평가 준거의 점수 결정에도 영향을 주면서 엄격한 채점이 이루어졌다고 볼 수 있다. R05와 마찬가지로 해당 수험자의 1번 문항 채점에서 엄격한 채점을 한 것으로 나타난 R09는 채점 과정에서 '전반적 수행' 점수를 다른 평가 준거의 점수를 종합하여 마지막에 결정한 것으로 나타났다.

이 문항은 수험자의 한국어 학습 경험을 말하는 서사적 말하기 기능을 평가하는 문항으로서, 상대적으로 가장 곤란도가 낮은 문항이었다. 엄격한 채점을 한 R05와 R09의 채점 과정이 평균적인 채점을 한 R03과 다른 점은 '담화 구성 능력'에 관한 근거(DD)를 여러 번 고려하고 있다는 점이다. '담화 구성 능력'은 과제의 요구를 따라 적절한 응답을 논리적으로 구성하였는지와 관련이 있다. 엄격한 채점이 이루어진 사례에서 응답에 나타난 담화적 특징에 대한 채점자의 반복적인 주목이 있었다는 것은 담화적인 측면에 대한 고려가 총체적인 채점의 경향을 가중할 수 있음을 나타낸 것이라고 볼 수 있다. 엄격한 채점 경향과 관련하여 채점자들은 점수 결정에 관한 가정을 고려하는 부분에서도

차이가 있었는데, R05는 어휘·문법에 대한 채점에서 과제의 화제를 고려(AT)를 하였다면, R03은 전반적 수행의 수준 판정과 관련하여 자신의 경험(AE)을 가정으로 삼은 것으로 나타났다. 종합적으로 채점자 특성이 0 logit, 즉 엄격한 특성과 관대한 특성이 나타나지 않았던 R09의 경우에는 해당 채점에서는 R05와 같이 엄격한 특성이 나타났으며 채점 척도(AR)를 가정하여 점수를 결정하였다는 점이 공통적이었다.

엄격한 채점이 이루어진 것과 관련하여 총체적 평가 준거인 '전반적 수행 능력'에 관한 채점 과정 보고의 내용을 살펴보면 평가 준거에 대한 채점자들의 관점에 차이가 있었다(#R051212, #R031212). R05는 전반적 수행에 관한 점수를 결정하는 과정에서 응답에 나타난 어휘의 수준을 고려한 것으로 보인다. 또한 담화적인 측면에서 응답에 나타난 미숙한 점을 다시 고려하면서 엄격한 채점이 이루어진 것이다. 그에 반해 R03은 응답에 나타난 전반적인 인상을 바탕으로 종합적인 준거의 채점을 하고 있었다. 채점자가 종합적인 접근을 취해야 하는 해당 평가 준거에 대해 분석적, 즉 계량적으로 채점을 할 때 엄격한 채점 경향이 나타날 수 있을 것이다.

#R051212

이 학생은 전반적으로 들었을 때는, 과제와, 과제와 관련해서 적절하게 과제를 어 수행하고 있어 보입니다
그런데 전반적인 표현 자체가 그렇게 고급 표현을 사용하는 게 아니라 중급 이하의 표현을 사용해서 유창하게 이야기를 하고 있는 것으로 봐서… 4점까지 주기에는 조금 어려움이 있을 것 같고
어, 그 다음에 끝나는 부분도 갑자기 완결되는 느낌이 조금 있어서
음 전반적 수행은 3점 정도를 주는 것이 적절해 보입니다.

#R031212

자 전반적으로 구어에 굉장히 강한 느낌이죠?
몇몇 표현들을 보면 구어에 워낙 강하기 때문에
노출이 많이 됐다라는 것이 딱 드러나면서
말하기가 좋습니다
전반적 5

이러한 특징은 채점자의 엄격한 채점 특성이 나타난 다른 사례에서도 찾아볼 수 있다. 채점자 R08의 수험자 E07에 대한 1번 문항 채점에서 엄격한 특성이 나타났는데, 채점 과정에서 총체적 평가 준거(S1)를 먼저 고려하였다는 점과 담화적인 측면에 관한 근거(DD)를 상대적으로 많이 고려하는 경향이 나타났다. <표 Ⅳ-23>에 제시한 채점 과정을 보면, E07의 1번 문항 채점에서 R08과 유사한 채점 특성이 나타난 R04의 경우에는, 총체적 평가 준거를 마지막에 부여하였지만, 담화적 측면의 근거를 많이 고려하고 있다는 점이 나타나고 있다(<표 Ⅳ-23> 참조).

<표 Ⅳ-23> 엄격한 채점 특성이 나타난 채점 과정2(R08-E07, 1번 문항)

채점자		채점 과정																								
엄격한	R08	DD1	DD2	DL1	DL2	DD3	DD4	DD5	AR1	S1:3	DP1	AR2	S2:2	DL3	DL4	S3:2	DD6	DD7	AR3	S4:2						
평균	R12	DP1	DP2	DP3	DP4	DP5	S2:4	DL1	DL2	S3:4	DD1	AR1	S4:4	DD2	AT1	DP6	DP7	DP8	S1:4							
비교	R04	DL1	DL2	DL3	DL4	DL5	DP1	DD1	DP2	AR1	S2:3	DL6	AR2	S3:3	AR3	DD2	DD3	DD4	DD5	S4:2	AR4	DD6	DD7	S1:2		

채점자 R08과 R04는 채점 과정에서 공통적으로 수험자(E07)의 1번 문항 응답에서 마무리가 잘 이루어지지 않은 부분을 주목하고 있음이 나타난다(#R080712, #R040712). 이러한 결과는 담화적인 측면을 넘어서 채점 영역 전반에서 체계적으로 영향을 나타내고 있었다.

#R080712

아 마무리가 없네, 마무리가 딱 문장이 끝나야 되는데. 쩝. 그리고 다음에 뭐 했을까?
그래도 일단 본인이 한국어 배울 때 어려운 점이 일본, 아니 그 발음과 관련된 거?
근데 그 뭐…학습경험이나 이런 것도 굉장히 구체적으로 얘기를 하지 않았으니까.
그래도 이해나 이런 것들은 괜찮았고
그래서 전반적 수행은 3점

#R040712

그리고 전반적인 수행으로 돌아가면

2점 볼까요 표에서

과제를 다루고 있지만 제한적으로 주제를 다루었고, 표현은 이해할 수는 있지만 전반적으로 발화 전달과 응집성에 문제가 있다.

왜냐하면 내용 조직이 제대로 안 된 상태에서 발화가 갑자기 끝나버려서 전반적인 수행에서 좋은 점수를 주기는 좀 어려울 것 같은데

그래서 2점 주겠습니다.

다음으로 채점 과정에서 나타난 엄격한 채점 경향이 문항 유형에 따라 차이가 있는지를 알아보기 위하여 3번 문항과 4번 문항에서 엄격한 채점이 이루어진 사례들을 살펴보았다. 3번 문항에서 엄격한 채점이 이루어진 사례는 채점자 R13이 수험자 E10을 채점한 결과에서 나타났다. 해당 결과를 살펴보기 위하여 E10에 대한 R13과 해당 피험자에 대해 평균적인 수준으로 채점을 한 R05의 채점 과정을 비교하여 보았다(<표 Ⅳ-24> 참조).

<표 Ⅳ-24> 엄격한 채점 특성이 나타난 채점 과정3(R13-E10, 3번 문항)

채점자		채점 과정																														
엄격한	R13	DL1	DL2	DL3	DL4	DL5	DL6	DL7	DD1	S1:3	DP1	S2:4	DL8	S3:3	DD2	DD3	S4:4															
평균	R05	DP1	DL1	DL2	DL3	DL4	DL5	DP2	DD1	AT1	DP3	S1:4	DP3	DP4	DP5	DP6	AR1	S2:4	DL6	DL7	DL8	DL9	S3:4	DD2	DD3	DD4	DD5	S4:4				

E10의 ‘자료 보고 설명하기’ 문항 응답에 대한 채점 사례에서 평균적인 수준으로 채점을 한 R05(M=4)와 엄격한 경향이 나타난 R13(M=3.5)의 평균 점수에는 큰 차이가 없었다. 그런데 <표 Ⅳ-24>를 통해 해당 사례의 채점 과정을 살펴보면, R13은 R05와 달리 ‘발음 구사 능력’에 관한 근거(DP)와 점수 결정을 위한 가정이 나타나지 않았음을 알 수 있다.

R13과 R05의 전반적 수행에 관한 채점 과정에서는 고려하는 근거의 차이가 나타나는데, R13(#R131032)은 발음에 관한 수험자 응답에 나타난 특징을 고려하지 않으면서 전반적인 응답에 대한 인상으로 점수를 결정하는 반면에,

R05(#R051032)는 발음에 관한 측면을 고려하고 있음이 채점 과정 보고에서 나타타났다.

#R131032

아 이게 상승 속도가 아닌데 속도로 본 거 같아
인상률의 그런 높아지고 낮아지고 얼만큼 인상이 되는지 많이 되는지 적게 되는지인데 그거를 조금 그렇게 때 그거랑
초반에는 밑에 막대그래프는 설명을 잘했는데 나중에 그 인상률 그래프 설명이조금 맞지 않아서
응답에서 과제와 관련하여 어느 정도 적절한 내용을 다루고 있다고 볼 수 있기때문에
보통 전반적인 수행은 보통을 3점을 줘야 될 것 같고

#R051032

어… 지금 이 학생의 경우에는 막대 그래프와 그 꺾은선 그래프의 2개의 내용도파악을 하면서 또 어 국내 어떤 정세와 관련된 내용과도 연관 지어서 이야기를 어 하고 있습니다
그래서 사실 전반적 수행 자체를 봤을 때 탁월 5점 아니면 우수 4점을 줘도 큰 문제는 없을 것 같은데
음 중간 중간에 발음과, 발음이 좀 어색하고 좀 매끄럽지 못한 부분이 조금 있어서 5점까지는 어려울 거 같고
전반적 수행은 그냥 4점 정도를 주는 것이 적절해 보입니다.

4번 문항(기사에 나타난 문제와 해결 방안 말하기)에서는 채점자 R11이 수험자 E08에 대한 채점에서 엄격한 경향이 나타났다. 이와 관련하여 평균적인 채점을 한 R06과의 채점 과정을 비교하였다(<표 Ⅳ-25> 참조).

<표 Ⅳ-25> 엄격한 채점 특성이 나타난 채점 과정4(R11-E08, 4번 문항)

채점자		채점 과정																														
엄격한	R11	DD1	DL1-6	DD2	DDL7	DDL8	DDL9	DDD3	DDL10	DDL11	DDL12	DDD4	DDL13	DDL14	S3:2	DDP1	DDP2	S2:3	DDD5-8	S4:3	DAR1-5	DDD9	S1:2									
평균	R06	DDD1	DDL1	DDD2	DDD3	DDD4	DDD5	DDD6	DDD7	DDD8	DDD9	DDD10	S1:4	DDP1	DDP2	S2:4	DDL2	DDL3	DDL4	DDL5	S3:4	DDD11	DDD12	DDD13	S4:4							

채점 과정 보고 분석의 결과, 채점자 R11은 E08의 응답을 채점하면서 어휘·문법적인 측면(DL1-14)을 주목하여 채점을 하는 경향이 나타났다. 해당 준거에 대한 채점을 통해 R11은 응답의 수준을 미흡 수준(2점)에 해당하는 것으로 인식하게 된 것으로 보인다. 평균적인 결과가 나타난 R06의 채점 과정에서는 응답의 어휘·문법적인 측면보다 담화적인 측면(DD1-13)을 중심으로 채점이 이루어졌음을 확인할 수 있다(<Ⅳ-25> 참조). 두 채점자는 채점 과정 중 동시적 보고에서부터 주목하는 근거 유형의 차이를 나타내고 있었다(#R110841, #060841).

#R110841

발생하고 있는 문제는.. 문제 먼저 얘기하고
높은 기온
죽는 사람이 있습니다
늘어나고 있습니다?
해결방법
가까운 일부터 시작하고.. 분리수거
일회용품..
작은 일부터 하자는 거지?
근데 어휘 문법이 매우 부족했고
개인적인 일.. 그 다음은 기업
커피 일회용컵..
시민?
쓰레기 버리지 못하게 하는 것?
갑자기 어휘와 문법이 굉장히 단순해 졌고
이거는 자기가 자신이 없으니까 쉬워진 거 같고
배려..
정부의 임무..

#R060841

신문의 문장을 그대로 읽고 있어요
기후 변화를 막아요?
분리수거
일회용품을 어떻게 재활용해요?
쓰레기를 어떻게 버리질 못해요?
어떻게 마련해요?

4번 문항에서 채점자가 엄격한 특성을 나타낸 사례(R11)에서는 어휘·문법적인 측면에 주목하는 경향이 나타났다. 해당 문항 유형은 수험자가 신문 기사를 바탕으로 나타난 문제가 무엇이며, 그에 대한 해결 방안이 무엇인지를 제시하는 문항이다. 이러한 읽기 자료를 사용하는 통합적인 말하기 능력을 평가하는 문항의 채점 과정에서 언어 사용의 정확성에 초점을 두었을 때 엄격한 채점이

이루어질 수 있음을 보여주는 사례이다.

② 인상 기반 접근으로 인한 관대한 채점 경향

MFRM 분석 결과에서 R09를 제외한 모든 채점자는 관대한 채점 경향을 나타냈다. 그 중에서도 관대한 특성이 두드러진 사례를 중심으로 채점 과정의 차이를 알아보았다. 먼저 채점자 R11은 수험자 E05의 1번 문항의 평가 결과에서 관대한 경향이 나타났다. 이와 관련하여 해당 문항에서 평균적인 채점을 한 R08과 상대적으로 엄격한 채점을 한 R04의 채점 과정의 차이를 알아보았다.

<표 Ⅳ-26> 관대한 채점 경향이 나타난 채점 과정1(R11-E05, 1번 문항)

채점자		채점 과정																					
관대한	R11	DP1	DL1-5	DP2	DP3	DP4	S2:4	AF1	AR2	DL6	S3:4	DC1	DC2	DC3	DC4	S4:4	DC5	AR3	DP5	DP6	S1:4		
평균	R09	DP1-5	AF1	S2:3	DL1	DL2	DL3	DL4	S3:3	DC1	DC2	DC3	AF2	S4:3	DP6	DP7	DL4	S1:3					
비교	R04	DL1-5	DL1	DP1	DP2	DP3	DP4	DP5	S2:2	DL6	DL7	AF1	S3:1	DL8	AR2	DC2-5	AF3	S4:1	DC6	AF4	S1:2		

<표 Ⅳ-26>에 제시된 채점자별 채점 과정을 비교해 보면, 관대한 채점을 한 R11은 평균적인 채점을 한 R09와 평가 준거별 근거 언급의 빈도와 관련하여 큰 차이를 나타내지 않았다. 다만 평균 집단 및 대조 집단이 채점 과정 초반에 발음에 관한 보고가 많았다면, R11은 어휘·문법에 주목한 것으로 나타났다. 관대한 채점 특성이 채점 과정에서 어떻게 나타난 것인지를 구체적으로 확인하기 위하여 관대한 채점을 한 R11과 대조적으로 엄격하게 채점한 R04의 채점 과정 보고의 내용을 비교하였다(#R110512, #R040512).

#R110512

그 다음에 전반적 수행 보면

완벽하진 않지만 적절하게 다루고 있었고

약간 실수 있지만 명료하고 유창하게..

유창.. 유창.. 유창성은 다소 부족했던 거 같은데

실수가 지속적으로 나타나진 않았으니까

4점

#R040512

전반적인 수행. 경험에 대해서 어려운 점 이야기..
그냥 발음이 어렵다라는 거 말고는 지금 뭐 경험 이야기한 게 없는데? 들은 게 없는데 경험을.. 그죠?
응답에서 과제를 다루고 있지만 제한적으로 주제를 다루고 있음, 표현을 이해할수 있으나 전반적으로 발화 전달과 담화 응집성에 문제가 있음..
1점까진 아닌 거 같아요 이 표를 보니까

2점 주면 되겠네 전반적인 수행은 이렇게 할게요

‘전반적 수행 능력’에 관한 채점 과정에서 R11은 구체적인 응답 내용을 고려하기 보다는 채점 척도에 기술된 기준만을 고려하려는 경향을 보였다면, R04는 과제에 대한 수험자의 응답 내용에 대한 이해를 바탕으로 채점 척도를 연결하는 것을 알 수 있다. R11의 경우에는 구체적인 응답 내용과 특징을 떠올리지 않으면서 감점의 원인인 문제 상황에 주목하지 않았다. 그러나 상대적으로 엄격한 채점을 한 R04의 경우에는 특정 문제(“그냥 발음이 어렵다는 거 말고는 지금 뭐 경험 이야기한 게 없는데?”)를 중심으로 채점 척도와 연계하여 점수를 결정한 것으로 나타났다. 이는 채점 과정에서 응답의 내용보다 인상이 중심이 되면서 감점을 받지 않은 것이라고 볼 수 있다.

2번 문항에서 관대한 채점 경향이 나타난 사례는 채점자 R11이 수험자 E12의 채점에서 확인할 수 있었다. R11의 채점 과정은 평균적으로 채점을 한 R10과 상대적으로 엄격한 채점을 한 R09와의 채점 과정 비교를 통하여 파악하였다.

<표 Ⅳ-27> 관대한 채점 특성이 나타난 채점 과정2(R11-E12, 2번 문항)

채점자		채점 과정																										
관대한	R11	DP1	DL1	DP2	DL2	DD1-4	AR1	AR2	DD5-7	S4:3	DP3	S2:5	AR3	DL3	S3:5	AT1	DD8	DP4	S2:4	AR4	DP5	DD9	AR5	DD10	S1:4			
평균	R10	DL1	DL2	DL3	DL4	DL5	AE1	DL6	DD1-6	S4:4	DL7	DL8	AR1	S3:3	DP1	DP2	S2:4	DD7	DD8	AR2	DD9	DP2	S1:3					
비교	R09	DD1	DP1	DP2	DP3	S2:3	DL1	DL2	AR1	S3:1	DD2	DD3	DD4	S4:1	DD5	DD6	S1:0											

<표 Ⅳ-27>에 제시한 2번 문항에서 관대한 채점을 한 R11의 채점 과정 보고에서는 응답에 나타난 어휘·문법 사용에 관한 언급(DL)이 평균 수준 채점자(R10)에 비하여 적은 것으로 나타났다. 또한 점수 결정과 관련하여 채점 척도(AR)를 자주 근거로 언급하는 모습도 나타났다. 대조적으로 같은 수험자의 응답에 대하여 엄격한 채점을 한 사례에서는 담화적인 측면에 대한 근거를 비율적으로 더 많이 고려하고 있는 것으로 나타났다. 관련하여 청취 과정에서 이루어진 동시적 보고를 확인한 결과, 관대한 채점 경향이 나타난 R11은 주로 내용적인 측면과 좋은 발음을 주목한 반면에 R10은 어휘·문법적인 오류를 주목한 것을 알 수 있다(#R111221, #R101221).

#R111221

와 있었고..
교육을 배우고 싶고
발음이 좋네
였잖아요.. 아는 선생님인가?
독일어
내가? 아.. 자기의 경험을 미리 얘기 했고
그 다음에 이제 해결.. 일을 해보고
아 자기의 경험을 얘기하느라고 조언을 제대로 못해줬네

#R101221

와 있었고 와 있고 라고 말해야 되고
문법 달르는데 문법 오류
가지게 됐는데 라고 말해야 됐는데 오류
조사 사용 오류
나을까 싶어서 마지막 문법 이상한데
나을 것 같습니다 이렇게 마무리를 하든가

자료 사용을 포함하는 문항에서 나타난 관대한 채점의 경향이 채점 과정을 통해 어떻게 나타나는지를 확인하기 위하여 3번 문항에서 관대한 채점을 한 채점자 R07과 같은 수험자 응답에 대해 평균적으로 채점한 R11의 채점 과정을 비교하였다.

<표 Ⅳ-28> 관대한 채점 특성이 나타난 채점 과정3(R07-E02, 3번 문항)

채점자		채점 과정																														
관대한	R07	DP1	DP2	DP3	DL1	DL2	AE1	DD1	DP4	S2:4	DL2	DL3	S3:5	DD2	DD3	S1:5	S4:5															
평균	R11	DL1	DD1	DD2	DP1	DD3	DD4	DL2-8	DP2	DP3	DL8	DL9	DD5	DP4-8	AR1	S2:3	DP9-2	중략	DL12	DL13	DD8	DD9	AR3	S4:3	AR4	S1:3					

<표 Ⅳ-28>에서 관대한 채점 경향이 나타난 채점자 R07과 평균적인 채점을 한 R11의 채점 과정을 비교해 보면, R07은 응답 청취 중에 발음 특성 주목을 하고 있는 것에 반해 R11은 담화와 어휘·문법적인 측면을 주목하고 있는 것으로 나타났다. 또한 점수를 결정할 때마다 채점 척도(AR)를 고려한 R11과 달리 R07은 채점 척도에 관한 언급이 나타나지 않고 있다.

앞서 3번 문항에서 채점의 엄격한 경향이 나타난 경우에는 응답의 정확성에 초점을 두고 발음 요소를 고려하지 않는 경향이 나타났었는데, 관대한 채점 경향이 나타난 R07은 평균적인 수준인 R11과 달리 발음 요소를 채점 근거로 활용하였으며, 어휘·문법이나 담화 구성과 같은 정확성을 측정할 수 있는 준거들은 구체적으로 고려하지 않는 양상이 나타났다.

③ 점수 결정의 가정 고려로 인한 특정 점수 집중 경향

본 연구에서 실시한 MFRM 분석에서 집중 경향이 나타난 점수는 3 ~ 5점으로 각각 문항에서 보통, 우수, 탁월 수준에 해당하였다. 특정 점수를 선택하려고 하는 채점자의 심리적인 경향인 집중 경향이 채점 과정에서 어떻게 나타나는지를 알아보기 집중 경향적 채점 경향이 나타난 채점자가 다른 채점자들과 달리 해당 점수에 집중적으로 부여한 사례를 확인하였다.

먼저 R05는 척도 중 '보통' 수준에 해당하는 3점을 부여한 비율이 가장 높은 채점자였다(36%). R05가 3점을 부여한 사례 가운데 다른 채점자들과 점수 열 상의 차이가 두드러졌던 수험자 E01의 1번 문항의 채점 과정을 확인하였다.

<표 Ⅳ-29> 특정 점수에 대한 집중 경향이 나타난 채점 과정(R05-E01, 3점)

채점자		채점 과정																				
집중된	R05	DL1	DL1	DL2	DL3	DL2	DL3	DP1	S1:3	AE1	DP2	DP3	S2:3	DL4	DL5	AR1	S3:3	AE2	DL4	DL5	AR2	S4:3
평균	R10	DL1-5	DL1	DL6	DL2	DL7	DL3-6	DL8	DL9	DP1	DL10	S3:4	S4:5	AE1	DL11	AR1	S2:4	DL7	AR2	AR3	DL8	S1:4

<표 Ⅳ-29>에서 R05는 해당 수험자 응답에 대한 채점 과정에서 모든 평가 준거에 3점을 부여하고 있었는데, R10보다 점수 결정의 가정으로 채점자의 경험(AE)을 더 고려하고 있었고, 가정으로서의 채점 척도는 더 적은 횟수로 고려한 것으로 나타났다. 또한 어휘·문법적인 측면(DL)에 대한 언급이 많았던 R10과 달리 R05는 상대적으로 적은 횟수를 언급한 것으로 나타났다. 채점자가 자신의 교육 및 평가 경험을 채점 과정에서 활용하는 경우는 기존 신념 체계가 작동하는 것으로 볼 수 있는데, R05는 그 중에서 가장 긴 교사 경력(18.7년)을 갖고 있었다. 3점에 집중하는 경향을 갖도록 하는 또 다른 측면은 '3점'이 중앙값이기 때문에 불확실한 상황에서 점수를 선택해야 할 때, 손해가 적은 대안이라는 점이다.[50] R05와 R10이 발음과 관련하여 사용한 가정은 서로 다른

결론을 내리는데 기여하고 있었는데, E01의 발음에 대해 R05가 가진 가정은 부정적인 것(“모국어의 영향을 많이 받는 거 같고...”)인데 반해 R10의 가정은 긍정적인 것(“발음은 특별히 고향 언어 때문에 방해 받는 아주 큰 요인들은 없었고”)이었다. 이러한 가정의 차이는 채점자가 이전의 교육 및 평가 경험 등으로부터 형성한 신념 체계로부터 나타난 것인데, 해당 문항의 채점자 평균 점수를 기준으로 보았을 때, R05는 척도 사용을 제한하면서 수험자의 응답에 대해 엄격한 채점 경향을 나타내게 되었다.

다음으로 전체 채점자가 결정한 점수 중에서 가장 집중적인 채점 경향이 나타난 것은 채점자 R13이 5점을 선택한 비율(42%)이었다. 5점은 중앙값이 아니며, 가장 높은 수준인 ‘탁월’에 해당하는데, 응시자들이 한국어 4~7급 수준의 학습자들이기 때문에 5점을 받을 수 있는 수험자가 제한적일 것이라는 예상과 다른 양상을 나타낸 것이었다.

<표 Ⅳ-30> 집중 경향이 나타난 채점자의 문항별 채점 과정(R13-E03, 5점)

문항	채점 과정																								
1	DP1	DL1	DL2	DL3	DL4	DP2	DP3	DL1	DL2	DL5	DL3	DL4	S3:5	S4:5	DP4	S2:5	S1:5								
2	DL1	DL1	DL2	DL2	DL3	DL4	DL5	DP1	DL6	S4:5	S1:5	DP1	DP2	DP3	S2:5	DL3	DL4	S3:5							
3	DP1-4	DL1	DL1	DL2	DL3	DL4	DL2	DL3	S1:5	DP5	DP6	S2:5	DL5-1	S3:4	DL12	DL4	DL5	DL6	AT1	S4:5					
4	DP1	DL1	DL2	DL1	DL3	DL2	DL3	AT1	DP2	DP3	S2:5	DL4	DL5	DL6	DL4	S3:5	DL7	DL8	S4:5	DL9	DL10	S1:5			

<표 Ⅳ-30>에 제시한 R13의 채점 과정에 나타난 가장 큰 특징은 채점 척도나 채점자 경험, 평가적 신념 등의 가정(AR, AE, AO)을 고려하지 않는다는 점이다. 또한 ‘담화 구성 능력’(S4)과 ‘전반적 수행 능력’(S1)의 점수가 모두 5점으로 같았다는 점, 그리고 1번과 2번 문항의 채점에서는 연이어서 평가 준거의 점수를 결정하는 모습도 나타났다. 채점 과정에서 채점자의 ‘가정’은 점수 결정을 포함하는 경험적인 근거로서 개인적으로 이루어지는 말하기 평가의 채점

50) 이러한 현상은 사람이 어떤 것을 선택할 때 이익보다는 손해를 먼저 고려한다는 신경경제학의 입장에서도 이해가 가능하다(이선애·박선철, 2013).

과정에서 기억 체계를 바탕으로 상호작용하면서 자신의 판단과 점수 결정을 반성하는 사고 과정과 관련이 있다. R13은 채점 과정에서 가정을 고려하지 않으면서 관대한 채점의 경향이 나타났는데, 이는 채점 과정에서 상호작용할 수 있는 평가적인 가정을 고려하지 않으면서, 자신의 직관적인 사고로 점수를 결정하였기 때문으로 볼 수 있다. 이와 관련하여 R13의 채점 과정에는 수험자의 응답에 관한 내용보다는 주로 채점자가 들으면서 직관적으로 판단한 내용이 주로 나타나고 있었다(예:“요목 조목 다 이야기했기 때문에...”, “유창성이 아주 뛰어났고 응집성도 뛰어났어”, “약간 당황스럽긴 했지만”, “주장에 흔들림이 없어.”)

④ 평가 결과를 제한하는 채점 척도 사용 경향

채점자가 0 ~ 5점의 척도에서 특정 점수만을 제한적으로 사용하는 것은 수험자의 응답을 평가할 수 있는 점수 사용을 제한하는 심리적인 기제로부터 발생한다. 채점자 가운데 R05는 전체 MFRM 분석 결과에서 내적합도가 가장 높은 채점자(Infit: MnSq: 0.63)였으며, 척도별 점수 빈도에서 5점을 한 번도 부여하지 않은 것으로 나타났었다. 이와 관련하여 전체 평균이 4.83으로 거의 대부분의 채점자가 모든 평가 준거 영역에서 5점을 부여한 수험자 E03의 1번 문항 채점 과정에서 채점자 R05가 제한적으로 척도를 사용하는 원인을 확인하여 보았다.

<표 Ⅳ-31> 제한적 척도 사용 경향이 나타난 채점 과정(R05-E03, 1번 문항)

채점자		채점 과정																							
제한된	R05	DP1	DP2	DD1	AT1	DD2	DP3	S1:4	DP4	DP5	DP6	S2:4	DL1	AE1	DL2	S3:4	DD3	DD4	DD5	AR1	S4:4				
평균	R09	DP1	AE1	DP2	DP3	S2:5	AE2	DL1	DL2	DL3	DL4	S3:5	DD1	DD2	DD3	S4:4	DD4	DD5	AR1	DD6	DD7	DD8	S1:5		

<표 Ⅳ-31>에서 R05는 수험자 E03의 1번 문항 응답에 대해 모든 평가 준거에 4점을 부여하고 있었으며, 점수 결정의 가정으로 과제(AT), 채점자 경험

(AE), 평가 준거(AR)를 고려하였으며, 평균 수준으로 채점한 R09와 비교하였을 때 어휘·문법(DL)을 상대적으로 적고, 발음(DP)을 근거로 고려한 횟수가 많은 것으로 나타났다. 실제 채점 과정 보고 내용을 보면, R05는 점수 결정 관련 발화에서 불확실한 느낌을 갖고 있지만, '탁월' 수준(5점)은 전혀 고려하지 않은 것으로 나타났다(#050312).

#050312
문법이 가장 어렵고 표현이 조금 어렵다는 이야기를 하고 있고
지금 이 학생 같은 경우에는 전반적 수행은 과제를 그래도 상당히 적절하게 수행하고 있는 것으로 보여집니다.
어...이 학생은 한국어를 배울 때 문법이 어려운데 그 이유가 문법이 굉장히 여러 가지 의미를 가지고 있기 때문에 어렵다라는 얘기도 했고
그 다음에 표현도 어렵지만 지금도 계속 계속 표현을 어렵다, 나름대로 명료하고 유창하게 이야기를 하고 있어서
전반적 수행은 4점까지 줘도 나쁘지 않을 것 같습니다
어 다음에 발음은 중간에 한두 번 정도 같은 말을 어휘를 반복하는 경우가 있었고
중간에 약간 휴지가 있기는 하지만
그게 이렇게 듣기에 어려운 정도가 아니었기 때문에
이것도 4점 정도를 줘도...될 거 같습니다.
다음에 어휘와 문법도 대체적으로 주제와 관련된 어휘를 사용하고 있고
중급 이상에서 사용 될 수 있는 부사적 부사 표현이라든가 좀 내용을 다양하게 사용하고 있어서
3급보다는 3점보다는 좀 더 줄 수 있…4? 우수하다고 보여집니다.
다음에 흐름이 대체적으로 일관되고
주제에 대해서도 대체적으로 논리적으로 전개를 하고 있기 때문에
그 다음에 마무리도 급하게 끝날 듯 하지만 그래도 어느 정도 문장 연결에 짜임새가 있는 구조라고 여겨져서
이 학생은 전반적으로 4점 정도 줘도 충분한 언어 실력을 가지고 있다고 생각합니다.

이러한 채점 척도의 제한적 사용은 해당 채점자의 채점 방식이 예측 가능하며, 따라서 평가 결과에 대한 예언 가능성(predictability)이 지나치게 높은 것(Linacre, 2002: 878)이며, 채점자는 수험자의 응답 사례에 나타난 여러 특징들을 자신의 채점 가능 범주로 조정하여 인식한 후에 채점을 수행하는 것을 의

미한다. 해당 채점자에게 배당된 수험자의 응답은 과제에 대하여 충분한 수행이 이루어졌음에도 최고 수준의 점수를 받지 못할 가능성이 있지만, 채점자는 이러한 점수 결정 경향을 통해 다른 채점자들과의 적당한 점수 거리를 유지할 수 있다는 이점이 있다. R05가 가장 한국어교육 경력이 많은 채점자라는 점을 고려하였을 때, 이러한 결과는 채점자 집단 내에서의 안정을 유지하기 위한 행위로도 해석할 수 있다. 즉, 복수의 채점자가 존재하는 채점 상황을 반복적으로 경험하면서 다른 채점자들과 비슷한 수준의 채점 일관성 및 신뢰도를 확보하기 위하여 수험자의 수행 결과를 극단 값으로 해석하지 않는 경향이 내재화되어 나타날 수 있다는 것이다. 이러한 내재된 채점 척도 사용의 경향은 R05와 비슷하게 채점 경력이 긴 편인 R06도 MFRM 분석 결과에서 과적합 경향, 즉 채점 척도를 제한적으로 사용하는 것으로 나타난 것을 통해서도 확인할 수 있다(Infit MnSq: 0.63).

⑤ 무선적인 평가 결과를 가져오는 직관 기반 채점 경향

문항 특성 측면에서는 나타나지 않았지만 R08은 어휘·문법에 대한 채점에서 상대적으로 더 관대한 채점을 하였으며, R04의 경우에는 발음과 관련하여 엄격한 채점을 한 것이 채점자×평가 준거의 상호작용에서 나타난 바 있다. 이와 관련하여 R08의 어휘·문법 채점 점수열과 R04의 발음 채점 점수열에서 다른 채점자들과 차이를 확인하고, 해당 채점 과정을 분석하였다(<표 Ⅳ-32> 참조).

<표 Ⅳ-32> 편향성이 나타난 채점자의 채점 과정(R08-E08, 1번 문항, 어휘·문법)

채점자		채점 과정																												
관대한	R08	D P 1	D L 1	D D 1	D P 2	D D 2	D D 3	D D 4	A R 1	A T 1	S 1: 5	D P 3	D P 4	D P 5	D P 6	S 2: 4	D L 2	S 3: 3	D D 5	D D 6	D D 7	S 4: 3								
평균	R06	D P 1	D P 2	D D 1	D D 2	S 1: 4	D P 2	D P 3	S 2: 5	A E 1	S 3 :4	D D 3	A R 1	D P 4	S 4: 4															
비교	R04 R01 R02	D D 1	D P 1	D D 2	D P 2	A O 1	D P 3	D P 4	D D 3	A R 1	A R 2	D L 1	D L 2	A O 2	D L 3	D L 4	S 3 :5	S 4 :5	D P 5	A R 3	D P 6	D P 7	S 2 :5	D D 6	D P 8	D D 7	S 1 :5

R08은 수험자 E08의 1번 문항에서 어휘·문법 점수가 MFRM 모형에서 기대하는 것과 다른 결과를 나타냈다. <표 Ⅲ-20>에서 채점자 R08은 어휘·문법 점수 결정 과정에서 적은 근거(DL1-2)를 바탕으로 점수 결정에 관하여 상호작용하는 가정의 고려가 없이 해당 평가 준거의 점수를 결정한 것으로 나타났다. 이는 결과적으로 어휘·문법 채점(S3)에서 MFRM 모형의 기댓값보다 1.8점이 낮은 점수를 부여한 것으로 나타났는데, 말하기 평가에서 어휘·문법 사용에 대한 판정은 오류를 중심으로 계량적인 접근이 가능한 준거라는 특징이 있다. 만약 주목한 사례 수가 적거나, 기억 체계가 바르게 작동하지 않았을 때는 불확실한 판정이 이루어질 수 있다. 따라서 언어적 측면에 관한 채점에서는 단기 기억을 연장시킬 수 있는 활동(예: 메모하기)이 도움이 될 수 있다.

R12는 수험자 E04의 1번 문항에서 '전반적 수행', '담화 구성 능력', '어휘·문법'에 대한 점수 결정에서 채점자 편향이 나타나면서 MFRM의 예상보다 약 1.8점 낮게 점수를 부여한 것으로 나타난 바 있다(<표 Ⅲ-20> 참조). 이와 관련하여 해당 채점자의 채점 과정을 확인하였다.

<표 Ⅳ-33> 채점자 편향이 나타난 채점 과정(R12-E04, '전반적 수행', '어휘·문법', '담화')

채점자		채점 과정															
편향된	R12	DP1-10	S2:4	AE1	DD1	DD2	S3:3	DD3	DD4	DD5	AE2	S4:3	DL1	DP11	S1:3		
비교	R11	DP1	DL1	DP2	DP3	DP4	DD1	DD2	S2:5	DL1-3	DP5	AE1	S3:5	DD3-5	S4:5	DD6	S1:5

<표 Ⅳ-33>의 MFRM 분석에서 모형의 예상과 거리가 먼 채점을 한 것으로 나타난 R12는 채점 과정에서 채점 근거로서 발음에 관한 고려가 절대적으로 많았으며(DP1-11), 어휘·문법에 관한 근거는 1회 고려하였다는 것이 나타났다. '담화'의 채점과 관련하여서는 고려한 근거의 수는 비슷했으나, 채점자의 관련 경험을 가정(AE)한 것에서 R11과 차이가 있었다.

#R120412

세 번째로 담화 구성은 어, 주제에 대한 아.. 발화를 하고 있기는 하지만
내용이 굉장히 짜임새가 있다든가
어, 굉장히 주제에 어울리는 적절한 정보를 담고만 있다고 느껴지지는 않아서
그냥 보통 수준의 학생, 보통 수준의 발화였다고 어, 느껴진다
따라서 마찬가지로 담화 구성 점수도 3점을 주도록 하겠다

#R110412

그 다음에 담화 구성을 보면
흐름이 일관되며
처음에는 뭐에 대해서 얘기했지
처음에는.. 어휘가 어려웠다 자긴 중국 원어민이지만 그래도 어휘가 어려웠다 그렇게 얘기했지
담화 구성도 적절했고..
그러면 담화도 5점

채점자 R12와 R11이 수험자 E04의 1번 문항 응답에 대한 담화 구성 능력을 채점하는 과정을 살펴보면, MFRM 모형의 예상에서 어긋난 채점이 이루어진 R12가 고려한 채점자의 경험(AE)이라는 가정이 전문적인 내용과는 거리가 먼 정보로 이루어져 있음을 알 수 있다("그냥 보통 수준의 학생, 보통 수준의 발화였다고, 어 느껴진다."). 채점자가 갖고 있는 '보통 수준'이라는 개념은 한국어 말하기 능력 수준에서, 그리고 담화 구성 능력에서 구체적으로 어떠한 것을 의미하는 것인지는 드러나지 않는다. 하지만 앞선 보고에서 '담화의 짜임새'와 '주제에 어울리는 적절한 정보'를 언급한 것으로 보았을 때 '담화 구성 능력'의 채점 기준을 읽으면서 무언가 부족하다고 느낀 점이 있었던 것으로 보인다. 그러나 구체적으로 그것이 무엇이며, 무엇으로부터 그러한 인식을 갖게 된 것인지를 언급한 내용이 없는 것으로 보아 직관에 의한 채점 수행이 이루어지는 중에 '보통 수준'이라는 모호한 개념을 가정으로 고려한 것으로 보인다.

R13의 경우에는 고전검사이론에 따른 신뢰도 분석에서 채점자 간 신뢰도에 낮은 것으로 나타났었다. 또한 앞서 이루어진 채점 과정 분석에서는 가정을 고려하지 않으면서 단정적인 채점을 하였으며, 5점을 집중적으로 부여하는 경향이 나타났다. 이와 관련하여 E02에 대한 채점 과정 보고에서 가정 고려에 따

른 채점 과정의 영향을 확인하기 위하여 평균적인 채점을 한 R04와의 R13의 채점 과정 보고를 비교하였다(#R130212, #R040212).

#R130212

어휘나 문법 그렇게 어려운 건 쓰지 않았지만 이 문제에 대해서는 어려운 단어가 필요하지 않고
문법도 근데 어떻게 받침을 하는지도 모르고 발음을 하는지도 모르고
이런 사소한 오류가 있었으니깐 그냥 5점을 줘도 될 것 같아

#R040212

뭔가 정확한 어휘를 사용하기보다 한국 사람들을 사랑..사따라하면 되는데 뭐 그렇게 하는 게 어려웠어요 이런 식으로 정확한 어휘가 생각나지 않아서 그렇게, 하면은 이런 식으로 그렇다, 뭐하다 이런 좀 뭐라고 하지 어휘들을 많이 사용해서
정확한 표현을 하지 못한 거 같..
그리고 문법 표현 수준도 좀 고급에 해당하는 표현은 거의 나오지 않은 거 같아서
뭐..그 한 2,3급정도? 수준의 어휘가 문법만 사용이 된 걸로 지금 들렸기 때문에
2점이나 3점을 주면 될 거 같은데 표를 보면 미흡 수준이 주제와 관련하여 사용은 어휘가 단순하고 사용한 문형이 제한적이다, 심각한 오류가 일부 나타나며 의미 이해해 어려움이 있다 그리고 3점 수준은 다소 제한적으로 주제와 관련된 어휘와 복잡한 문형을 사용함, 표현에 반복적인 오류가 있으면 의미 이해를 다소 방해함
그 어떤 문장에서 거슬릴 정도의 큰 오류가 나타났던 것은 아니지만
전체적인 표현 수준이 2,3급 정도 수준에 머물러 있었고
그래서 2점? 줄 수 있을 것 같아

R13이 1번 문항에서 ‘어휘·문법 사용 능력’에 관한 채점 과정을 보고한 내용을 살펴보면, 평가 맥락과 관련하여 존재하는 가정이나 채점과 관련된 전문적인 경험을 떠올리기 보다는 ‘그렇게 어려운 건’, ‘어려운 단어가 필요하지 않고’, ‘사소한 오류’와 같은 자신의 직관에 의한 가정을 중심으로 채점을 하는 경향이 나타나고 있었다. 채점 과정에서 이러한 접근이 이루어지면서 채점자들이 공유하고 있는 근거 및 가정의 체계가 흔들린 결과로 R13의 채점자 간 신뢰도가 낮은 것으로 나타난 것이다.

이상에서 살펴본 채점자 영향에 대한 말하기 평가의 채점 과정 분석의 결과를 정리하면 다음과 같다. 첫째, 특정 채점자의 엄격한 채점 경향은 채점 과정

에서 총체적인 평가 준거를 먼저 채점할 때 낮은 수준으로 판단이 이루어지면서 이후의 분석적인 접근을 요하는 평가 준거들에 대한 판단에도 영향을 미치면서 나타날 수 있다. 또한 분석적인 평가 준거 가운데 '담화 구성'에 관한 준거도 채점 과정에서 점수 결정의 근거를 여러 번에 걸쳐 인식하였을 때 엄격한 결과를 나타낼 수 있으며 이는 독립적인 말하기 과제의 사례에서 확인할 수 있었다. 둘째, 채점 과정에서 판단 근거를 확보하지 못한 평가 준거에 대해서 엄격한 채점이 이루어질 수 있으며, 이러한 경향은 '발음'에 대한 채점 과정의 분석에서 나타났다. 반대로 많은 판단 근거를 인식하면서 평가 준거에 대한 엄격한 채점 경향이 나타난 경우가 있었는데, 이는 자료를 바탕으로 담화를 구성하는 문항에서 어휘·문법 측면에서 확인되었다. 같은 사례에서 담화 구성 측면의 근거를 상대적으로 많이 고려한 채점자는 평균 수준의 채점을 한 것으로 나타났으며, 이러한 차이는 말하기 평가 과제의 형식에 따라서 채점에 미치는 영향이 평가 준거에 따라 다르게 나타날 수 있다는 것을 보인 것으로 판단된다. 셋째, 지나치게 관대한 채점을 한 것으로 나타난 사례의 채점 과정 분석에서는 수험자의 응답에 나타난 구체적인 내용이나 특징을 판단 근거로 고려하는 양이 적으면서 주관적으로 인상 중심의 채점을 하는 것으로 나타났다. 넷째, 특정 점수에 대한 집중 경향은 채점자가 채점 척도 사용에 관하여 적용하는 평가적인 가정으로부터 기인한 것으로 보인다. 척도의 중앙에 해당하는 3점에 집중하는 경향을 가진 채점자는 연구 참여자 중 가장 한국어교육 경력이 긴 채점자였으며, 수험자의 응답에 나타난 문제를 찾으면서 채점을 하는 경향이 있었다. 가장 높은 점수를 집중적으로 주는 경향도 확인할 수 있었는데, 이는 관대한 채점 경향과 마찬가지로 판단의 근거가 부족한 상황에서 평가적인 가정도 고려하지 않는 채점자의 사례를 통하여 확인할 수 있었다. 다섯째, 평가 척도를 제한적으로 사용하는 경향이 있는 채점자의 채점 과정 분석에서는 특정 점수를 한 번도 부여하지 않는 채점자의 사례에서 의식적으로 지나치게 후하거나 엄격한 채점을 하지 않으려는 경향이 있음을 확인할 수 있었다. 끝으로 직관에 의한 접근으로 무선적인 채점을 한 사례에서는 관대한 채점을 한 것과 마찬가지로 수준 판단을 위해 고려하는 근거의 양이 적은 것으로 나타났는데, 응답에 대한 인상과 가정을 고려하여 접근하기 보다는 자신의 불확실한 직관적인 인식을 바탕으로 접근한다는 점에서 차이가 있었다.

V. 말하기 평가의 채점자 교육을 위한 채점 과정 모형의 적용

본 장에서는 말하기 평가의 채점 과정에 관한 이론적 접근을 통해 구안한 채점 과정 모형과 말하기 평가 채점 과정과 관련하여 실시한 채점자 영향 및 채점 과정에 대한 분석 결과를 바탕으로, 채점 과정 기반의 한국어 말하기 평가 채점자 교육 방안을 제시하고자 한다. 채점 과정 기반의 말하기 평가 채점자 교육의 전제는 말하기 평가의 채점 과정이 실제적으로 평가 결과를 변화시키는 데 영향을 미치며, 채점자로 인한 영향이 채점 과정을 통해 문항 유형이나 평가 준거, 척도 등에 따라 체계적으로 나타날 수 있다는 점이다. 말하기 평가에서 채점자는 수험자의 과제 수행에 대하여 합리적인 접근을 통해 타당한 점수를 결정할 수 있어야 하는데, 이러한 접근을 위해서는 채점을 수행하기 이전에 자신의 채점 경향이 어떠한지를 성찰하고, 점수 결정을 위한 증거의 탐색과 판단을 위한 전문성을 갖추어야 할 필요가 있다.

1. 채점자 영향과 채점 과정 분석의 함의

본고에서 한국어 말하기 평가의 채점 과정에 실증적으로 접근하기 위하여 수행한 채점자 영향 및 채점 과정 분석의 결과에서 채점자 교육과 관련하여 시사하는 바를 살펴보았다. 채점자 영향에 관한 분석 결과에서는 평가의 결과를 바탕으로 채점자의 채점 경향을 파악할 때 고려해야 하는 평가 구성 요소와 영향의 의미가 무엇이며, 채점자 교육의 원리를 확인하고자 한다. 채점 과정에 관한 분석 결과에서는 문항 유형에 따른 채점 과정의 특징과 채점 과정의 차이를 가져오는 채점자의 채점 경향을 살펴봄으로서 채점 과정 기반 채점자 교육의 내용을 구안하고자 한다.

1) 채점자 영향 분석의 시사점

(1) 문항 유형에 따른 영향

Ⅲ장에서 수행한 MFRM 분석을 통해 확인한 문항 유형에 따른 말하기 평가의 채점자 영향은 다음과 같았다. 첫째, '한국어 말하기 능력 시험'의 문항 유형 중에서 한국어 학습 경험을 묻는 1번 문항의 logit 값이 -1로 전체 문항 가운데 가장 쉬웠으며, 조언하는 말하기 능력을 평가하는 2번 문항의 logit은 0에 위치하여 보통 수준이었으며, 그래프와 텍스트 자료가 주어진 통합형 3, 4번 문항은 logit이 0과 1 사이에 위치하여 상대적으로 어려운 수준으로 나타났다. 도표 상에서 logit의 위치가 같은 행으로 나타난 3, 4번 문항의 logit 값을 비교하였을 때는 3번 문항이 더 어려운 곤란도 수준인 것으로 나타났다([그림 Ⅲ-15], <표 Ⅴ-1> 참조).

<표 Ⅴ-1> 문항별 MFRM 분석

문항	문항	사례수	MFRM		
			측정값	표준 오차	곤란도 순위
전체	1. 경험 말하기	624	-1.17	.06	4
	2. 조언하는 말하기	624	-.05	.05	3
	3. 도표 보고 설명하기	624	.64	.05	1
	4. 기사 읽고 문제와 해결 방안 이야기하기	624	.59	.05	2

이러한 결과와 관련하여 카이제곱(고정된) 값과 그에 따른 유의도를 바탕으로 문항별 곤란도의 차이가 유의한지를 확인하였을 때 유의 확률이 .00으로 유의한 차이가 있는 것으로 나타났다. 문항 유형별 곤란도의 차이는 문항을 구성하는 말하기 기능과 세부 과제, 고려해야 하는 언어 사용 맥락, 활용하는 자료

등의 영향 가운데 나타나게 되는데, 특정 문항의 과제 수행에서 고려해야하는 요소가 많고, 복잡하게 구성된 경우는 채점 수행을 위해 채점자가 파악해야 하는 문항에 관한 정보가 많다는 것이며, 각각의 비중을 가늠하기 어려울 수 있으므로 '채점의 곤란도'가 높아질 수 있다. 위의 <표 Ⅴ-1>에서 문항 간의 곤란도 차이가 .05로 매우 비슷한 수준으로 나타난 3번 문항과 4번 문항은 과제 수행을 위한 자료의 측면에서 보았을 때 4번 문항이 더 복잡하고 긴 글로 되어 있으므로, 4번 문항이 곤란도가 더 높을 것으로 예상할 수 있다. 그런데 실제 평가 결과에서는 3번 문항이 미세한 차이지만 더 어려운 수준인 것으로 나타났으며, 이는 문항 유형에 따른 곤란도 차이를 표면적인 정보만으로 규정하는 것을 경계해야 함을 보여준다.

말하기 평가의 채점자 교육에서는 채점 과정에서 문항별 특징을 어떻게 고려하고, 그러한 인식이 평가 결과에 어떻게 반영되는지를 채점자가 파악할 수 있도록 접근이 제시되어야 할 것이다.

(2) 평가 준거 특성에 따른 영향

실시한 말하기 평가의 결과에서 평가 준거별 차이를 비교하였을 때, 점수를 획득할 확률이 가장 낮은 것은 '전반적 수행 능력'이었으며, 가장 높은 것은 '발음 구사 능력'이었다($p<.05$). 이러한 결과는 채점자들이 전반적 수행 능력을 발음에 비하여 상대적으로 더 엄격하게 채점하였으며, 발음은 상대적으로 더 관대하게 채점하였다는 것을 의미한다(<표 Ⅴ-2> 참조).

<표 Ⅴ-2> 평가 준거별 MFRM 분석

문항	능력 특성	사례수	MFRM		
			측정값	표준 오차	곤란도 순위
전체	전반적 수행	156	.11	.05	1
	발음	156	-.14	.05	4
	어휘·문법	156	.01	.05	3
	담화	156	.03	.05	2

[그림 Ⅲ-15]에서 살펴본 전체 MFRM 분석 도표에서는 모든 평가 준거가 서로 근접하여 있고, 평가 준거별 차이에 대한 분리비(Strata)와 분리유의도 분석에서도 변별적인 특징이 나타나지 않았다(p>.05)는 것은, 평가 준거들이 서로 유사한 채점 경향을 나타내는 후광성의 영향으로 볼 수 있다. 평가 결과에서는 후광성으로 인하여 평가 준거들이 서로 매우 밀접한 관계를 맺고 있는 것으로 나타나게 되는데, 이러한 현상은 채점자가 수험자의 응답에 관해 갖고 있는 주관적인 인상이 평가 준거별 채점에 체계적으로 영향을 끼쳤기 때문으로 볼 수 있다. 말하기 평가의 채점 과정에서는 채점자가 수험자의 과제 응답에 대한 청취를 통해 채점의 근거를 마련하기 때문에 응답을 반복 청취할 수 없는 상황에서는 기억 체계에 의존하여 채점을 할 수밖에 없다. 이런 과정에서 채점자가 구체적으로 평가 준거와 관련된 수행의 증거를 포착하지 못한 경우에는 주관적인 인상을 중심으로 채점이 이루어질 수밖에 없을 것이다. 특히 분석적 채점을 실시하여 각 평가 준거별로 독립적인 평가 결과만을 얻기 원하는 경우에는 각 평가 준거를 분리해서 채점할 수 있도록 방안을 마련해야 할 것이다. 이와 관련하여 채점자 교육에서는 평가 준거별로 채점 과정에서 주목해야 하는 수행의 증거는 무엇이며, 채점 과정을 효율적으로 이끌어 나가기 위하여 필요한 접근 방법은 무엇인지에 관한 파악이 이루어져야 할 것으로 보인다.

(3) 채점 척도 활용도에 따른 영향

MFRM에 의한 채점 경향 분석에서 전반적으로 모든 채점자가 관대한 경향을 나타낸 가운데, R09(.00 logit)만이 채점자 영향이 없는 채점을 한 것으로 나타났으며, R13(-2.71 logit)과 R08(-2.14 logit)은 관대성이 지나치게 높은 편인 것으로 나타났다. 그런데 채점자의 엄격하거나 관대한 채점 경향이 무엇으로부터 기인한 것인지를 확인할 수 없는 상황에서 평가 결과에 대한 통계 분석만으로 채점자의 자격을 결정하는 것은 '채점 특성'의 복합적인 성격을 충분하게 반영하지 못한 것이라고 볼 수 있다. 채점의 특성의 산출은 채점자가 채점 과정에서 채점 척도를 어떻게 활용하였는지와 밀접한 관련이 있는데, 이는 Facets에서 산출하는 채점자별 채점 적합도를 통해 확인할 수 있었다. 채점의 적합도는 분석 결과에 나타난 평균제곱값(MnSq)과 표준화점수값(ZStd)을 기준

으로 파악할 수 있다. 이와 관련하여 채점 척도 활용 양상을 파악하기 위하여 리니커(Linacre, 2002: 878)에서 제시한 적합도의 분류 기준을 바탕으로 연구에 참여한 채점자들의 적합도를 재정리하였다(<표 Ⅴ-3> 참조). 적합도 기준에서 평균제곱값은 채점 척도 사용의 생산성, 즉 척도 범주를 충분하게 활용하였는지와 관련된 것이라면, 표준화값은 정규분포 범주에 포함될 것이라는 예측 가능성과 관련이 있다. 따라서 MFRM 분석 모형에서 가정하는 우수한 채점은 채점 척도의 범주를 충분하게 사용하면서, 또한 일정한 범위 안에서 예측 가능성이 있는 것이라고 볼 수 있다.

<표 Ⅴ-3> 채점자별 적합도 분류표

평균제곱 (MnSq) / 표준점수 (z)	> 2.0	1.5 - 2.0	0.5 - 1.5	< 0.5
≥3	-	R13	R02	-
2.0 - 2.9	-	-	-	-
-1.9 - 1.9	-	-	R08, R12, R10, R03, R11, R04, R09, R07	-
≤ -2	-	-	R01, R05, R06	-

<표 Ⅴ-3>에서 13명의 채점자 가운데 8명(R03, R04, R07, R08, R09, R10, R11, R12)은 1을 기준으로 하는 평균제곱값이 0.5 ~ 1.5 사이에 위치하며, 0을 기준으로 산출하는 표준점수에서도 -1.9 ~ 1.9에 위치하는 것으로 나타나 생산적이고 합리적인 채점을 한 채점자로 추정되었다. 그밖에 생산적인 채점은 실시한 것으로 보이나(평균제곱: 0.5 - 1.5), 합리적인 채점 패턴이 나타나지 않은 4명의 채점자(R01, R02, R05, R06)들도 있었다. 그리고 채점의 품질은 유지하였으나 비생산적이며, 예측이 어려운 채점을 실시한 것으로 나타난 채점자도 있었다(R13).

이러한 채점 척도의 활용도 문제와 관련하여 나타난 채점자 영향은 특정 점

수에 대한 집중 경향성과 후광성, 무작위성으로 나타난 바 있는데, 이와 같은 경향은 채점 척도나 과제 특성 고려와 관련된 채점 척도의 구성 요소를 잘 이해하지 못한 상태에서 임의적인 방식으로 채점을 하였을 때 나타날 수 있다 (Myford & Wolfe, 2004). 이와 관련하여 채점자 교육에서는 채점자가 채점 척도를 충분하게 활용할 수 있도록 채점 과정에서 채점 척도를 내재화하기 위한 연습과 점검이 이루어져야 할 것이다.

2) 채점 과정 분석의 시사점

(1) 평가 문항 유형에 따른 변화

본 연구의 Ⅳ장에서는 말하기 평가 채점 과정의 전반적인 특징을 파악하기 위하여 각 문항의 평가 결과에서 중위, 최상위, 최하위 수험자를 선정하고, 이들에 대한 전체 채점자의 채점 과정 양상을 살펴보았다.

채점 과정 분석 결과, 채점자들은 전반적으로 채점 척도에 제시된 순서를 따라 점수를 부여하는 순차형 채점을 가장 많이 하고 있는 것으로 나타났다. 이러한 경향은 본 연구에서 활용한 4개 문항 유형 중에서 '경험 말하기'의 중·하위 수준을 제외하고 나머지 모든 경우의 채점 과정에서 나타났다.

이와 관련하여 '경험 말하기' 문항의 특징을 확인하였을 때에, 먼저 이 문항은 평가 문항 중 곤란도가 가장 낮은 문항으로서 다른 문항에 비해 중·하위 수준의 수험자들이 받은 점수가 상대적으로 더 높게 나타났다. 이는 중·하위 수험자의 응답에 대한 채점 과정에서 긍정적으로 평가할 수 있는 특성이 나타났음을 의미한다. 채점자는 채점 과정에서 수험자의 응답에 나타난 특징 가운데 점수 획득에 긍정적인 특징과 부정적인 특징을 모두 인식할 수 있는데, 상대적으로 곤란도가 높고, 문항의 변별도가 뛰어난 문항의 경우에는 수험자 수준에 따라 응답 양상의 차이가 분명하게 나타날 수 있으므로 상위나 하위 사례에 대한 구별이 간명하게 이루어질 수 있을 것으로 예상된다. 그러나 변별도가 상대적으로 낮은 쉬운 문항에서는 모든 수험자의 응답 수준의 차이가 크지 않고, 발화 내용에 대한 파악이 상대적으로 잘 이루어졌기 때문에 채점 과정에서 분석적인 평가 결과를 종합하는 접근이 더 많이 이루어진 것으로 보인다.

한편, 전체 13명의 채점자들의 채점 과정을 살펴보면, 수험자의 수준이나 문항의 유형에 상관없이 항상 순차형 채점을 하는 채점자(R06)와 항상 종합형 채점만을 하는 경우(R11, R12)도 나타났으며, 중위 수준에서는 종합형 채점을 하고, 상위나 하위 수준에서는 일부 순차형 채점을 하는 경우(R04) 등의 수준별로 채점 과정에서 평가 준거에 대한 결정의 순서를 다르게 접근하는 채점자의 사례를 확인할 수 있었다. 여기서 순차형 채점을 한 채점자는 채점 과정에서 수험자 응답에 대한 채점자의 인상 기반 접근을 하고 있음이 나타났으며, 종합형 채점 유형에서는 구체적인 응답 정보에 대한 판단을 포함하는 경향이 나타났다.

이와 관련하여 채점자 교육에서는 문항별 채점 과정이 문항의 곤란도와 그에 따른 변별도의 영향으로 차이가 나타날 수 있다는 점을 채점자들에게 안내하고, 실제적으로 자신의 문항 수준에 따른 채점 과정 접근의 차이가 있는지를 확인해 보는 연습이 이루어질 수 있을 것이다.

(2) 평가 준거 인식에 따른 변화

말하기 평가의 채점 과정 분석에서는 채점자가 평가 준거를 어떻게 인식하고 있는지에 따라서 채점 과정의 차이가 나타났으며, 결과적으로 평가 결과도 달라질 수 있다는 것을 확인하였다. 먼저 독립형 문항에서 '전반적 수행 능력'을 먼저 결정하는 현상과 '담화 구성 능력'에 관한 점수 결정의 근거를 많이 언급하는 것은 엄격한 평가 결과가 나타나게 하는 원인으로 나타났다(R05, R08-1번 문항).

이러한 현상은 채점 과정에서 평가 준거를 확장 또는 축소하여 적용하는 가운데 나타날 수 있는데, '전반적 수행 능력'과 같이 종합적인 성격의 평가 준거의 경우, 점수 결정 과정에서 '천장 효과', 즉 상한선을 고정하는 효과를 나타내면서 다른 평가 준거의 수준을 제한할 수 있으므로 이를 주의해야 할 것이다. 읽기 자료를 활용하는 통합형 문항의 채점 과정 분석에서는 '담화 구성 능력'에 관한 근거를 많이 고려하여도 엄격한 경향이 나타나지는 않았지만, '발음 구사 능력'에 관한 채점의 근거를 보고하지 않은 사례에서 엄격한 채점이 이루어진 것을 확인할 수 있었다.

한편, ‘어휘·문법 사용 능력’과 관련하여서는 채점 과정에서 관련 근거를 많이 확보하였을 때 엄격한 채점이 이루어진 경우와 반대로 근거를 적게 고려하였을 때는 관대한 경향이 나타난 경우가 있었다(R11). 이러한 현상은 해당 평가 준거가 언어적인 정확성을 중심으로 하기 때문에 채점자가 계량적인 접근을 하였을 가능성이 높다. 따라서 채점 과정에서 많은 오류를 주목하였을 때는 엄격한 결과가 나타날 수 있으나, 반대로 청취 과정에서 주목한 정보가 적어서 근거가 부족한 경우에는 점수를 깎을 근거가 없기 때문에 인상을 중심으로 접근이 이루어지면서 오히려 후한 점수를 부여한 것으로 판단된다.

이와 같이 채점자들마다 점수 결정을 위해 평가 준거에 따라 고려하는 근거와 가정의 차이는 채점자가 청취한 정보와 수집한 정보에 대한 판단, 그리고 점수 결정을 위한 접근 방법의 차이로 인하여 발생할 수 있다. 이러한 차이가 채점 과정을 통해 부여하는 점수의 차이로 나타날 경우에는 신뢰도 분석이나 문항 양호도 분석 정보를 바탕으로 해석이 이루어질 수 있지만, 같은 점수를 부여한 것으로 나타났을 경우에는 평가 준거에 대한 같은 이해를 공유하고 있었는지, 그리고 어떤 응답 특성이나 가정을 주목하여 점수를 결정하였는지에 관한 정보를 바탕으로 접근이 이루어져야 할 것이다.

(3) 평가 척도의 통제된 활용으로 인한 변화

본고에서는 평가 척도를 제한적으로 사용하는 경향과 관련하여 채점자 R05가 3점을 집중적으로 부여하고, 5점은 부여하지 않는 것과 관련하여 R05가 연구 참여자 중에 긴 교육 경험 속에서 평가 척도를 통제적으로 사용하는 경향을 갖게 되었을 것으로 해석하였다. 이와 관련하여 R05는 채점 과정에서 채점 척도(AR)가 아니라 채점자의 경험(AE)을 가정하여 점수를 결정하는 경향이 나타났으며, 이는 채점자의 인식 속에서 ‘경험’의 비중이 그만큼 크고, 효과적인 것으로 판단하고 있음을 보여주는 것으로 판단된다. 채점 과정에서 교사가 자신의 경험을 근거로 고려하는 것은 언어 교육 현장에서 보편적인 현상일 것이다. 이러한 측면에서 보았을 때, 이 채점자는 채점의 정확성을 높이기 위하여 다른 채점자와의 점수 일치도를 높일 수 있는 방향으로 채점을 오랜 시간 지속해 오면서 이러한 특성이 나타난 것으로 해석할 수 있다.[51] 이 채점자가 최

댓값(5점)을 한 번도 선택하지 않은 점과 관련하여 채점 과정 보고를 살펴보면, 4점과 5점 사이에서 고민하는 내용의 보고가 이루어졌음을 확인할 수 있다. 이 채점자는 고민 상황에서 응답에 나타난 특정 문제를 재인하면서 5점을 부여하지 않고 있었다(#R050422).

#R050422
거의 이해하는 데는 많은 노력을 기울이지 않아도 됐기 때문에 4점나 5점까지도 줄 수 있다고 볼 수 있겠습니다.
근데 약간 좀 비음이 좀 강하고 그래도 머뭇거림 있었기 때문에
4점 주도록 하겠습니다

점수 결정을 고민하면서 이 채점자가 최종적으로 점수를 결정하기 직전에 응답에 나타난 문제를 재인하면서 부여한 점수(4점)는 응답의 수준이 정확하게 '4점'에 해당한다고 판단해서 부여한 것이 아니라 '5점'에 미치지 못하였다고 판단하여 부여한 점수로 보인다. 만약 5점과 4점 사이에 선택할 수 있는 '4+ 또는 5-'와 같은 보다 세부적인 척도가 존재한다면 이 채점자의 평가 척도 활용 양상의 문제가 정확히 어떤 지점에서 비롯된 것인지를 파악할 수 있을 것이다. 이러한 채점자의 주관적인 채점 방법 사용에 따른 한계를 극복하기 위해서는 해당 평가에서 목표로 하는 언어 수준과 범주에 관한 인식을 마련할 수 있도록 해야 할 필요가 있으며, 이는 채점 과정에서 나타날 수 있는 제한적인 채점 척도 사용 현상을 해소하는데 기여할 수 있을 것이다.

(4) 상호작용 요소에 따른 변화

말하기 평가에서 채점은 개인별로 독립적으로 이루어지는 것이 일반적이지만, 채점 과정에서 수험자, 문항, 채점 척도 등 여러 평가 구성 요소들과 상호작용하는 가운데 점수 결정에 이르게 된다. 말하기 평가의 채점 과정은 이러한 상호작용 요소들을 언제 그리고 어떻게 고려하느냐에 따라서 그 양상이 달라

51) 신경경제학적인 관점에서는 이러한 현상을 바라보았을 때는 유사한 상황에 반복적인 노출이 이루어지면서 채점자가 5점을 부여하였을 때 발생할 수 있는 다른 수험자의 불균형 또는 채점자 간 신뢰도의 떨어짐과 같은 손해를 줄이기 위하여 한 선택으로도 볼 수 있을 것이다.

질 수 있을 것이다. 이와 관련하여 채점 과정 분석 결과에서는 독립형 문항과 통합형 문항 사이에 평가 준거에 따라 점수 결정에 관한 영향의 차이를 확인한 바 있다(예: '전반적 과제 수행 능력', '발음 구사 능력'). 또한 채점 과정을 통하여 점수를 부여하기 위하여 고려하는 평가적인 가정(AR, AT, AE, AO)의 측면에서도 채점자가 응답에 나타난 어떤 특징과 관련하여 어떤 가정을 어느 정도 혹은 빈도로 고려하는지에 따라 채점 과정의 양상에 차이가 있음을 확인하였다. 채점 과정에서 고려할 수 있는 여러 가정 중에 채점자의 전문성과 관련이 높은 경험에 기반을 둔 가정(AE)은 점수 결정에 있어서 편향적인 영향을 나타낼 수 있다는 점에서 주의가 필요할 것이다(R12의 사례). 점수 결정을 위한 가정은 그 자체가 선택하려는 점수의 의미를 포함한다고 볼 수 있는데, 만일 그 가정의 내용이 구체화가 필요한 모호한 상태의 정보라면 평가 결과 도출에 불확실성을 가중시킬 가능성이 있다. 채점 과정에서 채점자의 경험을 가정으로 고려할 경우에는 인상적인 판단에 그치지 않도록 구체적인 정보를 포함하고 있는 정보를 활용하는 것이 점수 결정의 타당성을 확보할 수 있는 방법이다.

이와 같은 현상은 단순히 평가 준거에 대한 채점자 이해의 문제가 아니라 해당 평가 문항의 구성적인 특징과 평가 준거가 채점자와 함께 상호작용 하면서 나타난 것이라는 점을 주목할 필요가 있다. 이러한 채점 과정에서 고려하는 평가 요소로 인한 상호작용적인 영향은 각각의 요소가 내포하고 있는 불확실성으로 인하여 예상하기가 어렵기 때문에 예비 시행을 통해 이를 점검할 필요가 있다. 이와 관련하여 채점자 교육에서는 실시하는 평가 맥락에서 고려할 필요가 있는 평가 요소들을 중심으로 채점 과정에서의 상호작용적인 영향을 알아보기 위한 연습 채점의 수행과 그에 대한 분석이 이루어져야 할 것이다.

2. 채점 과정 기반 한국어 말하기 평가 채점자 훈련의 원리

본 연구에서 제안하는 채점 과정 기반 말하기 평가 채점자 교육은 채점자가 자신의 채점 경향을 성찰할 수 있도록 채점 과정에 대한 분석을 통하여 자신의 채점에 관한 지식이나 신념, 관점 등을 돌아보고, 필요한 부분을 보완할 수

있도록 하는 것이 핵심이다. 기존의 언어 평가 분야에서 대표적인 채점자 교육 방법은 대규모 평가의 관습을 따라 일치도 확보를 위한 훈련(training)으로서 기준(benchmark)을 따라 자신의 채점 경향을 임의적으로 조정(norming)하는 방식으로 이루어져 왔다. 이는 효율성과 경제성을 고려해야 하는 대규모 평가에서 선택할 수 있는 방식이지만, 채점자가 갖고 있는 채점에 관한 전문성을 제한하고, 결과적으로 채점 과정을 형식화 및 기계화 시킨다는 점에서 결과 해석의 타당성을 훼손할 수 있다는 문제가 있다고 판단된다. 이러한 일치도 확보를 목표로 하는 채점자 훈련은 결과적으로 채점자가 이전에 형성한 '내재된 인식'(inbuilt perception)인 주관적 경향을 제거하지 못하며(Brown, 1995; Lumley & McNamara, 1995; Myford & Wolfe, 2000; Weigle, 1998), 훈련 이후에도 채점자별로 상이한 평가 준거 친숙성의 영향(Xi & Mollaun, 2009)이 나타난다는 점은 기계적인 채점자의 양성을 목표로 하는 채점자 훈련의 한계를 극복하기 위하여 새로운 접근의 필요성을 방증하는 것으로 볼 수 있다. 이와 관련하여 본 연구에서는 채점 과정 분석을 중심으로 하는 채점자 교육이 평가 결과의 일치도 향상만이 아니라 채점 과정에 대한 채점자의 인지적 활성화를 가져올 수 있다는 점(Davis, 2016)에서 이를 주목하였다.

1) 채점 척도의 내재화

채점자는 말하기 평가의 채점 과정에서 다양한 형태의 채점 지식을 동원하여 인지적인 처리를 통하여 채점 정보를 인식 및 규정하고, 관계를 형성하고, 판단을 내리고, 최종적으로 점수를 결정한다. 평가 맥락에서 채점 지식의 표상에 직접적으로 관여하는 것은 채점 척도라고 할 수 있다. 채점 척도는 점수를 결정해야 하는 측정 영역인 평가 준거와 점수의 간격인 척도, 그리고 각 준거별 척도에 해당하는 수준에 관한 기술인 채점 기준으로 이루어져있다. 채점 척도 상의 여러 정보에 대한 이해는 채점자의 배경 지식에 영향을 받으며, 채점 기준을 적용하는 범주를 축소하거나 채점의 속도를 변화시키는 등에 관여할 수 있다(Bejar, 2012: 5).

채점자는 채점 과정에서 자신이 갖고 있는 채점 지식을 바탕으로 채점 척도를 재해석하고, 자신의 수험자 응답에 관한 표상과 비교하면서 채점을 수행한

다. 채점자들이 공동의 채점 지식 표상을 갖기 위해서는 채점자들이 무엇을 목적으로 채점을 하며, 그 목적을 성취하기 위해 적합한 방법은 무엇인지, 그리고 해당 평가 맥락에서 고려해야 할 것은 무엇인지에 관한 체계적인 이해가 마련되어야 한다. 이는 채점 규칙과 채점 척도만을 제공하는 형식적인 접근이 아니라 채점자가 스스로 채점 수행의 이유를 깨닫고, 능동적으로 채점 지식을 활용하여 채점 관련 정보들을 비교하며, 자신의 채점 과정을 개선하려는 노력 가운데 이루어질 수 있을 것이다.

말하기 평가의 채점에서 이루어지는 채점 지식의 표상은 수험자 발화 청취와 채점 척도에 대한 독해를 통해 이루어지는 선언적 지식 표상과, 그리고 채점 과정에서 이루어지는 상호작용의 영향에 관한 절차적 지식 표상이 있다. 채점자는 자신의 채점 전문성이 다른 사람과 같은 점수를 선택하는 능력이 아니라 채점 과정에서 고려해야 할 여러 평가 맥락 요소와 개인적 요소에 대한 바른 이해를 갖추는 데 있다는 것을 알고 채점 과정에서 일어나는 인지적인 측면에 대한 상호작용의 영향이 어떤 절차에서 어떻게 나타나는지에 관한 개념을 갖도록 해야 한다.

2) 채점 수행 증거의 타당성 확보

말하기 평가의 채점자는 채점 과정에서 고려하는 여러 증거들이 무엇이며, 다른 증거들과 어떻게 구별할 수 있는가에 관한 인지적인 능력을 갖추고 있어야 한다. 증거는 채점자가 채점 과정을 통해 결정한 점수의 타당성을 입증하는 것인데, 증거의 가치는 증거를 발견하는 과정의 가치와 관련이 있다(Cronbach, 1971, 483). 채점 과정에서 채점자가 사전에 갖고 있던 신념 체계에 의해 주목한 정보는 다른 정보들보다 무비판적으로 지각되고 저장하였을 가능성이 있다. 이처럼 익숙한 정보를 주목하거나, 해당 정보를 익숙한 방식으로 처리하려는 경향은 채점 과정에서 자연스럽게 일어날 수 있는 일인데, 이는 채점 과정 분석에서 나타난 바와 같이 채점 척도를 제한적으로 사용하거나 평가 준거를 구별하지 않고 적용하는 경향의 원인이 될 수 있다.

말하기 평가 채점 과정에서 채점자가 고려할 수 있는 증거는 인지적인 한계의 영향을 받는데, 이러한 한계를 극복하기 위해서 수험자의 응답을 의미 단위

로 연결시켜 기억하고, 장기 기억과 연계할 수 있는 가정을 고려함으로서 효과적으로 증거를 처리하는 방법을 적용할 수 있다. 채점 과정 분석에서 확인한 바와 같이 채점자가 점수 결정을 위하여 어떤 가정을 고려할 것인지에 따라 평가 결과의 변화가 나타났다. 채점 과정에서 고려하는 증거에 대한 판별을 잘 하기 위해서는 채점하는 준거에 따라 계량적인 접근이 가능한지, 인상적인 접근이 필요한지와 같이 각 평가 준거가 가진 특징을 고려하여 수행 증거에 대한 판별이 이루어져야 한다.

3) 채점 경향에 관한 일관성 유지

채점 과정을 중심으로 이루어지는 말하기 평가의 채점자 교육은 지속적인 채점 경향에 대한 성찰을 중심으로 채점자 경향의 일관성 유지와 개발을 목표로 한다. 이는 말하기 평가가 특정 이익 집단의 편의를 위한 것이 아니라 평가 참여자와 이해 관계자가 함께 만족할 수 있는 교육적 활동으로 기능하기 위하여 지향해야 하는 중요한 목표라고 할 수 있다. 공식적으로 이루어지는 언어 평가의 특성상 시험을 계획하고 개발한 평가 개발자 혹은 교사는 평가 상황에서 평가 결과를 얻기 위해 응시한 수험자(학습자)에 비해 상대적인 우월성을 갖고 있다. 또한 이들이 정한 채점 방법 또는 채점 과정을 통해 산출한 평가 결과는 '평가'의 결과라는 점에서 수험자가 신뢰할 수 있는 정보로 인식될 가능성이 크다. 따라서 채점자는 자신에게 평가 결과에 대한 책임이 있음을 기억하고, 수험자와 그 밖의 평가 결과 사용자가 이해할 수 있는 채점의 근거를 제시할 수 있어야 한다는 사실을 유념해야 할 것이다.

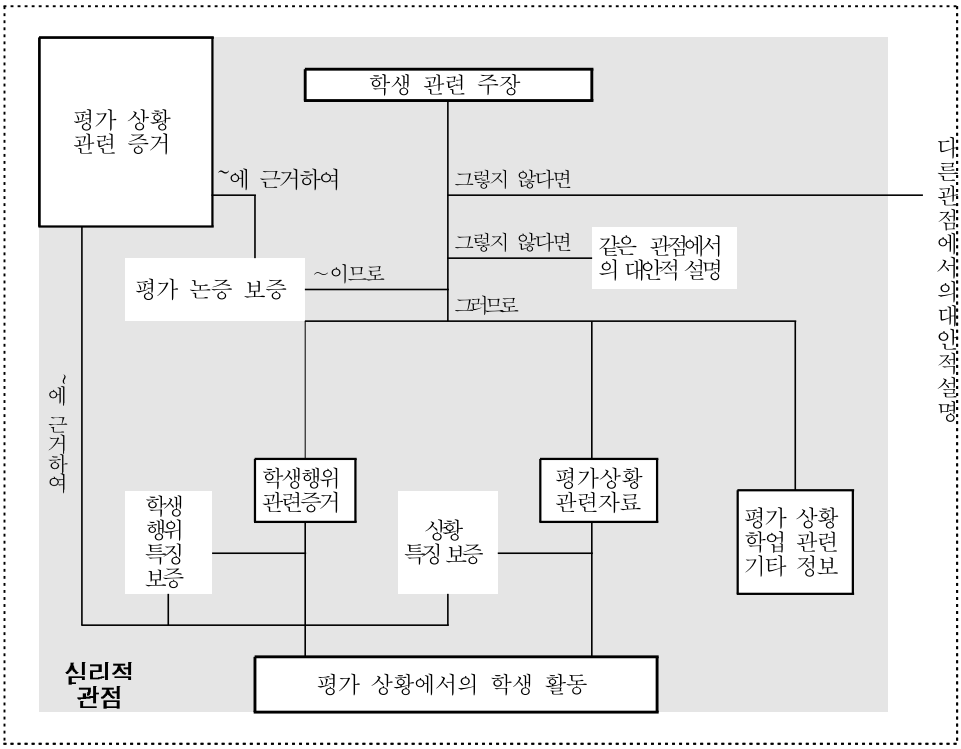
말하기 평가의 채점 과정에서 이루어지는 채점 경향에 대한 성찰은 정보에 대한 판단을 내린 이후에 점수를 결정하기까지의 과정에서 이루어진다. 채점자는 자신의 기억 체계에 저장한 수험자 응답에 관한 정보와 채점 척도 정보를 연계하여 점수를 결정하며, 주관적인 채점 경향성을 파악하기 위하여 채점 과정 분석을 통하여 자신의 채점 과정을 평가할 필요가 있다. 이러한 접근을 통해 평가의 일관성을 떨어뜨리는 불확실하거나 제한적인 채점 경향을 극복할 수 있도록 해야 할 것이다.

3. 채점 과정 기반 말하기 평가 채점자 교육의 설계

채점 과정 기반 말하기 평가의 채점자 교육은 채점자가 자신의 채점 특성을 파악하기 위하여 정보 수집 과정과 정보 판단 과정, 점수 결정 과정으로 이루어지는 채점 과정을 성찰하면서 자신의 채점 경향과 채점 과정에서 고려해야 하는 평가 구성 요소들을 적절하고 충분한 수준으로 다룰 수 있는지에 관한 평가를 중심으로 구성하였다.

1) 채점자 교육의 근거와 절차

채점 과정에 개입하는 핵심적인 영향 요인은 중 하나로 주목받아 왔던 채점 척도에 대한 채점자의 주관적인 이해(Fulcher, 1993; Fulcher, 2003)는 채점자 영향에 관한 선행 연구 결과에서 채점자 훈련으로 제거할 수 없는 성질의 것이며, 잠시 사라진 것처럼 보인다고 하여도 지속력이 없다는 사실이 밝혀져 왔다. 그런데 최근에 수행 기반의 언어 평가에 대한 인지심리학적인 관점을 적용하려는 접근이 이루어지면서 평가를 구성하는 요소들이 고정된 것이 아니라 실제적으로 다양한 요소에 의하여 지속적으로 상호작용이 일어나는 것으로 보는 관점이 나타나기 시작하였다(Bejar, 2012).



[그림 Ⅴ-1] 평가 논증의 구조(Mislevy, 2006: 465)

[그림 Ⅴ-1]은 언어 평가에서 결정한 점수를 바탕으로 수험자에 관한 특정한 주장을 이끌어내기까지의 논리적인 과정을 나타낸 것이다. 이 과정에서 고려하는 여러 가지 증거들은 평가 결과의 타당성을 지지하는 일련의 구조를 형성하고 있는데, 이러한 '평가 논증'은 평가 상황에서 새로운 해석의 가능성을 확보하기 위하여 다양한 반론과 새로운 증거에 관한 탐색 가능성을 반영하기 위하여 평가 상황적인 맥락과 외부적 관점을 고려한다는 특징을 갖고 있다(Mislevy, 2006: 466). 언어 평가에서 이러한 접근이 가능하기 위해서는 평가에 참여하는 채점자들 간에 고려해야 하는 평가 맥락과 상황에 대한 정보와 이해를 공유해야 하며, 채점자 개인으로서는 이와 같은 논리적인 절차를 따라 학습자(수험자)에 관한 주장을 할 수 있도록 타당한 근거를 마련할 수 있어야 할 것이다.

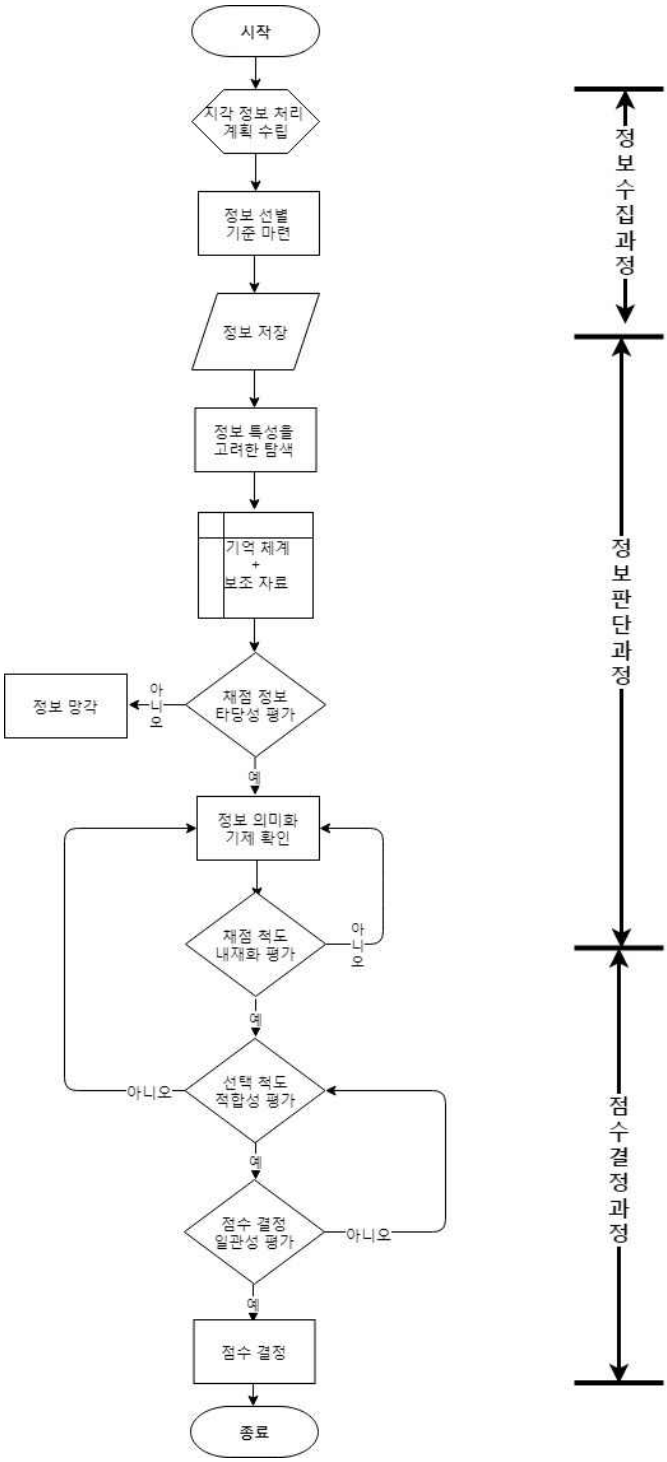
본고에서 제안하는 채점 과정 기반 말하기 평가 채점자 교육에서 채점자는

자신이 청취한 수험자의 발화에 나타난 특징과 채점 척도를 연계하여 해당하는 말하기 능력 수준에 관한 주장을 하며, 이 과정에서 점수 결정을 위한 자료들의 타당성을 확보하고 채점에서 불확실한 결과를 도출할 가능성을 줄이기 위하여 자신의 채점 경향을 성찰할 필요가 있다.

이와 관련하여 [그림 Ⅴ-2]는 채점 과정 기반 채점자 교육이 이루어지는 절차를 채점 과정 모형을 기초로 구성한 것이다. 채점 과정 기반 채점자 교육은 크게 수험자의 응답으로부터 채점을 위한 정보를 수집하는 과정과 수집한 정보의 내용과 성격을 규정하는 정보 판단 과정, 그리고 점수를 결정하는 과정으로 구성하였다. 본 채점자 교육의 특징은 채점자가 자신의 채점 경향을 파악하기 위하여 채점 과정 보고를 수행하면서 진행된다는 점이다. 녹음한 채점 과정 보고 자료는 채점 정보 타당성 평가와 채점 척도 내재화 평가, 선택 척도 적합성 평가, 점수 결정 일관성 평가와 관련하여 인지적인 특성을 성찰하기 위한 자료로 활용하기 위한 것이다.

수험자의 응답을 청취하면서 이루어지는 정보 수집 과정에서는 수험자의 응답으로부터 지각한 정보 가운데 특정 정보를 어떻게 처리할 것인지에 대한 계획을 먼저 마련하고, 그 중에서 채점 과정에서 중요하게 다루어야 하는 정보를 선별하기 위한 기준을 마련하는 활동이 핵심이다

다음으로 정보 판단 과정에서는 단기 기억 체계에 저장된 정보에 대한 특성에 따라 주관적인 인상 중심의 접근이 이루어졌는지 아니면 계량적인 접근을 취하였는지 등을 기준으로 탐색이 이루어진 다음에 장기 기억과 이를 보조할 수 있는 메모 등의 보조 자료의 정보를 바탕으로 해당 정보를 구체화 한다. 구체화한 정보는 그에 대한 타당성을 판단하기 위하여 평가되며, 타당성이 없는 정보는 제거하고, 타당성이 있는 것으로 판단한 정보는 해석을 위하여 고려할 수 있는 관련 맥락을 탐색하도록 한다.



[그림 Ⅴ-2] 채점 과정 기반 말하기 평가 채점자 교육의 진행도

마지막 과정은 점수 결정을 위하여 기준을 제시하고 있는 채점 척도의 내재화 수준을 평가하는 것으로부터 시작한다. 내재화하지 못한 채점 척도의 정보는 수험자 응답에 대한 해석의 맥락으로서 기능할 수 있도록 재검토가 이루어질 수 있다. 다음으로 평가 맥락을 바탕으로 의미화환 수험자 응답의 특징과 내용을 채점 척도에 적용하고자 하였을 때 적합한 것인지를 판단해야 하는데, 이 과정에서는 수험자 응답의 특징과 채점 척도의 연계가 적합하지 않다고 판단하였을 때에는 각각의 해석에 문제가 없는지를 다시 살펴보도록 해야 한다. 마지막 과정에서 채점자는 자신의 점수 결정이 일관성을 갖고 있는지를 평가해야 하는데, 이는 다른 채점자와의 비교 또는 내적 신뢰도 검정으로 이루어질 수 있다.

2) 채점자 교육의 목표와 내용

말하기 평가의 채점 과정을 중심으로 이루어지는 채점자 교육은 채점자가 자신의 채점 과정 보고를 산출하는 경험을 중심으로 <표 62>에서 제시한 4개의 교육 목표를 달성하기 위하여 이루어져야 한다고 보았다.

<표 Ⅴ-4> 채점 과정 기반 말하기 평가 채점자 교육의 목표와 내용

교육 목표	교육 내용
말하기 평가 채점 과정의 구성적 특징을 이해한다.	추론을 통한 채점 과정 추단을 통한 채점 과정 채점 과정의 상호작용 요인
채점 과정을 통해 나타난 채점자 영향의 유형과 산출 방법을 안다	집단적인 채점자 영향 개인적인 채점자 영향 관찰점수에 의한 채점자 영향 확인 조정점수에 의한 채점자 영향 확인
채점 과정을 보고하는 방법과 보고한 자료를 바탕으로 채점 과정 양상을 분석한다.	채점 과정 보고 방법과 특징 채점 과정 보고의 분석 절차 채점 과정 보고의 형식화 채점 과정 보고의 코딩
말하기 평가 채점 과정의 의미를 확장한다.	채점자 집단별 협의 전체 집단 종합 협의

<표 Ⅴ-4>에서 제시한 첫 번째 교육 목표는 채점자 교육을 통하여 말하기

평가 채점 과정의 구성에 대한 이해를 갖도록 하는 것이다. 이와 관련하여 교육 내용으로는 채점 과정에 관한 추론과 추단의 관점, 그리고 채점 과정에 개입할 수 있는 상호작용 요소에 관한 검토가 이루어질 수 있다. 두 번째 교육 목표는 평가 결과에 체계적으로 영향을 미치는 채점자 영향을 파악하는 것이다. 이와 관련하여 채점자 교육에서는 집단적인 수준에서 나타나는 채점자 영향과 개인 수준에서 나타나는 채점자 영향은 어떤 것이 있는지 알아본다. 그리고 이를 실제적으로 파악하기 위하여 관찰 점수를 바탕으로 접근하는 방법과 척도 점수를 바탕으로 접근하는 방법을 알아보고, 수치에 관한 해석 방법에 관한 내용도 포함하였다. 세 번째 교육 목표는 채점 과정 보고의 수집과 분석의 방법을 파악하는 것이다. 이와 관련하여 채점자 교육에서는 채점 과정 보고의 방법과 특징 이해, 채점 과정 보고를 분석하는 절차, 채점 과정 보고를 형식화하는 방법, 채점 과정 보고의 코딩 방법 등의 내용을 다룰 수 있을 것이다. 마지막 교육 목표는 채점 과정에 관한 의미를 개인별 성찰과 집단적인 협의를 바탕으로 확장하는 것이다. 이는 채점자들이 갖고 있는 채점 경향을 상호작용을 통해 보완하기 위한 것이다.

VI. 결론

1. 연구의 요약

본 연구에서는 한국어 말하기 평가의 채점에 관한 실증적인 접근을 위하여 먼저 문헌 연구를 바탕으로 채점 과정의 모형을 구성하였다.

말하기 평가의 채점 과정은 인지적인 정보 처리 과정으로 이루어진다는 점에서 채점자의 추론과 추단으로 이루어진다고 판단된다. 이와 관련하여 이론적 검토를 바탕으로 말하기 평가의 채점 과정은 채점자의 채점 과정은 평가에 관한 외적·내적 요인의 영향 속에서 수험자의 응답을 청취하고 그로부터 수용한 정보를 판단하며, 점수를 결정하는 체계로 구축하였다.

다음으로 채점 과정에 관한 실제적인 특징과 양상을 확인하기 위하여 중·고급 한국어 학습자를 대상으로 구성형 응답 문항으로 이루어진 준직접식 평가인 '한국어 말하기 능력 시험'을 구안하였다. 평가의 문항은 '국제통용한국어표준교육과정'과 한국어 교재에 대한 내용 분석을 기초로 작성하였다. 채점 척도는 국내외 말하기 평가 연구에 대한 검토를 바탕으로 4개 평가 준거에 따라 0~5점의 척도별 채점 기준으로 구성하였다. 채점 과정의 심층적인 양상을 파악하기 위하여 채점의 방법은 총체적 채점 구인과 분석적 채점 구인을 함께 채점하는 혼합식 채점으로 이루어졌다. 말하기 평가에는 18명의 수험자가 참여하였으며, 연습 채점 대상을 제외하고 12명의 응답에 대하여 채점을 실시하였다. 연구에 참여한 채점자는 경력 5년 이상의 한국어 교사 13인이었다. 채점 과정에 관한 실증적인 접근을 위해 채점자들이 부여한 말하기 평가 점수와 녹음한 채점 과정 보고를 수집하였다. 수집한 자료 중에서 평가 결과에 나타난 채점자 영향 분석은 고전검사이론과 문항반응이론의 분석으로 이루어졌다. 분석 결과, 높은 내적 일관성 신뢰도가 나타났으나, 채점자 간 신뢰도를 확인하였을 때는 상대적으로 낮은 상관을 나타내는 채점자가 나타났다. 채점자 배경 변인 분석에서 채점자의 교육 경력, 평가 경험, 평가 관련 교육 경험, 구사 가능한 외국어 수 등의 배경 변인에 대하여 유의한 차이가 나타나지 않았다.

MFRM에 따른 문항반응이론 분석에서는 등급 척도 모형을 적용한 분석 결과에서 채점자 사이에 유의한 채점 경향성의 차이가 나타났으며, 전반적으로 관대한 채점 경향이 나타났다. 평가 문항 유형 중에 수험자 능력 수준에 대하여 어려운 수준으로 나타난 '도표 보고 설명하기' 문항과 '기사를 읽고 문제와 해결 방안 이야기하기' 문항이 같은 곤란도 수준으로 logit 척도 상에 나타났으며, '조언하는 말하기' 문항은 보통 수준, '경험 말하기' 문항은 쉬운 수준으로 나타났다. 혼합형 채점 척도 사용에서는 총체적 채점 영역인 '전반적 수행 능력'이 어려운 수준으로 나타났으며, '발음 구사 능력'은 상대적으로 쉬운 수준으로 나타났다.

채점자별 채점자 영향 분석에서는 평가 결과에 나타난 채점 경향성, 집중 경향성, 채점 척도 및 준거 사용 경향성을 확인하였다. 채점 경향성 분석 결과, 평가 문항과 채점 영역에 따라 채점의 엄격하거나 관대한 경향이 MFRM 모형의 기대 수준을 벗어난 경우가 있음을 확인하였다. 특정 점수에 대한 집중 경향성의 측면에서도 한 채점자가 중앙값인 3점을 상대적으로 많이 사용하는 경향을 확인할 수 있었으며, 평가 준거별로는 발음(46%)을 중심으로 3점을 부여하고 있음을 확인하였다. 채점 척도의 주관적 사용 경향인 무작위성·후광성과 관련하여 전체 집단 수준 분석에서는 확인할 수 없었던 특징을 채점자 개인 수준 분석을 통하여 확인할 수 있었다. 무작위성이 나타난 채점자는 편향 상호작용 분석 결과에서 MFRM 기댓값과 가장 차이가 큰 채점자의 채점 사례를 통하여 확인할 수 있었으며, 후광성은 채점자의 척도 사용의 한계치 분석에서 척도 거리가 가장 넓은 경우와 내적합도가 높은 채점자의 점수열 비교에서 확인할 수 있었다.

다음으로 채점자들의 채점 과정 보고 자료를 바탕으로 채점 과정을 분석하였다. 채점 과정 분석은 구성형 응답 문항 유형 및 MFRM 분석을 통하여 확인한 채점자 영향을 바탕으로 이루어졌다. 구성형 응답 문항 유형별 채점 과정 분석에서 채점자들은 '경험 말하기' 문항의 중위 수준과 하위 수준을 제외하고 대부분 순차적 채점을 한 것으로 나타났다. 또한 채점자들에게서 점수 결정의 근거와 가정에 대한 접근의 차이가 나타났는데. 이는 채점자의 선별적 지각 능력의 차이와 채점 자원에 대한 인식 차이에서 비롯된 것으로 판단된다. 문항 유형 중에서는 자료를 바탕으로 응답을 하는 통합형 말하기 문항에서 점수 결

정의 근거로 담화적인 응답의 담화적 특징을 고려하는 빈도가 높고, 그러한 결과가 다른 평가 준거의 채점에도 영향을 끼치는 것을 확인하였다.

채점자 영향에 따른 채점 과정 분석에서 채점 경향이 엄격한 것으로 나타난 채점자의 채점 과정에서는 총체적 채점 기준에 대한 점수를 먼저 부여하고, 담화 구성 능력을 중복하여 고려하면서 엄격한 채점 결과가 나타났다. 이러한 경향은 문항의 유형에 따라 차이가 있었는데, 자료를 활용하는 기능 통합형 문항에서는 채점 과정에서 담화 구성 능력에 대한 근거를 고려하는 것이 엄격한 채점 결과로 이어지지 않았으며, 발음에 관한 근거를 주목하지 않는 경향이 나타났다. 관대한 채점 경향이 나타난 채점 사례의 채점 과정 분석에서는 채점자들이 평가 준거에 대해 인상적인 접근을 하였다는 점과 점수 결정의 근거로 가정을 고려하지 않는 경향이 나타났으며, 말하기 기능을 독립적으로 평가하는 문항에서는 채점 과정에서 발음 구사 능력을 많이 주목한 경우에 상대적으로 관대한 채점 경향이 나타난다는 것을 확인할 수 있었다.

집중 경향성이 나타난 채점자의 채점 과정 분석에서는 중앙값인 3점을 집중적으로 부여한 채점자가 채점 과정에서 채점자 경험을 가정하는 경향이 나타났으며, 계량적인 접근이 가능한 어휘·문법적인 측면을 상대적으로 주목하지 않는다는 것을 확인할 수 있었다. 채점 척도 사용에서 최고 수준인 5점에 집중하는 경향이 나타난 채점자는 채점 과정에서 채점 척도, 경험, 신념 등의 점수 결정을 위한 가정을 고려하지 않는 것으로 나타났으며, 담화 구성 능력과 전반적 수행 능력을 같은 준거로 인식하여 점수를 결정하는 특징이 나타났다. 또한 보고 내용에서 수험자 발화에 대한 구체적인 내용보다는 직관에 의한 주관적 판단을 근거로 삼고 있었다.

채점 척도의 평가 준거나 평가 척도에 대한 주관적인 인식과 관련하여 채점 과정 보고 내용을 분석한 결과, 계량적인 접근이 가능한 준거를 충분하게 고려하지 않았으며, 모호한 수준의 가정을 고려하면서 임의적인 점수 결정이 이루어졌다는 것을 확인할 수 있었다.

채점 과정에 대한 분석을 통해 말하기 평가의 채점 과정과 관련하여 얻을 수 있는 시사는 다음과 같다. 첫째, 말하기 평가 채점 과정에서 독립형 말하기 과제의 총체적 평가 준거는 분석적 평가 준거에 대한 채점에 영향을 미쳐 측정값을 통제하는 효과를 나타낼 수 있으므로, 이를 고려하여 결과를 해석할 필

요가 있다고 판단된다.

둘째, 말하기 평가의 채점에서 특정 문항에서 관대한 채점이 이루어진 경우에는 수험자 응답에 대한 증거 수집이 잘 이루어지지 않는 경향이 나타났다. 말하기 평가에서 채점자가 겪을 수 있는 가장 우선적인 어려움은 발화 청취에 관한 것이며, 청취 과정에서 주목한 정보가 적거나 없을 경우에는 문제 발생을 회피하기 위하여 좋은 점수를 부여하거나 무작위적인 채점이 이루어질 가능성이 있다고 판단된다.

셋째, 채점 척도를 제한적으로 사용하여 특정 점수의 선택이 집중적으로 나타나는 경향과 척도를 제한적으로 사용하는 경향은 채점자가 주목한 정보를 처리하는 특정한 방식이 있으며, 그에 따라 새로운 정보는 기존의 체계에 통합하여 처리하고 있음(Rust et al., 2005: 232; Anson, 2006: 104)을 나타낸 것으로 해석하였다. 이러한 경향은 카너먼과 프레더릭(2002)이 제시한 추단의 첫 번째 체계인 직관을 따른 것으로 볼 수 있는데, 채점자가 평가 경험을 통하여 형성한 직관이 청취를 통해 얻은 정보에 대한 인식을 제한하고 있는 것으로 판단된다. 채점 척도의 주관적 인식 경향은 채점 방법이나 점수 결정에 대해 단순한 접근을 취하는 경우에 나타났다. 이는 채점자가 채점 과정에서 여러 요소를 고려할 때 발생하는 인지적 부담을 떨어드리려는 경향으로 판단된다. 말하기 평가에서 채점자는 평가 개발자가 제시한 채점 척도를 사용하지만, 실제로는 자신이 심리적으로 재구성한 기준을 따라 평가한다고 볼 수 있으며, 무작위적인 채점 경향의 경우에는 그 기준 체계가 임의적인 성격을 띤다면, 평가 준거를 구분하지 않는 후광성 경향의 경우는 경계를 설정하지 않으면서 보상적·연계적 접근을 취하였기 때문에 나타난 것으로 해석하였다.

끝으로 본 연구에서는 이상의 연구 수행 내용을 바탕으로 말하기 평가의 채점 과정 기반 접근이 채점의 타당성과 책무성을 확보할 수 있는 방법이라고 보고 채점자 교육에 대한 적용이 필요하다고 보았다. 이에 채점 척도에 대한 내재화의 원리와 수행 정보에 대한 타당성 확보의 원리, 채점 경향의 일관성 유지의 원리를 바탕으로 이루어지는 채점 과정 기반 채점자 교육의 방안을 제시하였다.

2. 후속 연구 제언

본고에서 수행한 말하기 평가에서의 채점자 영향과 채점 과정 연구를 바탕으로 후속 연구에서 다루어야 할 연구의 내용과 방향을 양적 분석 방법, 질적 자료 수집, 수험자의 응답 과정에 관한 통합방법연구, 채점 중심의 '평가 전문성' 연구, 실험 연구를 통한 말하기 평가 채점 과정 모형의 규명으로 제안하였다.

첫째, 본 연구에서 수행한 양적 자료 분석 방법인 MFRM은 1모수 문항반응이론을 바탕으로 문항 곤란도만을 고려하고, 모든 문항의 변별도를 1로 고정한다는 특징을 갖고 있다. 또한 MFRM을 적용한 소프트웨어인 Facets는 분석 모형에 대한 선택에 따라서 측정값의 미세한 변화가 나타난다. Facets에서 다분 문항 분석의 기본 모형인 등급 척도 모형을 적용할 경우가 부분 점수 모형을 적용할 경우보다 미세하게 수치가 낮게 나타날 수 있다. 이는 각 분석 모형에서 곤란도에 대한 정의가 다르기 때문인데, 등급 반응 모형에서 곤란도(F_{ik})를 모든 문항에서 모든 판정자가 평균적으로 특정 범주의 점수를 얻을 확률로 정의한다면, 부분점수 모형에서는 각 문항에서 판정자가 평균적으로 특정 범주의 점수를 얻을 확률로 고려하기 때문이다. 따라서 부분점수모형에서는 각 점수 범주별 확률 가운데 중복하여 계산되는 경우가 있었기 때문에 미세한 차이가 나타난 것으로 보인다. 이러한 차이는 분석 결과의 해석을 뒤집거나, 바꾸는 결과들은 아니었으나, 능력 추정에 대한 정확도를 떨어뜨릴 수 있기 때문에, 반드시 어떤 목적으로 어떤 모형을 선택하여 분석을 실시하였는지를 분석 과정에서 밝혀야 하며, 또한 참여자 수를 충분히 확보하여 실행한 대규모 양적 연구를 통하여 양적 연구의 분석 결과에 대한 재검증 작업을 할 수도 있을 것이다.

둘째, 채점 과정 보고의 수집 과정에서 일부 연구 참여자들은 점수 결정 과정에 대한 언어 프로토콜 보고의 수행의 어려움을 언급하기도 하였다. 이와 관련하여 해당 채점자는 기초 면담 과정에서 평소 말을 잘 하지 않는 편이고, 내성적인 편이어서 구두로 보고를 하는 것에 대한 큰 부담을 갖고 있다는 언급을 한 경우가 있었다. 분석 결과에서도 해당 참여자의 보고량은 다른 참여자에 비하여 적은 것으로 나타났는데, 언어 프로토콜 보고에서 참여자의 성향으로 인한 영향에 대하여 에릭슨과 사이먼(1993)에서도 말을 잘 하는 참여자와 그렇지 않은 참여자의 차이가 있을 수 있다는 의견을 밝힌 바 있으며, 따라서 질적 자료

의 타당성을 더 확보하기 위해서는 양질의 언어 프로토콜 보고 수준을 나타낼 수 있는 추가적인 절차가 필요할 것으로 보인다. 본 연구에서는 채점자들의 심리적 영향과 일반적인 한국어 말하기 평가 상황을 고려하여, 연구자가 없는 상황에서 두 차례의 연습 채점과 그에 대한 피드백을 제공하는 것으로 보고 연습을 할 수밖에 없었으나, 언어 프로토콜 보고를 중심으로 연구를 진행할 경우에는, 발화 산출을 위한 추가 연습 과정이 필요할 것으로 보인다. 또한 언어 프로토콜 보고가 학습 효과를 나타낸다는 선행 연구의 의견(Swain, 2006)과 관련하여서도 본 연구의 결과를 바탕으로 채점자들이 채점을 진행하면서 나타내는 시간에 따른 변화가 있었는지를 알아보기 위한 사전-사후 검증 또는 시계열 측정 등을 바탕으로 하는 추가적인 연구가 이루어질 수 있을 것이다.

셋째, 본 연구는 채점자의 채점 과정을 중심으로 말하기 평가에 관한 이해를 확장하고 타당화 근거를 마련하고자 접근한 것이었다. 이러한 접근은 수험자가 구성형 응답 문항의 과제를 어떻게 이해하고, 이를 바탕으로 어떻게 응답을 계획하여 어떻게 전달하는지의 응답 과정에 관한 연구와 연계될 때 실제적으로 말하기 교육을 위한 내용과 방안 등이 제시될 수 있을 것으로 예상된다.

다섯째, 채점자인 교사가 현실적으로 평가 결과를 활용할 때 겪는 어려움으로는 적용할 수 있는 교육의 체계가 잡혀 있지 않다는 점(Price et al., 2012)과 교사마다 해석한 평가 기준의 범주가 상이하다는 점(Barkaoui, 2007), 그리고 기본적으로 교사들의 ‘평가 전문성(assessment literacy)’ 부족의 문제를 지적할 수 있을 것이다. 말하기 평가의 채점 과정 연구에서 살펴본 채점 과정의 양상과 특징이 교육적 환류로서 나타나기 위해서는 한국어교육 현장에서 학생들을 평가하고 있는 교사가 자신의 채점 특성을 파악하고, 교육의 수월성을 넘어 책무성과 공평성을 나타낼 수 있는 바람직한 평가가 이루어질 수 있도록 평가에 대한 인식 파악 및 개선을 위한 광범위한 기초 연구가 이루어져야 할 것이다.

끝으로 연구 수행 결과를 바탕으로 구안한 채점 과정 기반 한국어 말하기 평가의 채점자 교육 방안이 실제적으로 어떤 효과를 나타낼 것인지와 관련하여 기존의 채점자 교육의 방식과의 비교를 위한 실험 연구가 수행되어 검증이 이루어져야 할 것이다.

<참고문헌>

1. 국내 저서

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

강현주(2013). 한국어 말하기 평가의 구인으로서 상호작용 능력 연구, 고려대학교 박사학위 논문.

강승호·김양분(2004). 신뢰도, 서울: 교육과학사.

국립국어원(2017), 국제 통용 한국어 표준 교육과정 적용 연구, 서울: 국립국어원.

김경선·이규민·강승혜(2010). 일반화가능도 이론을 적용한 한국어 말하기 성취도 평가의 신뢰도와 오차요인 분석. 한국어 교육, 21(4), 51-75.

김상경(2015). 학문 목적 한국어 학습자의 말하기 능력 평가 방안 연구. 경희대학교 박사학위논문.

김상수(2011). 한국어 학습자의 말하기 평가 담화에 나타난 결속장치 사용 연구. 언어와 문화, 7(2), 33-53.

김정숙·원진숙(1993). 한국어 말하기 능력 평가기준 설정을 위한 연구. 이중언어학, 10, 24-33.

김정숙(2014). 한국어 말하기 능력 구인 설정을 위한 기초 연구. 한국어 교육, 25(4), 1-21.

김지영(2018). 한국어 말하기 평가 채점자의 채점 경향 연구. 연세대학교 박사학위논문.

김평원(2007). 성인화자의 말하기 평가 방법; 의사소통과정으로서의 말하기 평가: 대규모 성인 말하기 평가 시스템을 중심으로. 화법연구, 11, 9-34.

김평원(2010). 말하기 평가의 분석 모형 연구. 서울대학교 박사학위논문.

김현아(2016). Comparing Native and Non-native Rater Assessments of Korean Oral Proficiency: A FACETS analysis. 국어교육학연구, 51(5), 84-113.

김현주(2007). 영어 말하기 평가의 채점과정 연구. 영어영문학연구, 49(4), 169-186.

나카가와 마사오미(2014). 한국어 말하기 수행 평가 연구. 서울대학교 박사학위논문.

민병곤(2004). 논증 교육의 내용 연구. 서울대학교 박사학위논문.

민병곤·조수진·홍은실·박현정·강석한·이성준·오예림·이승원·안현기(2017). 학문 목적 한국어 말하기 평가 도구 개발 연구. 국어교육, 157, 309-340.

박현정·박민호·이성준·강석한(2017). 문항반응이론을 활용한 학문 목적 한국어 말하기 평가 문항 및 채점자 특성 분석. 이중언어학, 73, 177-203.

백순근(2002). 수행평가. 서울: 교육과학사.

송민영·이용상(2015). 영어 말하기 채점자의 행동 특성 분석. 교과교육학연구, 19(4), 1081-1101.

신동일(2004). 한국의 영어평가학. 서울: 한국문화사.

원미진·김지영(2017). 한국어 말하기 평가 개발을 위한 채점 경향 분석 연구. 외국어로서의 한

국어교육, 47, 169-192.
오승영(2019). 이주민 대상 한국어시험의 맥락 타당도 연구. 배재대학교 박사학위논문.
윤은경(2008). 말하기 숙달도 평가에 관한 논평: ACTFL OPI를 대상으로. 한국어 교육, 19(1), 1-25.
이선애·박선철(2013). 신경경제학과 정신의학. 신경정신의학, 52(5), 301-310.
이성준(2015). 한국어 말하기 평가 연구의 연구 방법에 대한 고찰, 한국화법학회 제32회 전국학술대회 자료집. 21-37.
이성준(2018). 한국어 말하기 평가 타당화를 위한 논거 기반 접근법의 이해와 적용. 화법연구, 41, 85-116.
이영식(2004). 한국어 말하기 시험의 유형 및 채점 기준 설정을 위한 기초 연구. 한국어교육, 15(3), 209-230.
이영식(2006). 영어 말하기 평가의 개발. Studies in English education, 11(2), 1-18.
이정모(2001). 인지심리학. 서울: 아카넷.
이정모(2009). 인지과학. 서울: 성균관대학교 출판부
이준호(2009). 한국어 수행 평가의 원리 및 방안 연구. 고려대학교 박사학위논문.
이향(2013). 한국어 말하기 수행 평가의 발음 범주 채점에 대한 타당성 검증. 이화여자대학교 박사학위논문.
전나영·한상미·윤은미·홍윤혜·배문경·정혜진·김수진·박보경·양수향(2007). 한국어 말하기 능력 평가 도구 개발 연구. 외국어로서의 한국어교육, 32, 259-338.
전은주(1997). 한국어 능력 평가. 한국어학, 6, 153-173.
정광·고창수·김정숙·원진숙(1994). 한국어 능력 평가 방안 연구: 언어 숙달도 (proficiency)의 측정을 중심으로. 한국어학, 1(0), 481-538.
정화영(2000). 한국어 말하기 숙달도 평가 방안. 연세대학교 교육대학원.
주미진(2014). 영어 말하기 평가의 채점자 신뢰성과 편향성 조사. ENGLISH TEACHING, 69, 247-270.
지은림·채선희(2000). Rasch 모형의 이론과 실제. 서울: 교육과학사.
지현숙(2005). 인터뷰 시험 담화 분석을 통한 한국어 구어 능력 평가의 구인 연구. 국어교육연구, 16, 79-104.
지현숙(2006). 한국어 구어 문법 능력의 과제 기반 평가 연구. 서울대학교 박사학위논문.
최은규(2006). 유형별로 본 한국어 능력 평가의 실제와 과제: 배치 시험과 성취도 시험을 중심으로. 한국어 교육, 17(2), 289-319.

2. 국외 저서

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2006). *Standards for educational and psychological testing*. Washington, DC: American Educational Research

Association.

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.

Anson, C. M. (2006). Assessing writing in cross-curricular programs: Determining the locus of activity. *Assessing Writing*, 11(2), 100-112.

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge: Cambridge University Press.

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

Bachman, L. F. & Palmer, A. S.(2010), *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Baker, F. B. & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Boca Raton, FL: CRC Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing*, 12(2), 86-107.

Baars, B. J. & Nicole M. G. (2010). 인지, 뇌, 의식: 인지신경과학 입문서. 강봉균 (역). 파주: 교보문고. (원서출판 2007)

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In Williamson, D. M., Mislevy, R. J. & Bejar, I. I., (eds) *Automated scoring of complex tasks in computer-based testing*. (pp. 49-81). Hillsdale, NJ: Lawrence Erlbaum Association.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9.

Bohrnstedt, G., Rossi, P., Wright, J., & Anderson, A. (1983). Handbook of survey research. *Measurement*. San Diego: Academic press.

Branthwaite, A., trueman, M., & Berrisford, T. (1981). Unreliability of Marking: further evidence and a possible explanation. *Educational Review*, 33(1), 41-46.

Brown, A. & Hill, K. (1998). *Interviewer style and candidate performance in the IELTS oral interview*. International English Language Testing System (IELTS) Research Reports 1998, 1, 1.

Brown, A. (1995). The effect of rater variables in the development of an

occupation-specific language performance test. *Language Testing*, 12(1), 1-15.

Brown, A. (2000). *An investigation of the rating process in the IELTS oral interview.* International English Language Testing System (IELTS) Research Reports 2000, 3, 49.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks*. ETS Research Report Series, 2005(1), i-157.

Brown, A. (2006). *An examination of the rating process in the revised IELTS Speaking Test*. International English Language Testing System (IELTS) Research Reports 2006, 6, 1.

Brown, G. D. & Yule, G. (1983). *Teaching the spoken language.* Cambridge: Cambridge university press.

Brownell, J. (2007). 듣기: 태도, 원리 그리고 기술. 이시훈, 한주리 (역).. 서울: 커뮤니케이션북스. (원서출판 2007)

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition, & assessment* (pp. 55-73). Cambridge: Cambridge University Press.

Chapelle, C. A., Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.

Charness, N. (1981). Aging and skilled problem solving. *Journal of Experimental Psychology: General*, 110(1), 21.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6(3), 271-315.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Creswell, J. W. & Plano Clark, V. L. (2007), *Designing and Conducting Mixed Methods Research*, Thousand Oaks, CA: Sage.

Creswell, J. W. (2011), 연구방법: 질적, 양적 및 혼합적 연구의 설계, 서울: 시그마프레스. (원서출판 2009)

Creswell, J. W. & Zhou, Y. (2016). What is mixed methods research. In A. J. Moeller, J. W. Creswell & N. Saville, (Eds.). (2016). *Second language assessment and mixed methods research*. (pp. 30-50). Cambridge: Cambridge University Press.

Cronbach, L. J. (1971). Test validation in education measurement, In R. L. Thomdike (ed.). *Educational Measurement* (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education,

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cumming, A., Kantor, R., & Powers, D. E. (2001). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117-135.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience*. (Unpublished doctoral dissertation). University of Hawaii at Manoa,

De Groot, A. D. (1978). *Thought and choice in chess*. The Hague: Mouton Publishers.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt: Peter Lang.

Embretson, S. E. & Reise, S. P.(2000), *Item Response Theory for Psychologists.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A., Ericsson, R. R., Hoffman, A., Kozbelt, & A. M. Williams, (Eds.). *The Cambridge handbook of expertise and expert performance*, (pp. 223-242). Cambridge: Cambridge University Press.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3), 215.

Ericsson, K. A. & Simon, H. A. (1984). P*rotocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K. A. & Crutcher, R. J. (1991). Introspection and verbal reports on cognitive processes-two approaches to the study of thought processes: A response to Howe. *New Ideas in Psychology*, 9, 57-71.

Ericsson, K. A. & Simon H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press,

Ericsson, K. A. & Kintsch, W. (1995). Long-term working memory. *Psychological*

review, 102(2), 211.

Ericsson, K. A., Patel, V. L., & Kintsch, W. (2000). How experts' adaptations to representative task demands account for the expertise effect in memory recall: Comment on Vicente and Wang(1998). *Psychological Review*, 107, 578−592.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. (Unpublished doctoral dissertation). University of Lancaster.

Fulcher, G. (2003). *Testing second language speaking*. Glasgow: Pearson Education Limited.

Gass, S. & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65−87.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. New York, NY: Viking.

Green, A. (1998). V*erbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed−method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255−274.

Greene, J. C. (2007). *Mixed methods in social inquiry* (Vol. 9). New York, NY: John Wiley & Sons.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika*, 26(1), 93−107.

Haertel, E. H. (2006). Reliability. In B. Brennan (ed.). *Educational measurement* (4th ed., pp. 65−110). Westport, CT: Praeger.

Hamp−Lyons, L. (1991). Scoring procedures for ESL contexts. In *Assessing second language writing in academic contexts*. (pp. 241−276). Norwood, NJ: Ablex.

Howell, W. S. (1982). *The empathic communicator. Prospect Heights*, IL: Waveland Press Inc.

Hubley, A. M. & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D., Zumbo, & A. M. Hubley, (Eds.). *Understanding and investigating response processes in validation research* (pp. 1−12). Cham: Springer.

Hughes, A. (2003) *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

International English Language Testing System. (2019). *Speaking assessment criteria*

[PDF file]. Retrieved from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

International Language Testing Association (2018). *Code of Ethics for ILTA*(Translated in Korean), Retrieved March 3, 2019 from www.iltaonline.com/resource/resmgr/docs/code_of_ethics/ilta_coe_korean_2018.pdf

Isaacs, T. (2016). Assessing speaking. In G. Fulcher. *Handbook of second language assessment* (pp. 131-146). London: Rutledge.

Isaacs, T. & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135-159.

Isaacs, T. & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, 32(1), 113-140.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice,* 18(3), 239-258.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). New York, NY: Cambridge University Press.

Kim, H. J. (2011). I*nvestigating raters' development of rating ability on a second language speaking assessment*. (Unpublished doctoral dissertation). Teachers College, Columbia University.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448-457.

Kiddle, T. & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, 8(4), 342-360.

Kienpointner, M. (1992). How to classify arguments. In F. H. van Eemeren, R. Grootendorst, J.A. Blair & C.A. Willard, *Argumentation Illuminated* (pp. 178–188), Amsterdam: Amsterdam University Press.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment.

Language Assessment Quarterly, 12(3), 239-261.

Kim, Y. (2009a). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics*, 30, 435-440.

Kim, Y. (2009b). An investigation into native and non-native teachers' judgments of oral English performance: A mixed-methods approach. *Language Testing*, 26, 187-217.

Kim, Y. (2009c). Exploring rater and task variability in second language oral performance assessment. In A. Brown, & K. Hill, (Eds.), Tasks and criteria in performance assessment (pp. 91-109). *Proceedings of the 28th Language Testing Research Colloquium*. Frankfurt am Main: Peter Lang.

Lane, S. & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement* (pp. 387-431). Westport, CT: American Council on Education & Praeger.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language testing*, 13(2), 151-172.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23(2), 131-166.

Lee, Y. J. (2018) Investigating Rater Effects Using Many-Facets Rasch Measurement: An Application of Myford and Wolfe (2003, 2004). *Secondary English Education*, 11(4), 165-192.

Leighton, J. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement*. (Unpublished doctoral dissertation). University of Chicago.

Linacre J. M. (1994). *Many-Facets Rasch Measurement* (2nd Ed.). Chicago, IL: MESA Press www.rasch.org/books.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, 16(2), 878.

Linacre, J. M. (2012). *Many-Facets Rasch measurement: Facets tutorial*. Retrieved from http://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2019) *Facets computer program for Many-Facets Rasch measurement*, version 3.81.0. Beaverton, Oregon: Winsteps.com

Linacre, J. M, Wright, B. D. and Lunz, M. E. (1990). *A Facets model for judgmental scoring.* MESA Memo 61. http://www.rasch.org/memo61.htm.

Llosa, L. M. (2005). B*uilding and supporting a validity argument for a standards-based classroom assessment of English proficiency*. (Unpublished doctoral dissertation). University of California.

Lumley, T. (2000). *The process of the assessment of writing performance: the rater's perspective*. (Unpublished doctoral dissertation). University of Melbourne.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.

Lynch. B. K. & McNamara, T. F. (1998). Using G-Theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Matsugu, S. (2013). *Effects of rater characteristics and scoring methods on speaking assessment*. (Unpublished doctoral dissertaion). Nothern Arizona University.

McNamara, T. F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational i settings. *Language Testing*, 14(2), 140-156.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes*. (Unpublished masterís thesis), University of California at Los Angeles.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Studies in Language Testing*, 3, 92-111.

Mislevy, R. J. & Brennan, R. (2006). Cognitive psychology and educational assessment. In B. Brennan (Ed.). *Educational measurement* 4, (pp. 257-305). Westport, CT: American Council on Education & Praeger

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1), 3-62.

Myford, C. M. & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system*. Princeton, NJ: Educational Testing Service.

Myford, C. M. & Wolfe. E. W. (2000). *Monitoring sources of variability within the test of spoken English assessment system*. Princeton, NJ: ETS Research Report Series. Educational Testing Service.

Myford C. M. & Wolfe, E. W. (2003) Detecting and measuring rater effects using Many-Facets Rasch measurement: Part I. J*ournal of Applied Measurement*, 4(4), 386-421.

Myford C. M. & Wolfe, E. W. (2004) Detecting and measuring rater effects using Many-Facets Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227

Newton, P. E. (1996). The reliability of marking of general certificate of secondary education scripts: Mathematics and English. *British Educational Research Journal,* 22(4), 405-420.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, 30(2), 143-154.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, 36(2), 219-233.

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1994). Diagnostic reasoning and medical expertise. Psychology of Learning and Motivation-advances in *Research and Theory*, 31(C), 187-252.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. M. Gass (Eds.), Tasks and language learning: Integrating theory and practice (pp. 9-34). Clevedon: Multilingual Matters.

Pollitt, A. & Murray, N. L. (1996). What raters really pay attention to. *Studies in Language Testing,* 3, 74-91.

Pollitt, A. & Crisp, V. (2004). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?*. Paper presented at the British Educational Research Association Annual Conference. 16-18.

Purpura. J. E. (2013). Cognition and language assessment. In A. J. Kunnan. (Ed.). *The companion to language assessmen*t (pp. 1452-1476). Oxford: Wiley/Blackwell.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, 65(S1), 37-75.

Price, M., Rust, C., O'Donovan, B., Handley, K., & Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning*. Oxford, Oxford Brookes University.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment:

Affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113-125.

Quellmalz, E. & Burry, J. (1983). *Analytic scales for assessing students' expository and narrative writing skills* (CSE Resource Paper No. 5). Los Angeles, CA: Center for the Study of Evaluation, University of California.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford: Nielsen & Lydiche.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In Elder, C., Brown, A., Grove, F., Hill, K., Iwashita, N., Lumley, T., McNamara, T., & O'Loughlin, K. (Eds.) *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Reed, S. K. (2012). *Cognition: Theories and applications*. Belmont, CA: Wadsworth Cengage Learning.

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: how the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education*, 30(3), 231-240.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses* (Vol. 6). Frankfrut am Main: Peter Lang Pub Incorporated.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, 26(1-2), 113-125.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology,* 85(6), 956.

Seong, Y. & Bottcher, E. (2011). Assessing academic presentation performance: Does the rater matter? Multi-Facetsed Rasch analysis of ESL learners' academic presentation task. *Paper session presented at the meeting of the 33rd Language Testing Research Colloquium*, Ann Arbor, MI.

Simpson, S. A. & Gilhooly, K. J. (1997). Diagnostic thinking processes: Evidence from a constructive interaction study of electrocardiogram (ECG) interpretation. *Applied Cognitive Psychology*, 11(6), 543-554.

Slater, S. J. (1980). Introduction to performance testing. In J. E. Spirer (Ed.), *Performance testing: issues facing vocational education* (pp. 3-17). Columbus, OH: National Center for Research in Vocational Education.

Spolsky, B. (1978). Introduction: Linguists and language testers. Approaches to

Language Testing. In *Applied Linguistics. Advances in Language Testing Series: 2*.(pp. v-x). Arlington, VA: Center for Applied Linguistics.

Stansfield, C. W. & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20(3), 347-364.

Suto, W. M. I. & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. In Green, S. (Ed.). *Research matters: 2* (pp. 7-11), Cambridge, UK: Cambridge Assessment.

Suto, W. I. & Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. Assessment in Education: Principles, *Policy & Practice*, 15(1), 73-89.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice*, 31(3), 21-30.

Swain, M. (2006). Verbal protocols. In Chalhoub-Deville, M., Chapelle, C. A., & Duff, P. A. (Eds.). *Inference and generalizability in applied linguistics*: Multiple perspectives (Vol. 12, pp. 97-114). Amsterdam: John Benjamins Publishing.

Tashakkori, A. & Teddlie, C. (2003). Issues and dilemmas in teaching research methods courses in social and behavioural sciences: US perspective. *International Journal of Social Research Methodology*, 6(1), 61-77.

Tashakkori, A. & Teddlie, C. (2008). Introduction to mixed method and mixed model studies in the social and behavioral sciences. *The Mixed Methods Reader*, 7-26.

Thissen, D. & Orlando M. (2001). Item Response Theory for items scored in two categories. In: Thissen D, Wainer H (Eds), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Earlbaum,

Teddlie, C., & Tashakkori, A. (2015). **통합방법 연구의 기초: 사회.행동과학에서 양적 접근과 질적 접근의 통합**. 강현석 외(역). 서울: 아카데미프레스. (원서출판 2009)

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological review, 90(4), 293.

Tversky, A. & Kahneman, D. (Eds.). (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Van Moere, A. (2013). Raters and ratings. In Kunnan, J. (Ed.). *The companion to*

language assessment (Vol 3., pp. 1358-1374). London: Rutledge.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In Hamp-Lyons, L. (Ed.). *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation

Vygotsky, L. S. (1987). Thinking and speech. *The collected works of LS Vygotsky*, 1, 39-285.

Watson, J. B. (1920). Is thinking merely the action of language mechanisms?. *British Journal of Psychology*, 11, 87-104.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.

Wolfe, E. W., Kao, C.W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465-492.

Wolfe, E., Chiu, C., & Myford, C. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses*. Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Wolfe, E. W. & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31-37.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: Praeger

Wilson, M. & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrik*a, 58, 87–99.

Xi, X. & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST)*. ETS research report series, Princeton. NJ: Educational Testing Service.

Xi, X. & Mollaun, P. (2009). H*ow Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?* ETS Research Report Series, Princeton, NJ: Educational Testing Service.

Zhang, Y. & Elder, C. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET–SET solution. *Language Assessment Quarterly*, 6(4), 298-314.

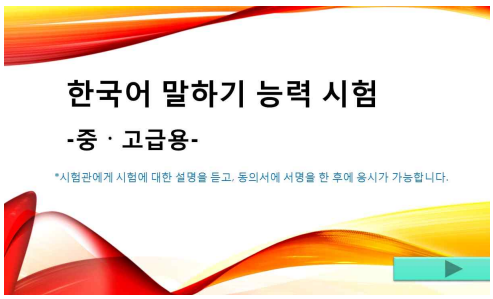
3. 교재

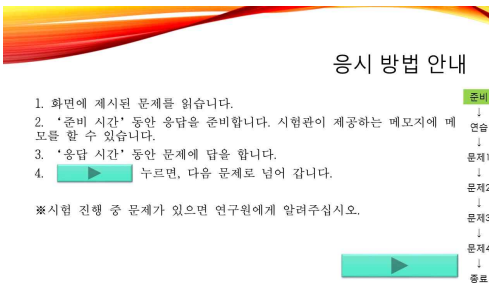
서울대학교 한국어교육센터(2015). 서울대 한국어 4A. 서울: TWOPONDS.
서울대학교 한국어교육센터(2015). 서울대 한국어 4B. 서울: TWOPONDS.
서울대학교 한국어교육센터(2012). 서울대 한국어 5A. 서울: 문진미디어.
서울대학교 한국어교육센터(2012). 서울대 한국어 5B. 서울: 문진미디어.
서울대학교 한국어교육센터(2015). 서울대 한국어 6A. 서울: TWOPONDS.
서울대학교 한국어교육센터(2015). 서울대 한국어 6B. 서울: TWOPONDS.
이화여자대학교 언어교육원(2011) 이화 한국어 4. 서울: 이화여자대학교출판문화원.
이화여자대학교 언어교육원(2012) 이화 한국어 5. 서울: 이화여자대학교출판문화원.
이화여자대학교 언어교육원(2012) 이화 한국어 6. 서울: 이화여자대학교출판문화원.
고려대학교 한국어문화교육센터(2010). 재미있는 한국어 4. 서울: 교보문고.
고려대학교 한국어문화교육센터(2010). 재미있는 한국어 5. 서울: 교보문고.
고려대학교 한국어문화교육센터(2011). 재미있는 한국어 6. 서울: 교보문고.
국립국어원(2014) 세종한국어 5. 서울: 하우.
국립국어원(2014) 세종한국어 6. 서울: 하우.
국립국어원(2014) 세종한국어 7. 서울: 하우.
국립국어원(2014) 세종한국어 8. 서울: 하우.

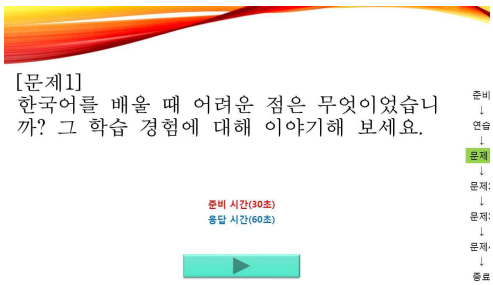
<부록>

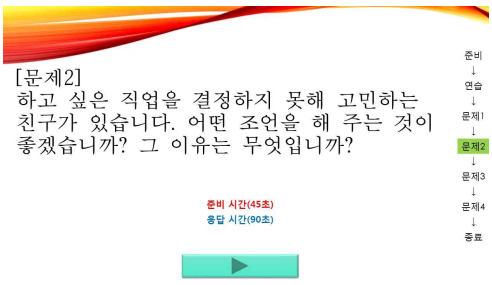
1. ‘한국어 말하기 능력 시험’의 문항
2. ‘한국어 말하기 능력 시험’의 채점 척도
3. 채점 과정 보고 수행을 위한 사전 교육 자료
4. 채점 과정 보고 절차 안내 자료
5. 채점 과정 보고 사례

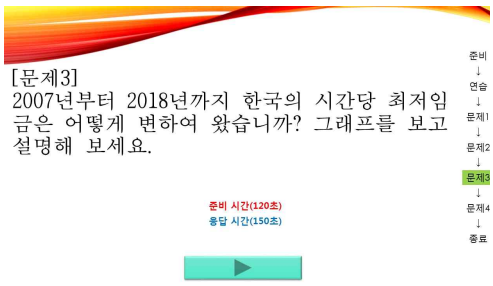
<부록 1> '한국어 말하기 능력 시험'의 문항

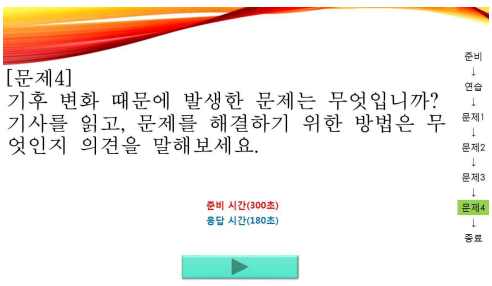












<부록 2> '한국어 말하기 능력 시험'의 채점 척도

영역	전반적 수행	발음	어휘·문법	담화 구성
5 탁월	●응답에 사소한 실수가 있지만 과제를 충족하며, 적절한 세부 설명을 포함함. ●표현의 실수가 거의 없고, 잘 이해할 수 있으며, 담화의 응집성이 있음.	●발화가 명확하며, 자연스럽게 이어짐. ●발음의 사소한 문제가 있으나 이해를 방해하지 않음.	●다양한 주제 관련 어휘와 복잡한 문형을 사용하여 적절하고 효과적으로 표현함. ●매우 적은 오류가 발견되지만 의미 이해를 방해하지 않음	●전체 내용이 주제에 관한 적절한 정보를 충분하게 포함하고 있음. ●흐름이 일관되며, 주제를 논리적으로 잘 전개함. ●완결된 구조를 갖추었으며, 문장 연결에 짜임새가 있음.
4 우수	●응답이 완벽하지는 않지만 과제를 상당히 적절하게 다루고 있음. ●표현의 몇 가지 실수가 있지만 명료하고 유창하게 응집성 있는 담화를 구성함.	●발화가 명확한 편이며, 반복과 수정이 거의 없지만, 머뭇거리거나 주저하는 일이 있음. ●발음의 문제가 종종 나타나지만, 이해를 위해 조금만 노력하면 됨.	●대체로 주제 관련 어휘와 복잡한 문형을 사용하여 적절하게 표현함. ●몇 가지 오류가 나타나지만, 의미 이해를 방해하지 않음.	●내용이 대체로 주제에 관한 적절한 정보를 포함하고 있음. ●흐름이 대체로 일관되며, 주제를 대체로 논리적으로 전개함. ●대체로 완결된 구조가 나타나며, 문장 연결에 짜임새가 있음.
3 보통	●응답에서 과제와 관련하여 어느 정도 적절한 내용을 다루고 있음. ●표현에 실수가 지속적으로 나타나고, 유창성이 다소 부족하지만 이해가능한 편임.	●발화가 덜 명확한 편이며, 반복과 수정, 주저하는 일이 종종 나타남. ●발음의 문제가 나타나며, 이해를 위해 노력해야 함.	●다소 제한적으로 주제와 관련된 어휘와 복잡한 문형을 사용하여 표현함. ●반복적인 오류가 있으며, 의미 이해를 다소 방해함.	●내용이 주제와 관련된 적절한 정보를 일부 포함하고 있음. ●흐름의 일관성이 부족하며, 주제 전개의 논리가 다소 부족함. ●구조가 잘 드러나지 않으며, 문장 연결의 짜임새가 다소 부족함.
2 미흡	●응답에서 과제를 다루고 있지만, 제한적으로 주제를 다루고 있음. ●표현을 이해할 수는 있으나, 전반적으로 발화 전달과 담화 응집성에 문제가 있음.	●발화에 반복과 수정, 주저하는 일이 자주 나타남. ●발음의 심각한 문제가 반복해서 나타나며, 이해를 위해 상당히 노력해야 함.	●주제와 관련하여 사용한 어휘가 단순하고 사용한 문형이 제한적임. ●심각한 오류가 일부 나타나며, 의미 이해에 어려움이 있음.	●내용에서 주제와 관련이 없는 정보가 많이 나타남. ●흐름의 일관성이 매우 부족하며, 주제 전개의 논리가 빈약함. ●구조가 거의 드러나지 않으며, 문장 연결의

				짜임새가 매우 부족함.
1 부 족	● 응답이 과제와 거의 관련이 없음. ● 표현을 대체로 이해하기 어렵고, 응집성이 매우 부족함.	● 발화에 긴 휴지와 주저하는 일이 자주 나타남. ● 발음의 심각한 문제가 전체적으로 나타나며, 이해를 위해 심각하게 노력해야 함.	● 주제와 관련하여 사용한 어휘가 매우 제한적이며, 단순 문형만을 사용함. ● 심각한 오류가 나타나며, 의미 이해에 큰 어려움이 있음.	● 내용이 대부분 주제와 관련이 없음. ● 흐름의 일관성이 없고, 주제 전개의 논리가 매우 빈약함. ● 구조가 드러나지 않으며, 문장 연결의 짜임새가 거의 없음.
0 채 점 불 가	● 응답한 내용이 과제와 전혀 관련이 없음. ● 다른 응답 내용이 없이 문제나 자료의 내용을 그대로 읽기만 하였음.			

<부록 3> 채점 과정 보고 수행을 위한 사전 교육 자료

연구 참여 방법 안내

안녕하세요, 본 연구에서 연구자는 선생님께서 어떤 생각을 하면서 말하기 시험 채점을 하시는 지에 관심이 있습니다. 이를 위하여 채점을 하는 과정에서 떠오르는 생각을 구술하여 주시기를 부탁 드립니다. 연구자를 통해 전달 받은 음성 파일들을 채점하면서 시작할 때부터 마칠 때까지의 생각을 소리 내어 모두 말해 주십시오. 가능하다면 녹음을 시작할 때부터 마칠 때까지 말이 멈추지 않기를 바랍니다. 채점 중에는 시험에 사용한 문제와 채점 기준표를 참고할 수 있으며, 그 밖의 자료는 참고할 수 없습니다. 학생 응답을 들은 후에는 기억을 하면서 채점을 해야 할 것입니다. 기억을 하기 어려운 경우에는 그렇다는 보고를 하시면 됩니다. 선생님의 생각을 연구자에게 설명하려고 하지는 마십시오. 그리고 서두르지 마십시오. 참여 방법을 잘 이해하셨습니까? 질문이 있으시다면 언제든 연구자에게 연락 주십시오. 선생님의 참여에 진심으로 감사 드립니다.

2018년 월 일 서울대 국어교육과 박사과정 이성준 올림

1. 사전 모임	● **목적 및 참여 방법 확인** ● **참여 중 주의 사항 확인** ● **사례를 통한 보고 방법 확인**
↓	
2. 연습 채점	● **3명의 학습자 자료에 대한 채점 (문제 및 채점 기준표 이해)** ● **2명의 학습자 자료에 대한 구두 보고 채점(구두 보고 채점에 대한 연습)** ● **채점 결과를 기록한 엑셀 파일 및 채점 과정에서 녹음한 음성 파일 파일 제출** ● **연구자 피드백 확인**
↓	
3. 본 채점	● **12명의 학습자 사례에 대한 채점 및 구두 보고** ● **채점 결과를 기록한 엑셀 파일 및 채점 과정에서 녹음한 음성 파일 파일 제출**

<부록 4> 채점 과정 보고 절차 안내 자료

	안녕하세요?채점을 시작하기 전에 다음과 같은 준비를 부탁드립니다.
1	채점 장소는 조용한 방이나 연구실이며, 다른 사람의 방해가 전혀 없는 곳이어야합니다.
2	컴퓨터와 헤드셋, 이메일로 전달 받은 채점용 음성 파을 준비합니다.
3	컴퓨터를 켜고, '녹음기' 또는 '음성 녹음기'를 실행하고, 녹음이 잘 되는지 확인한 후에, 녹음 버튼을 누릅니다.
4	평가 문항 및 채점 기준표를 준비하고, 채점 폴더를 엽니다.
5	준비가 되었으면, 폴더별 문제 순서에 따라 채점과 구두 보고를 시작합니다.
6	먼저 학습자의 응답을 들으면서 떠오르는 생각을 보고합니다.
7	학습자의 응답을 모두 들은 후에 모든 채점 영역 별로점수를결정하기까지의과정을보고하고,결정한점수는엑셀파일에기록합니다.
8	같은 방식으로 폴더별 1-4번 문제에 대한 응답을 채점합니다.
9	채점중간에쉬거나,다른일정이있어여러번에걸쳐나누어하실경우, 시작시간과종료시간을각각기록해주십시오.폴더별채점중간에채점이중단되지않도록주의해주십시오.

<부록 5> 채점 과정 보고 사례

엄격한 채점 경향 관련
#R05121 #R051211 유창하다 사용하는 어휘는... 약간 중급스럽다, 중급 이하의 표현을 쓰고…있다 #R051212 이 학생은 전반적으로 들었을 때는, 과제와, 과제와 관련해서 적절하게 과제를 어 수행하고 있어 보입니다 그런데 전반적인 표현 자체가 그렇게 고급 표현을 사용하는 게 아니라 중급 이하의 표현을 사용해서 유창하게 이야기를 하고 있는 것으로 봐서… 4점까지 주기에는 조금 어려움이 있을 것 같고 어, 그 다음에 끝나는 부분도 갑자기 완결되는 느낌이 조금 있어서 음 전반적 수행은 3점 정도를 주는 것이 적절해 보입니다. 발음은 상당히 자연스럽고 어 매끄럽습니다 반복도 별로 없었고 주저하는 일도 없고 그래서 사실 발음 같은 경우에는 4점 정도를 줘도 큰 문제가 없을 것 같습니다 다음 어휘와 문법은 어 내용을 잘 들어보면 사실 주제와 관련한 이야기를 하고 있기는 하지만 그 주제와 관련된 이야기를 할 때 사용하는 어휘 수준이 그렇게 높지 않고 어 중급 이하의 표현들을 주로 사용하고 있습니다. 그렇지만 어떤 복잡한 문형 자체를 사용하고 있기 때문에 4급을 주는 것은 어려울 것 같고 이 정도 수준이면 3급을 주는 것이 주는 것이 적절하, 아니 3점을 주는 것이 적절해 보입니다. 다음에 담화 구성도 음 지금 어느 정도 흐름... 그 과제에 대한 주제에 관련된 내용을 이야기 하고 있고 그에 대한 정보를 포함하고 있지만 그 끝나는 부분에서 갑자기 매듭 짓는 느낌이 있으면서 어 전체 문장 구조가 매끄럽지는 않아보였기 때문에 4점까지 주는 것은 어려워 보이고 이것도 3점을 주는 것이 적절해 보입니다. #R03121 #R031211 저 같은 경우에는 으로 시작을 하면 참 훌륭하다

되게

굉장히 구어에 친숙한 느낌?

#R031212

자 전반적으로 구어에 굉장히 강한 느낌이죠? 몇몇 표현들을 보면 구어에 워낙 강하기 때문에 노출이 많이 됐다라는 것이 딱 드러나면서

말하기가 좋습니다

전반적 5

어, 담화도 뭐 큰 무리 없는 수준 정도는 되겠고

담화 4

잘한 건 아니겠지만 어휘나 발음이 큰 문제가 없었(으)리라고 봅니다

어휘문법4,

발음5

#R091211

#R091212

먼저 발음이나 억양에는 크게 문제가 없는 것 같은데

약간 머뭇거림이 있어서

발음은4점을..우수를 주고

어휘와 문법은

지금 너무 구어에다가 복잡한 문형 이런 거 없었고 단순한 거 같았어

어휘와 문법은 2점

담화구성은

구성에 문제.. 근데 짜임새 있게 말한 것은 아니라서

유창하긴 했지만

담화구성 3점

적절한 정보를 포함하고 있지만 논리적으로 말하지는 않았어

생각나는 대로 열거한 것 같아서

3점을 주고

전반적인 수행은

유창성은 부족하지 않은데

응집성 있게 구성하는 것 같진 않아

전반적인 수행 3점

관대한 채점 경향 관련
#R110511 일본 학생 처음부터 좀 주저함이 있고 발음입니다.. 발음이가 내가.. 저는.. 수정을 했는데 오히려 더 틀렸고 해지 못한? 이해하지 못한? 발음이가.. 오류가 지속적으로 있구나 #R110512 그러면 일단 이 학생을 어휘 문법을 보면 발화가 명확한 편이며 반복과 수정이 거의 없지만 머뭇거리며 주저하는 일이 있었지 발음 문제도 종종 나타났고 그래서 이 학생은 어휘..발음은 일단 4점을 주면 되겠고 어휘 같은 경우는 대체로 주제와 관련한 어휘와 복잡한 문형을 사용해서 적절하게 사용.. 표현함 오류 나타나지만 의미를 방해하지 않음 조사 오류가 좀 있었지만 의미 이해를 방해할 정도는 아니었고 4점 그 다음에 담화 구성 적절한 정보는 포함하고 있었고 흐름이 일관됐고 논리적으로 전개했고 4점? 5점이 안 되는 이유는 충분하게 포함되고..뭐 그리고..충분하다? 이런 표현들은 이 학생에게 어울리지 않으니까 4점 정도 주고 그 다음아 전반적 수행 보면 완벽하진 않지만 적절하게 다루고 있었고 약간 실수 있지만 명료하고 유창하게.. 유창.. 유창.. 유창성은 다소 부족했던 거 같은데

실수가 지속적으로 나타나진 않았으니까

4점

#R090512

음 먼저 발음은요

받침 발음이 잘 안되는 것 같고요 받침 발음이 정확하지 않고

명확하게 발화하지 않는 거 같네요

자신감이 없는지 명확하지 않아서 전달력이 좀 떨어지고
중간에 계속 자기가 말하는 걸 확실하지 않은지 수정을 하고 싶은 건지 응? 이런 거 많이 해서
좀 중간 중간 이해하기가 어려웠던 것 같습니다.
그래서 발음은.. 심각한 문제가 있는 것은 아닌데 수정하고 주저하는 것이 좀 많았던 것 같아요
그래서 3점 보통을 줄게요.

그리고 어휘는

주제와 관련 없는 어휘를 사용하거나 하진 않는데

어려운 어휘나 복잡한 문형이 있었던 것 같지는 않아요

고급 표현이 있었던 것.. 고급 표현도 없었던 것 같고

그렇지만 이게 오류가 있어서 의미를 방해한 것은 아니라서

3점 보통을 주겠습니다.

담화 구성은

발음이 어려운데 일본어에는 없는 발음이 있고 내가 말하면 상대방이 이해를 못한다 이런 것은 이유를 들어서 말했네요

일관성이 부족하지는 않아요
그런데 우수한 구조를 뒀다고는.. 아 우수한 구조로 과제를 수행했다고는 할 수 없을 것 같아요
조금 더 논리적으로 이야기를 하면 좋지 않았을까 해서

네 우수는 줄 수 없을 것 같고요

보통 3점을 주겠습니다

전반적인 수행은

마찬가지로 3점이고요

계속 발화하는데 수정하는 것도 그랬고

유창성도 좀 떨어졌던 것 같아요

그런데 어느 정도는 질문에 맞는 과제를 대답했다고 생각해서

3점을 주겠습니다.

#R040512

발음이

뭐 특별한 문제가 있는 건 아닌데 발화가 굉장히 불분명한 게 많네요.
계속 응? 응? 이러면서 자기 물음을 표시하는 듯한.. 그런걸 굉장히 많이 하고.. 말하다가 갑자기 응? 응? 이런 소리를 엄청 많이 냈고
메모를 할 시간이 없을 정도로 굉장히 빨리 끝났는데 지금 뭘 얘기해야되나

발음 발화는

그러면.. 2번 아 2점 미흡

뭐 발음에 문제가 있어서 못 알아듣거나 이런 건 아닌데

발화면에서 반복, 수정, 주저하는 일인데 매우 자주 나타났어요.

그래서 2점밖에 못 줄 거 같아요

어휘문법?

뭐 고급 수준의 어휘라고 할 게 거의 하나도 없어요 지금
심지어 자기가 말하고 싶은 거를 전해.. '전해지 못하고' 이렇게.. '전하다'를 '전해지'.. 이거 굉장히.. 고급이면은 이 정도로 틀리면 안 되는 거 같은데.. 아닌가.. 안 되면.. 그리고 '내가' 왜 말을 '내가'라고.. '제가'도 아니고 '저가'도 아니고 '내가'? 그리고 '내가'로 말하다가 갑자기 '저는' 이러고.. 너무 못하는데 지금..
그래서 어휘 문법 점수를 주면 2점 줄까 1점 줄까.. 지금 보고 있는데 기준표.

주제와 관련하여 사용한 어휘가 매우 제한적이며 단순 문형..

그래도 말을 안 한 건 아니니까 2점 줘야 되나.. 1점을 줘야 되나..

1점도 줄 수 있을 것 같아요 얘는
왜냐하면 제가 지금 머리에 남은 게 없어요 얘가 뭘 말하는지..내용이 지금 기억이 나지 않을 정도로.. 표현이 너무 자연..제한적이었거든요 지금
이해에 큰 어려움이 있었다.. 네

담화구성, 구성..

뭐 구성이라 할 게 없었지 지금

그냥 내가 가장 어려운 거는 발음이다 이러고선.. 일본사람이라서..그래서 뭐 어쨌다고..
발화가 명확하게 끝나지 않아서 일본사람이라서 뭐 어떻게 했다는 건지 이유를 제대로 제가 이해를 못 했어요
내용 조직도 잘 안됐지..

그럼 미흡이나 부족인데

1점 줘도 될 거 같은데

표를 보니까? 흐름에 일관성 없고, 논리도 빈약하고, 구조도 드러나지 않고, 짜임새도 없고..

1점

전반적인 수행. 경험에 대해서 어려운 점 이야기..

그냥 발음이 어렵다라는 거 말고는 지금 뭐 경험 이야기 한 게 없는데? 들은 게 없는데 경험을.. 그죠?

응답에서 과제를 다루고 있지만 제한적으로 주제를 다루고 있음, 표현을 이해할 수 있으나 전반적으로 발화 전달과 담화 응집성에 문제가 있음..

1점까진 아닌 거 같아요 이 표를 보니까

2점 주면 되겠네 전반적인 수행은 이렇게 할게요

집중 경향 관련

#R050111

시작 괜찮음

음. 소통을 사통으로 말함

표정을 표현으로 자가 수정

발음이도

선생님보다 발음이 나빠질 수 있다?

#R050112

자 우선 이 학생은

주제와 관련해서 어느 정도 음…적절한 내용을 다루고 있고

어... 중간 중간에 실수가 조금 지속적으로 나타나기 때문에 유창성이 다소 부족하다고 보여져서

전반적 수행은... 3점. 줄 수 있을 것 같습니다

다음에 발음은

어. 모국어의 영향을 많이 받는 거 같고.

중간에 약간? 주저? 그러고 수정? 수정하는 경우도 있었고.

그렇기 때문에 지금 발음에서도 어, 뭐 아까 사,소통을 사통이라고 얘기한다거나 이것은 약간 주의를 기울여야 하는 부분이기 때문에

발음에서도 3점 줄 수 있을 거 같고.

어휘와 문법에서는

어... 지금 아까 뭐였더라? 발음이도처럼 명사에다가 2개의 조사를 결합해서 말하기는 하는데,

이게 한국인이 사용하지 않는 그런 문법 결합 형태로 이야기를 하는 경우가. 어. 있어서...

어휘 문법 부분에서도 의미 이해를 약간 다소 방해한다고 보고
음...3점.
어...심각한 오류는 아니기 때문에
3점 정도를 주는 것이 적절…한 것 같습니다.
다음에 담화 구성은…
약간 처음 시작은 괜찮았는데 뒤쪽으로 갈수록 전개가 약간 어색하고 논리성의 부족…
논리가 약간 다소 부족하기 때문에
어, 그렇다고 해서 주제와 관련이 없는 정보가 있는 것은 아니고
어. 그냥 논리가 다소 부족하다고 보여지고
전체 문장에 종결이 제대로 안돼서 짜임새가 부족하다라고 봤을 때
담화 구성도 3점을 주는 것이 어... 적절하다고 보여집니다.
#R020111
분야
사투
#R020112
음 첫 번째 말하기 말하기가 많이 어렵다고 했고
이유도 충분하고…
전반적으로… 실수없이 과제 내용을 충분히 잘 설명하고 있고 논리적인 포장
발음. 약간…나이가 있는 사람 같은데
발음… 정확한 편이지만 일본사람인가? 약간 일본인 억양이 있는데
약간의 머뭇거림이 있었음
4점
어휘, 간단한 질문이지만 고급 어휘를 많이 사용하려고 했음
고급 문법을 많이 없었는데
오류 조사 오류 몇 개 정도 의미를 방해하지 않음
어휘 문법 4점
담화 구성, 내용이 주제를 주제와 관계가 있고 중구난방 아니고 논리적으로 전개를 잘 했음
5점
#R100111
분야인 것 같습니다로 말했어야 하는데 일 것 같습니다로 말했고
시제오류?

수정이 한 번 있었고
발음이도
조사 오류 두 번이나 오류 발생했고
#R10112
전체적으로 내용은 한국어 배울 때 말하기가 좀 어렵다
문장이 문장을 구사할 때 문법에 오류는 거의 없었는데
전체적인 내용이 주제에 대해서 적절한 정보를 충분하게 포함하고 있다고 보고
사용 어휘도 뭐 소통에 필요하니까 좀 이해할 수 있어야 되는데
알아들을 수 없는 발음이나 사투리 이런 게 어렵다라고 말했으니까
예를 들어 설명하는 것
그리고 주제에 어긋나는 뭐 흐름을 방해하는 건 없었고
대체로 담화 구성면에서는 괜찮았다고 봐야 될 것 같은데
조사 사용 오류 그거 하나 정도만
그리고 문법사용 하나 오류 있었던 거 그 정도 외에는
발음도 명확했고
그래서 어휘 문법에서는 오류가 있었으니까
4점으로 주고
담화 구성은 5점
발음은 특별히 고향 언어 때문에 방해 받는 아주 큰 요인들은 없었고
중간에 휴지나 아니면 수정 수정하는 부분들이 있었지만 그 거의 없다고 봐야되지
보통 3점을 줄 수는 없을 거 같은데
종종 나타났다고 보기는 어려우니까
4점 이것도 4점
전반적인 수행은
조금 더 풍부했으면 좋겠다
그렇지만 과제를 아주 적절하게 잘 다루고 있었고
표현의 몇 가지 실수가 있었으니까
담화는 잘 구성했고
그러면 이것도 4점으로 줘야 되겠네

채점 척도 사용 제한 관련

#R050311

나름유창함
전반적으로 자연스럽게 이야기를 하고 있다
#R050312
문법이 가장 어렵고 표현이 조금 어렵다는 이야기를 하고 있고
지금 이 학생 같은 경우에는 전반적 수행은 과제를 그래도 상당히 적절하게 수행하고 있는 것으로 보여집니다.
어... 이 학생은 한국어를 배울 때 문법이 어려운데 그 이유가 문법이 굉장히 여러 가지 의미를 가지고 있기 때문에 어렵다라는 얘기도 했고
그 다음에 표현도 어렵지만 지금도 계속 계속 표현을 어렵다, 나름대로 명료하고 유창하게 이야기를 하고 있어서
전반적 수행은 4점까지 줘도 나쁘지 않을 것 같습니다
어 다음에 발음은 중간에 한 두 번 정도 같은 말을 어휘를 반복하는 경우가 있었고
중간에 약간 휴지가 있기는 하지만
그게 이케 듣기에 제일 정도가 아니었기 때문에
이것도 4점 정도를 줘도...될 거 같습니다.
다음에 어휘와 문법도 대체적으로 주제와 관련된 어휘를 사용하고 있고
중급 이상에서 사용될 수 있는 부사적 부사 표현이라던가 좀 내용을 다양하게 사용하고 있어서
3급보다는 3점보다는 좀 더 줄 수 있…4? 우수하다고 보여집니다.
다음에 흐름이 대체적으로 일관되고
주제에 대해서도 대체적으로 논리적으로 전개를 하고 있기 때문에
그 다음에 마무리도 급하게 끝날 듯하지만 그래도 어느 정도 문장 연결에 짜임새가 있는 구조라고 여겨져서
이 학생은 전반적으로 4점 정도 줘도 충분한 언어 실력을 가지고 있다고 생각합니다.
#R090312
어 먼저 발음은
발화가 명확하고 대체적으로 자연스럽게, 매끄럽게 이야기를 했던 거 같습니다.
일본인 화자라서
약간 그 받침 발음이 어색한 것들이 있는데
의미 이해를 방해하진 않았던 것 같습니다.
그래서 저는 발음은 5점을 주겠습니다.
어휘와 문법은
고급.. 중급 수준의 표현들을
적절하게 잘 사용했던 것 같고요
효과적이었던 것 같습니다

오류가 없었던 건 아닌데
의미 이해하는 데 전혀 방해가 되지 않았고요
그래서 어휘와 문법도 5점을 주고 싶습니다.
담화구성은
지금 두 가지가 문법과 표현이라고 했죠
표현이라면 어휘.. 이런 것들을 말하는 거죠?
두 번째.. 이렇게 둘을 나눌 거라면 두 번째 것도 좀.. 뭐가 어려웠는지 설명했으면 좋았을 텐데 약간 아쉬웠던 것 같아서
담화 구성은 4점
전반적인 수행은 꽤 대답을 잘 했던 것 같아요
자기의 어려운 점을 충분히 설명했고
세부 설명을 포함하고 있었음(AR)
응집성도 있었던 거 같고
마지막이 약간 아쉽긴 한데
전반적으로 대답을 잘했던 것 같습니다
그래서 전반적인 수행도 5점을 주겠습니다.

채점자 편향 관련

#R120411
오후, 어휘?
속도가 느린편이다
#R120412
일단 발음은 전반적으로 깔끔한 편이기는 한데
처음에 어휘라고 말하는지 오후라고 말하는 건지
모… 음 발음이 좀 명확하게 들리지 않았다
어휘..라는 표현을 그 다음에 다시 사용했기 때문에
그리고 나서는 명확하게 들렸고
어, 중간에 조금 주저하는 pause가 어..있었기 때문에
음.. 아주 우수하다고는 볼 수 없을 것같다
하지만 발,발화를 이해하기 어려운 수준은 아니었기 때문에
어, 3점 아니면 4점 줄 수 있을 것같고
모음에서 나타났던 오류 외에는 전반적으로 깔끔한 발화, 발음이었던 거 같아서
어, 발음 점수는 4점 주겠다
다음으로 어,어휘와 문법 있어서는 전체적으로 어, 뭐.. 그냥 보통 수준의 어휘와 문법을

사용을 했고
지금 어휘가 어렵다고 말을 하면서 예로 든 것은 거울과 겨울이라는 발음 문제로 인해 구분이 안 되는 영역이었기 때문에
어.. 좋은 예를 든거 같지는 않다
따라서 들어서 어휘 문법 점수는
3점을 주도록 하겠다
세 번째로 담화 구성은 어, 주제에 대한 아.. 발화를 하고 있기는 하지만
내용이 굉장히 짜임새가 있다든가
어, 굉장히 주제에 어울리는 적절한 정보를 담고만 있다고 느껴지지는 않아서
그냥 보통 수준의 학생, 보통 수준의 발화였다고 어,느껴진다
따라서 마찬가지로 담화 구성 점수도 3점을 주도록 하겠다
마지막으로 전반적인 수행은 어..
유창성이나 정확성에서 우수한 수준은 아니었고
또 이런저런 실수들이 중간에 조금씩 나타났..기는 하지만
뭐 이해 가능한 편이었기 때문에
보통 해당되는 3점을 주도록 하겠다
#R110411
자연스럽네
원어민.. 어휘
있는데..
발음이 좋네
#R110312
발음이랑.. 발음이.. 발음이랑 억양이 아주 깨끗하네
그럼일단발음은
듣기어려운게하나도없었고
맞다 자기가 거울이랑 겨울이 헷갈린다고 했는데
그거 발음도 명확했고
그러니깐 발음은 일단 5점을 주고 시작하고
그 다음에 어휘 문법
어휘 문법 오류도 별로 없었던 거 같은데
오류 별로 없었고
적절하고 효과적으로 표현함
조금 포즈는 있었지만 그 정도 포즈는 네이티브들도 만드는 포즈들이니까 괜찮았고
그 다음에 담화 구성을 보면
흐름이 일관되며

처음에는 뭐에 대해서 얘기했지
처음에는.. 어휘가 어려웠다 자긴 중국 원어민이지만 그래도 어휘가 어려웠다 그렇게 얘기했지
담화 구성도 적절했고..
그러면 담화도5점
전반적으로
사소한 실수 실수가 있었나 거의 없었던 거 같고
응집성도 있었고
잘했음
5점, 5점, 5점, 5점
#R040212
다음에 어휘 문법 수준을 보면
뭔가 정확한 어휘를 사용하기보다 한국 사람들을 사랑.. 사따라 하면 되는데 뭐 그렇게 하는 게 어려웠어요 이런 식으로 정확한 어휘가 생각나지 않아서 그렇게, 하면은 이런 식으로 그렇다, 뭐 하다 이런 좀 뭐라고 하지 어휘들을 많이 사용해서
정확한 표현을 하지 못한 거 같..
그리고 문법 표현 수준도 좀 고급에 해당하는 표현은 거의 나오지 않은 거 같아서
뭐..그 한 2, 3급 정도? 수준의 어휘가 문법만 사용이 된 걸로 지금 들렸기 때문에
2점이나 3점을 주면 될 거 같은데 표를 보면 미흡 수준이 주제와 관련하여 사용은 어휘가 단순하고 사용한 문형이 제한적이다, 심각한 오류가 일부 나타나며 의미 이해에 어려움이 있다 그리고 3점 수준은 다소 제한적으로 주제와 관련된 어휘와 복잡한 문형을 사용함, 표현에 반복적인 오류가 있으며 의미 이해를 다소 방해함
그 어떤 문장에서 거슬릴 정도의 큰 오류가 나타났던 것은 아니지만
전체적인 표현 수준이 2, 3급 정도 수준에 머물러 있었고
그래서 2점? 줄 수 있을 것 같아
담화 구성으로 넘어가 보면
자기가 말하기..말할 내용을 준비할 시간에 그 내용을 전체적인 내용을 다 구성하지 못했다는 느낌이 들었어요
크게 뭐 처음에 이렇게 시작해서 중간에 어떤 이야기를 좀 더 구체적으로 하고 그 다음에 마지막 이 어떻게 정리를 할지에 대한 그런 전체적인 흐름이 전혀 계획되어 있지 않은 그런 말하기였다
그래서 보면 몇 점을 줄 수 있을까 그니까 흐름에 일관성이 부족하고 전개 논리가 부족했다라는 게 들어가야 되는데
한 2점 줄 수 있을 것 같은데요
#R130212

어휘나 문법 그렇게 어려운 건 쓰지 않았지만 이 문제에 대해서는 어려운 단어가 필요하지 않고
문법도 근데 어떻게 받침을 하는지도 모르고 발음을 하는지도 모르고
이런 사소한 오류가 있었으니깐 그냥 5점을 줘도 될 것 같아

Abstract

An Analysis of the Rating Process for a Korean Speaking Assessment

Lee, Sungjun
Department of Korean Language Education, College of Education
The Graduate School
Seoul National University

Various factors may intervene in the rating process of the speaking assessment that affects the progress of the rating and causes different test results among the raters. A variable noted in previous studies is the effect on the characteristics of the rater, which can be understood by explaining the rating process in regards to both how the rating method was applied in some situations and how the scores were determined based on the rating process. With this variable in mind, the present study uses a mixed method research design to empirically explore the rating process as a direct variable involved in rating speaking assessment. The study seeks to identify the rating process of a speaking assessment by analyzing the rater effect and the rating process.

To this end, we establish a theoretical model to explain the rating process of the speaking assessment, which is made up of perception, the judgment of information, and a score decision. This is a hypothetical model that entails what the various variables involved in the rating process are and how they influence the rating process. We explain that the rater's cognition is based on the inferencing and heuristics points, which are presented in the discussion of the language assessment and cognitive psychology field during the rating process. In addition, we discuss changes in the rating process that could occur due to external and internal factors concerned with the rater.

Next, to find out the specific features and patterns of the rating process for a

speaking assessment, a computer-based, Korean-speaking achievement test collected the test-takers' responses. The speaking assessment consists of a mid- to high-level structured response item to specifically identify how the scorers perform. The writing of the tasks is based on the “International Standard Curriculum of Korean Language” and Korean language textbooks. In order to look at the combined rating patterns of the raters, the rating scale is based on the previous study for “task perform, pronunciation, language use, and discourse organizing.” After test-takers (N=18) applied for the test, twelve of their responses, which exclude the main material, were scored. Teachers (N=13) participated as raters and were asked the verbal report during their scoring process.

The test result data was collected from the raters, and it was first analyzed with the Classical Test Theory (CTT) and then in a Many-Facet Rasch Model (MFRM) to determine the rater effect. The specific analyses of the rater effect are based on subjective rating tendency, central tendency, randomness, halo effect, and interaction bias. The results of the CTT analysis show high internal consistency confidence among the raters in the rating, but the raters show a relatively low correlation. The ANOVA on variables of rater characteristics does not show significant differences among groups. The analysis of the MFRM shows some instances where the rating tendency is either not consistent or it is outside of the expectations of the analysis model. This factor depends on the test items, criteria, and scales.

Overall, the results of the analysis of the rater effect show that raters are generally lenient and that some raters score rigidly or leniently depending on specific items, criteria, and scale. The central tendency analysis shows a rater concentrating on the median of scale and another rater focusing on the maximum of scale. The randomness and halo effect analysis show corresponding cases with the analysis of the MFRM and the comparison of the score line. Rater bias analysis shows that interactions with the raters and tasks have the greatest probability of error with the highest frequency in the "experience speaking" item.

Next, to identify the flow and characteristics of the speaking assessment rating process, the raters themselves analyzed the rating process report. To analyze the overall aspects of the rating process, we selected examinees from the test results of the four item types at the median, the highest, and the lowest levels, and looked at the differences in reporting the rating process. The analysis shows that the raters score sequentially in most of the items and groups, while the ratio of comprehensive rating remains relatively high in the rating process for mid- and lower-level test-takers in the "experience speaking" item. The rating process report of a sequential scoring case shows that the rater tends to determine the score based on the overall impression of the response and in the case of a summative rating, the score is determined by considering the response information or characteristics. In addition, the differences in the amount and type of warrants or assumptions that were formed to determine scores in the rating process were seen by each rater because of the effect of perceptual level and memory system operation on the responses of the test-takers. In particular, reports of the rating process for median-level responses often show a quantitative approach based on perceptual information. This is interpreted as an active use of perceptual information to overcome the difficulty in determining the level and score compared to those at the highest or lowest levels. The report of the rating process of an integrated question using data shows a tendency to determine the overall assessment criteria score based on the consideration of a response's level of ability regarding discourse composition.

Next, we examine the rating process report for cases in which the rater effect is shown in the analysis of the rater effect. The rating process report analysis of cases showing strict scoring trends confirms that when "total performance" is first scored, the same trend applies to the scoring of the remaining criteria thereby present to the severity of the rating trend. Reports of rating processes with lenient rating trends indicate relatively lenient scoring or random grading in situations where evidence collection for responses is not well-performed. If there is a tendency to use the assessment scale on a limited basis, the rating process considers impression-based approaches and the assumptions, which are interpreted

as a phenomenon. This is because information that is not aggregated by test-takers is not considered in the rating process while trying to integrate their responses into the existing memory system.

Finally, the rating process, which shows a random scoring tendency, is characterized by a simple structure compared to the other raters, which is considered random because the rating process relies on subjective interpretations of the rating scale without considering the judgment of the response and the theoretical and empirical assumptions for the scoring.

In order to utilize the research results in the training of raters for the Korean speaking assessment, the implications of the raters' education are derived from the results of the research and based on this, the principles and methods of a process-based approach are presented.

keyword: Korean Language, speaking assessment, rating variance, rating process, rater effect, rating process report, rating process-based rater education.

Student Numner: 2013-30424