

# 학문 목적 한국어 말하기 평가 도구 개발의 쟁점과 과제\*

민 병 곤\*\* · 이 성 준\*\*\*

— < 次 例 > —

- I. 논의의 배경
- II. 선행연구 검토
- III. 학문 목적 말하기 평가 도구 개발 사례 분석
- IV. 학문 목적 한국어 말하기 평가 도구 개발의 쟁점
- V. 학문 목적 한국어 말하기 평가 도구 개발의 과제

## I. 논의의 배경

이 글은 유학생을 대상으로 한 학문 목적 한국어 말하기 평가 도구 개발의 주요 쟁점과 과제를 제시하는 것을 목적으로 한다.

교육부의 ‘2016 교육 기본 통계’에 따르면 국내 외국인 유학생은 104,262명에 이르며, 그 중 대학이나 대학원 과정에서 유학 중인 이들은 63,104명으로 한국어를 배우기 위해 온 어학 연수생에 비해 2.3 배나 많은 것으로 나타났다. 정부는 2004년에 발표한 ‘스터디 코리아 프로젝트’를 필두로 고등교육 신뢰도 및 경쟁력 강화를 위해 글

\* 이 연구는 한국연구재단의 지원(NRF-2015S1A5A2A03049743)을 받아 수행된 것이다.

\*\* 서울대학교 사범대학 국어교육과 부교수 / 국어교육연구소 겸무연구원

\*\*\* 서울대학교 사범대학 국어교육과 박사과정

로별 인재 유치를 대학에 장려하고 있으며, 한국의 대외적 신인도 상승도 유학생 유치의 동력을 제공하고 있다.

국내 대학(원)의 유학생 숫자가 지속적으로 증가하는 가운데, 유학생이 겪는 큰 어려움은 대학 생활을 수행하기에 한국어 실력이 부족하다는 점이다. 이와 관련하여 유학생의 성공적인 대학 생활을 위해 학문 목적 한국어 교육이 필요하다는 논의는 한국어교육 연구에서 꾸준히 제기되어 왔다(김정숙, 2000; 이해영, 2004; 나임순, 2006; 박선옥, 2009; 민진영, 2013, 2014; 유해준, 2015). 기존의 연구들은 주로 유학생의 한국어 학습 현황이나 인식에 대한 기술에 초점을 맞추고 있으며, 교재 분석이나 교육과정 등에 대한 거시적인 논의에 그치고 있다.

유학생을 위한 학문 목적 한국어교육이 성공적으로 이루어지기 위해서는 유학생과 현장의 요구를 반영한 학문 목적 한국어 교육과정과 교수·학습 방법, 교재, 평가 등이 체계적으로 개발되어야 하며, 이를 위한 기초 연구들이 뒷받침 되어야 한다. 지금까지 학문 목적 한국어교육과 관련한 연구들은 학습자의 맥락보다는 교육 필요성을 근거로 교육 내용과 방법을 제시하고, 쓰기 관련 내용을 다루는 것이 대부분을 차지하고 있다(최정순·윤지원, 2012).

유학생의 입장에서 학문 목적 한국어에 대한 필요가 있는 것은 대학(원) 입학과 관련하여 한국어 능력을 증명하는 일이다. 현재 국내 대학에서는 유학생의 선발을 위해 한국어능력시험(TOPIK)이나 한국어 교육기관의 수료 등급을 기준으로 사용하는 것이 일반적이다. 그런데 TOPIK은 일상생활 중심의 일반 목적 한국어를 대상으로 하며, 말하기 시험은 시행하지 않는 한계가 있다<sup>1)</sup>. 또한 교육기

1) TOPIK 관련 보고서(김정숙 외, 2010; 윤희원·서혁, 2010)에서도 A-TOPIK(학문 목적 한국어 능력 평가)과 TOPIK-Speaking을 추가한 TOPIK을 제안한 바 있다.

관이 부여한 한국어 등급은 기관마다 교육과정이 서로 다르고, 학문 목적 한국어 교육에 해당하는 내용을 다루고 있지 않는 경우가 많아 적용에 무리가 따른다. 더 문제가 되는 것은 현재 유학생의 입학 조건으로 제시된 한국어 능력은 필수 사항이 아니며, 일부 대학에서는 유학생 유치를 늘리기 위해 TOPIK 성적을 확인하지 않거나, 응시 결과만 제출하기도 한다는 점이다.<sup>2)</sup> 유학생의 한국어 말하기 능력은 공인된 평가 도구가 없기 때문에 한국어능력시험의 평가 대상인 어휘나 문법, 읽기, 듣기, 쓰기와 달리 학습 동기와 자원이 제공되지 않는 상황이어서 유학생을 위한 학문 목적 한국어의 말하기 능력과 평가에 대한 논의가 필요한 상황이다.

말하기 평가에 관한 논의는 측정과 분석, 해석의 어려움으로 인하여 한국어교육에서도 아직 활발하지 않은 연구 영역이며, 타당한 자료 수집을 바탕으로 분석 및 보고가 이루어져야 하는 상황이다. 따라서 본 연구에서는 학문 목적 말하기 평가의 핵심 문제를 도구 개발에 있다고 보고, 이와 관련한 이론적 배경과 사례를 검토하여 관련된 쟁점들을 분석하고, 학문 목적 한국어 말하기 평가 도구 개발의 방향을 제안하고자 한다.

## II. 선행연구 검토

### 1. 학문 목적 언어 평가

---

2) 국내 대학의 유학생 선발 기준 중 한국어 능력에 관하여 대학의 유학생 담당자를 대상으로 조사를 실시하여 확인한 내용이다. 조사는 2016년 4월부터 5월까지 실시하였으며, 해당 결과는 관련 학술지를 통해 게재될 예정이다.

학문 목적 언어 평가는 학문적인 맥락에서 이루어지는 외국어 사용 능력에 대한 평가를 의미한다. 평가 대상인 학문 목적 언어는 특수 목적 언어(language for specific purposes, LSP)에 해당하며, 특정한 언어 사용 목적을 반영한 내용과 맥락이 포함된다. 존스와 더들리 에반스(Johns & Dudley-Evans, 1991)는 특수 목적 언어<sup>3)</sup>가 학문적으로 정립되는 과정에서 가장 큰 영향을 끼친 요소로 요구평가(needs assessment)와 담화분석(discourse analysis)을 제시하였다. 이들은 특수 목적 언어활동을 파악하기 위한 방법이다. 요구평가는 특수 목적 언어 분야에서 사용되는 대표적인 연구 방법 중 하나로서, 학습자의 요구를 조사하여 교수·학습을 개선하기 위한 용도로 사용한다. 담화분석은 전형적이고 실제적으로 특수 목적 언어를 파악하기 위해 사용되는 방법으로서 분석의 주체와 분석 기준, 접근 방법 등이 정해져야 한다. 학문 목적 언어 평가는 특수 목적 언어로서의 특성을 반영하여 평가 대상 언어 사용자의 요구와 의사소통 현장의 실제적인 측면을 바탕으로 구성해야 하며, 평가로서 갖춰야 할 타당성을 확보해야 한다.

스케한(Skehan)은 특수 목적 언어 평가를 둘러싼 여러 쟁점들을 제기하는 연구에서 특수 목적 언어 평가 도구의 개발을 위해 해결해야 할 여러 과제를 제시하였는데, 다음의 네 가지 문제로 요약할 수 있다(Skehan, 1984: 204-214).

첫째, 특수 목적 언어가 일반적인 언어 능력과 어떤 관계를 맺고 있는지를 밝혀야 한다. 이는 보편적으로 수용되고 있는 일반적인 언어 능력과의 비교를 통해 평가 대상 언어의 특징을 드러낼 수 있다.

3) 존스와 더들리 에반스(1991)에서 제시한 용어는 ‘특수 목적 영어’(English for Specific Purposes, ESP)이지만, 이때의 영어가 학습 대상인 언어 자체이기 때문에 본 연구에서는 특수 목적 영어를 일반화한 ‘특수 목적 언어’를 사용하였다.

전통적인 관점에서는 언어 능력을 하나의 통합된 언어(unity of language)로 보려는 관점이 우세하였으나, 근래에는 언어 사용(의사소통)에 내재된 여러 능력이 존재한다는 관점이 중심을 이루게 되었다(Canale & Swain, 1980; Bachman, 1990). 이러한 변화는 특수 목적 언어 능력이 의사소통의 목적과 맥락에 따라 독특한 내용과 수준으로 구성될 수 있다는 것을 의미하며, 평가 대상인 언어의 내용과 상황이 불확실하고 포괄적인 일반 목적 외국어에 비해 구체적이고 실제적일 수 있게 하였다.

둘째, 특수 목적 언어 능력을 구성하는 요소는 무엇이며, 이들을 구별하는 특징이 무엇인지 파악해야 한다. 이는 평가 도구 개발의 핵심인 구인(construct)에 대한 타당도 확보의 문제와 관련되어 있으며, 스케한도 구인을 통해 평가 도구가 갖는 의미가 드러날 수 있다고 보았다.(Skehan, 1984: 209) 평가 도구의 구인타당도를 확보하기 위해서는 이론적인 측면에서 충분한 검토가 이루어져야 하며, 이를 통해 평가 도구는 정당성을 부여 받게 된다. 결국 구인타당도를 확보한다는 것은 언어 능력 및 평가에 관한 이론과 실제 수행을 연결하고자 하는 것이며, 이를 바탕으로 평가의 과정과 도구 자체의 타당도를 확보할 수 있다.

셋째, 특수 목적 언어 능력의 평가 내용을 구성하는 언어 기능(function)이 타당한 것이어야 한다. 스케한(1984)은 ELTS 시행 당시에 과제 선정과 관련하여 사용된 먼비(Munby, 1978)의 세부 언어 기술 분류 항목 238개가 실제로 평가되기 위해서는 수험자의 기저 능력을 측정할 수 있어야 하며, 일반화하여 적용할 수 있어야 한다고 보았다. 이는 안면타당도(face validity)의 문제로 귀결되는데, 전통적인 평가에서는 안면타당도 조사를 통해 평가 도구에 관한 외형적인 의견만을 수집하지만, 실제적인 관찰을 통해 평가 대상 언어와

의사소통에 대한 안면타당도로 발전되어야 한다고 보았다.

넷째, 특수 목적 언어 능력 평가에서 다루어지는 언어 기술(skill)은 무엇인지 그리고 어떻게 접근할 것인지를 명확히 해야 한다. 언어 기술은 과제 수행의 조건이 되는데, 이러한 언어 기술 항목들이 서로 유사하거나 더 세분할 수 있는 경우에는 수험자들이 과제 수행에서 혼란을 겪을 뿐만 아니라 평가를 하는 주체로서도 이 기술에 대한 평가 가능성을 다시 검토해야 할 수밖에 없다. 또한 특수 목적 언어의 평가라는 점에서 현재의 특수성이 언제까지 유지될 것이며, 특히 학문 목적 외국어 평가의 경우에는 담화가 구성되는 강의 맥락에서 양식(style)에 따라 어떤 차이를 두어야 하는지에 관한 고려도 이루어져야 한다.

## 2. 학문 목적 한국어 말하기 교육

지금까지 학문 목적 한국어 말하기 교육 연구는 학문 목적 한국어교육의 세부 분야로 포함되어 교육과정과 과제에 관한 논의가 주로 이루어졌다(이덕희, 2003; 황현주, 2005; 송지현, 2005; 윤신애, 2008; 옥건화, 2014).

이덕희(2003)는 한국어 교육과정 설계를 위해 학부생 94명과 대학원생 26명을 대상으로 설문 조사를 하여 학문적 언어 기술 및 일상 언어 기술과 일반적인 언어 요소에 대한 의견을 조사하였는데, 그 결과 학부생의 경우에는 듣기나 읽기와 관련되는 이해 기술에 대한 요구가 나타났으며, 대학원생의 경우 발표나 보고서 쓰기 등의 표현 기술에 대한 요구와 어려움을 나타냈다. 송지현(2005)에서는 학부생 147명을 대상으로 하여 학문 목적 과제에 대한 설문 조사를 실시하였는데, 그 결과 중요도 측면에서는 이해도가 높게 나타났지

만, 어려움에 대한 인식 조사에서는 말하기와 쓰기 같은 표현 기술을 더 어렵게 생각하는 것으로 나타났다. 윤신애(2008)는 학부생 102명과 대학원생 20명을 대상으로 학문 목적 말하기·듣기 과제에 대한 설문 조사를 실시하였다. 조사 결과에서는 강의 중 노트 필기와 교수가 학생에게 질문하기, 구두 수업 참여, 수업 중 토론에 대한 필요도가 상대적으로 높은 것으로 나타났으며, 전공별로는 인문대 학생들이 타 전공에 비해 교수에게 질문하기에 대한 필요성을 더 느끼는 것으로 나타났다. 그리고 대학원생은 발표하기와 발표 후 질문, 수업 중 토론, 교수에게 질문하기에 대한 필요가 상대적으로 높게 나타났으며, 전체 집단에서 가장 어렵게 생각하는 과제는 발표하기였다. 옥건화(2014)는 학문 목적 한국어 교재 분석을 바탕으로 ‘과업 기반 학문 말하기 교육’을 위한 교수·학습 방안을 제시하였다.

이상의 학문 목적 한국어교육 연구에서 말하기 평가에 대한 시사점을 살펴보면 다음과 같다. 첫째, 학문 목적 한국어 말하기에 대한 요구는 재학 상태에 따라 차이가 있으나, 공통적으로 표현에 대한 어려움을 느끼고 있었다. 이는 학생들이 실제 학문 목적 맥락에서 요구되는 과제를 수행하기 위해 필요한 표현 능력을 점검받지 못하였으며, 말하기 능력의 경우에는 TOPIK에서도 평가가 이루어지지 않고 있어 대학 입학 후에 한국어로 표현하는 데 어려움이 크다는 것을 의미한다. 둘째, 학생들은 수업 중에 이루어지는 교수의 질문에 대한 부담을 갖고 있으며, 이는 즉석에서 논리를 파악하고 내용을 생성하여 표현해야 하는 학문 목적 말하기 과제에 대한 요구를 나타내는 것으로 볼 수 있다. 셋째, 학문 목적 한국어 학습자의 실제적인 요구를 반영한 교육과 평가가 이루어져야 한다. 기존의 연구들은 말하기 교육의 과제를 일반적인 말하기 기술 중심으로 접근하여 전공 강의와 같은 실제적인 학문 목적 한국어 말하기 과제를 제시

하는 데 한계가 있었다. 학문 목적 한국어 말하기 평가에는 대학에서 유학생들이 노출되는 한국어 말하기 상황과 그 속에서 파악된 실제적인 요구가 반영되어야 한다.

### 3. 학문 목적 한국어 말하기 평가

학문 목적 한국어 말하기 평가에 관한 연구는 평가 도구와 과제, 채점 기준 개발을 중심으로 2000년대 후반부터 꾸준히 이어지고 있다(한상미, 2009; 박광진, 2010; 김상경·박동호, 2014; 심재경, 2014; 서동선, 2015; 김상경, 2015; 박선운, 2015; 권순용, 2015; 박자영, 2016).

한상미(2009)와 박광진(2010)은 평가 범주와 과제 및 채점에 이르는 학문 목적 말하기 평가 도구의 구성과 세부 내용을 제시하였는데, 평가 대상의 측면에서 한상미(2009)는 국내 대학 입학 준비생을, 박광진(2010)은 학문 목적 한국어 말하기 교육을 받은 유학생을 대상으로 하였다는 차이가 있다.

박선운(2015)과 권순용(2015), 박자영(2016)도 학문 목적 한국어 말하기 평가개발에 관한 연구를 수행하였는데, 모두 국외 말하기 평가 도구에 대한 분석을 기초로 하고 있지만, 나머지 연구들이 도구 개발과 관련하여 과제 개발을 중심으로 접근한 것과 달리 권순용(2015)은 평가 기준을 중심으로 접근하고 있었다.

학문 목적 말하기 평가와 관련하여 양적 연구를 통한 접근을 한 김상경·박동호(2014)와 김상경(2015)는 평가 도구의 타당도 확보를 위해 전문가 조사를 실시하고, 평가 결과를 다국면 라쉬 모형(multifaceted Rasch analysis)과 일반화가능도(generalizability) 연구를 통해 타당화하였다. 이 두 연구는 매우 유사한 형식을 취하고 있는데, 양적 연구를 통하여 문항 개발과 채점, 그리고 결과 활용 과



정의 타당도를 검증 하였다는 점에서 주목할 만하다.

평가 도구 설계뿐만 아니라 평가 실행 계획과 연결한 연구 중에서 서동선(2015)은 말하기 시험의 대표 과제로 ‘듣고 말하기’를 들고, 시험 설명서(test specification)를 작성하여 제시하였다.

학문 목적 한국어 말하기 평가 연구의 특징은 다음과 같다.

첫째, 학문 목적 말하기 평가 연구의 주요 과제로 평가 과제 및 틀 구성을 바탕으로 도구 개발을 목표로 삼고 있다. 말하기 평가는 여러 단계로 구성된 순환적인 활동이며, 도구 개발은 그 중의 중요한 한 단계가 되며, 과제 개발 또한 도구 개발의 한 부분이 되는 것이다(Luoma, 2004: 4-5). 기존의 연구들은 이러한 순환적인 평가 과정보다 형식적인 측면에서 과제나 평가 틀(framework) 구성에 집중하고 있었다.

둘째, 국외 평가 도구 분석과 언어 능력 평가 이론 분석을 바탕으로 학문 목적 말하기 평가 도구 개발 과제에 접근하고 있다. 검증된 평가 과제와 측정 및 채점 방법 등을 알아보기 위해 기존의 평가 도구에 대한 검토가 필요한 것은 당연한 사실이다. 그러나 해당 평가 도구들이 개발된 원리와 실제로 어떻게 활용되고 있는지를 함께 검토하지 않고 도구 자체를 항목별로 정리하여 비교만 할 경우에는 용어와 해석, 활용 측면에서 오해를 불러일으킬 수 있으며, 해당 도구의 언어와 사용 환경과 맥락 등이 한국어와는 다르다는 점도 고려하기 어렵다.

셋째, 사용하는 용어의 개념이 서로 다르며, 혼재되어 있다. 선행 연구들에서는 학문 목적 말하기 평가의 대상에 대해, 입학의 목적으로 하는 학습자와 현재 학문 목적으로 수학 중인 학습자, 그리고 이미 유학생생활을 시작한 학습자 등 여러 대상을 포괄하여 학문 목적 한국어 학습자로 부르고 있다. 또한 이들에게 요구되는 학문 목적

한국어 말하기 평가 영역을 주로 학문적 기술에 초점을 맞추고 있어 평가 목적과 대상이 불분명한 것으로 보인다. 이와 같은 혼란은 평가와 관련된 용어에서도 나타나는데, 사용 맥락에 따라 구별하여 사용해야 하는 구인, 평가 범주, 평가 기준 등이 혼재되어 나타나고 있어 이에 대한 정리가 필요하다.

### Ⅲ. 학문 목적 말하기 평가 도구 개발 사례 분석

#### 1. 토플(TOEFL)

ETS(Educational Testing Service)에서 주관하는 학문 목적 영어 능력 평가 시험인 토플(The Test of English as a Foreign Language™, TOEFL®)은 1963년에 미국 내의 여러 기관에 지원하는 비영어민들의 영어 능력을 확인하기 위한 목적으로 ‘외국어로서의 영어 시험 위원회’(the National Council on the Testing of English as a Foreign Language)에 의해 개발되었다.

토플의 말하기 시험에 대한 논의는 토플과 함께 필요에 따라 선택적으로 응시하던 TSE(Test of Spoken English)와 현재의 인터넷 기반 토플(TOEFL iBT)에서 시행하고 있는 말하기 시험(TOEFL® Academic Speaking Test, TAST)에 이르는 개발 과정을 통해 살펴볼 수 있다.

##### 1) 토플의 말하기 시험 개발 과정

TSE는 영어의 구어 의사소통 능력 측정을 위한 것으로 1980년에

ETS가 개발하였으며, 주로 미국과 캐나다의 대학에서 비원어민 수업 조교(international teaching assistant, ITA)의 선발이나 기타 특수 직종<sup>4)</sup>에서의 활동 자격을 인증하기 위한 목적으로 사용되었다. 시험은 테이프 녹음기를 사용하여 미리 녹음된 고등 교육 수준의 언어 기능을 확인하기 위한 문제를 듣고 응답을 녹음하는 방식으로 이루어졌다.

TSE는 개발 과정에서 도구의 타당화를 위해 FSI(Foreign Service Institute)의 말하기 시험 점수를 바탕으로 공인타당도 연구와 실제 응시자의 교실 의사소통에 대한 예측타당도<sup>5)</sup>를 알아보는 연구를 실시하였다(Clark et al, 1980). TSE는 이해가능도(comprehensibility), 발음(pronunciation), 문법(grammar), 유창성(flucy)으로 채점이 이루어지는데, 초기 시행 과정에서 실시한 채점자 내 신뢰도에 대한 분석에서는 전체 채점 기준에서 신뢰도가 .77~.85로 나타났으며, 각 채점 기준의 상관관계에서는 이해가능도가 발음이나 유창성 점수와 밀접한 것으로 나타났다. TSE 시험의 점수와 FSI 점수의 상관은 다른 토플 평가 영역(청해, 어휘와 문장 구조, 독해)보다 훨씬 높게 나타났다. 예측타당도 연구는 실제 TSE와 FSI 시험의 결과를 각각 강의 조교로서의 영어 의사소통 능력과 그 밖의 교육적인 말하기 수행으로 보고, 학생 조교를 통해 여러 관련 정보와 실제 교육 현장에서의 말하기 능력에 관해 조사과 다중 회귀 분석을 실시하였다. 그 결과 FSI와 TSE 점수는 응시자들의 학

4) 루오마(Luoma, 2001: 225)에서는 이와 관련된 직업으로 간호사, 의사, 약사, 물리 치료사, 수의사 등을 예로 들었다.

5) 클락 외(Clark et al, 1980)에서는 예측타당도에 관한 연구를 '실제 사용 타당화'(use-validation)라고 하였으나, 본고에서는 현재의 교육측정학적 개념을 따라 예측타당도로 명명하였다.

문적인 영어 의사소통 수행과 수업 중 질문 상황, 일대일 대화 상황 등의 말하기 능력을 잘 예측하는 것으로 나타났으며, 이는 거주 기간이나 학습 기간, 유사 경험에 비해 매우 높은 상관을 나타냈다. 채점 기준에서 나타나듯이 초기의 TSE는 주로 발음과 문법, 즉 언어 자원에 관련된 요소를 평가 대상으로 하여 실제 의사소통능력에 대한 평가 도구가 아니었다. 이에 따라 당시의 의사소통능력 이론에 대한 검토를 기반으로 하여 TSE 개편을 위한 연구를 실시하였다.

볼트와 올트만(Boldt & Oltman, 1993)은 초기 TSE에 대한 요인 분석과 다차원척도법을 적용한 분석을 실시하였는데, 그 결과 문항의 형식이 측정하고자 하는 능력에 영향을 미치며, 한명의 채점자가 한 피험자의 여러 과제를 모두 채점하면서 발생하는 ‘후광 효과’(halo effect)로 인해 구인별 상관이 높게 나타남을 확인하였다. 또한 여러 평가 기준으로 채점을 하는 것이 유용하지만, 효과적이지는 않다는 지적도 하였다.

초기 TSE는 시행 과정에서 함께 수행된 연구 결과들을 반영하여 개정되었는데, 헨닝 외(Henning et al, 1995: 2)는 개정된 TSE의 특징을 다음과 같이 제시하였다. 첫째, 기존의 TSE의 일곱 평가 과제 중 낭독과 문장 완성, 그림 보고 말하기를 내용타당도 및 안면 타당도의 문제와 시험의 효율성 확보를 위해 제거하였다. 둘째, 평가 척도의 신뢰도와 차별성을 높이기 위해 기존의 4단계 점수 체계를 5단계로 변환하였다. 셋째, 수험자의 응답 발화 양이 늘어날 수 있도록 지시하는 내용을 줄이고, 문항도 줄여서 응답자의 공백이나 채점에서의 공백도 줄였다. 넷째, 시험의 개발 배경과 발화 사례 채점에 관한 체계적인 근거를 제시하여 수험자와의 소통을 강화하였다. 다섯째, 점수 보고 척도(score-reporting scale)의 폭을 줄이고, 이를 통해 준거 참조 평가로서 각 등급에 해당하는 말하기 수행의 일관성

을 높이고자 하였다. 여섯째, 유창성, 발음, 문법 정확성 등의 채점 하위 영역의 평균이 전반적인 이해 가능도와 맞지 않으며, 채점이나 결과 사용 측면에서 효율적이지 않다는 기존의 연구 결과를 수용하여 총괄적 척도인 의사소통 언어 능력 척도(scale of communicative language ability)로 대체하여 사용하였다. 개정된 TSE는 1995년부터 도입되었으며, 도입과 동시에 이에 대한 타당성을 확인하고 실제 의사소통 맥락이 강화된 토플을 준비하기 위한 연구들이 이루어졌다(Lazaraton & Wagner, 1996; Douglas, 1997; Douglas & Smith, 1997; Butler et al, 2000).

라자라톤과 와그너(1996)는 개정된 TSE 시험에 대한 원어민과 비원어민의 발화를 분석하였는데, 높은 수준의 비원어민 화자 발화에서 원어민 화자와 같은 수준의 수행이 나타났다고 보고하였다. 다만 상호작용적인 기능과 관련한 평가가 매우 한정적으로 이루어지고 있어 실제 환경에서 청자의 인식 수준에 대한 고려가 채점 척도에 실제적으로 반영되어야 한다고 지적하였다. 더글라스(1997)는 말하기 기능의 실현과 관련한 이론적인 검토와 사용 맥락을 고려한 말하기 과제의 기능 통합적인 접근이 필요하다고 주장하였다. 이 연구에서는 토플 시험이 미국의 대학이나 대학원의 입학 자격으로서 사용되는 것을 고려하여 학생 맥락에서의 말하기 능력의 구인과 언어 이해 및 생산의 체계적인 통합, 그리고 언어 지식의 복합 요소적인 특성, 전략적 능력의 중요성 등을 다루고 있으며, 이러한 이론적 검토는 새로운 토플의 과제에서 기능 통합적인 측면의 고려를 통해 실제적인 능력 측정이 이루어져야 한다는 점을 강조하고 있다. 더글라스와 스미스(1997)에서는 TSE가 학문적 맥락에서의 구어 의사소통 능력을 측정하는 평가 도구인 것은 맞지만, 실제 시험을 구성하고 실행하는 측면에서 수험자가 접하게 되는 일반적인 맥락을 다루

도록 하여 시험 개발의 경제성과 시행의 효율성을 추구해야 한다고 주장하였다.

버틀러 외(2000)는 의사소통능력을 중심으로 개정된 토플의 개발에 따라 말하기 능력 검사가 고려해야 할 사항들을 종합적으로 검토하여 제시하였다. 이 연구는 토플이 실제적인 의사소통능력 평가가 되기 위해서는 말하기 시험을 토플과 통합하고, 실제 의사소통의 특징을 반영할 수 있는 구성형 문항(constructed-response item)을 사용해야 한다는 주장을 담고 있다.

## 2) 토플의 학문 목적 말하기 시험의 구성과 특징

토플의 학문 목적 말하기 시험(TOEFL® Academic Speaking Test, TAST)은 토플이 2005년에 인터넷 기반 시험을 실시하게 되면서 학문 목적 평가 영역으로 포함되었다. 토플의 학문 목적 말하기 시험의 목적은 미국 대학이나 대학원에서 수학하기 위한 적절한 수준의 구어 의사소통 기술을 갖추고 있는지를 확인하기 위한 것이며, 또한 이를 바탕으로 여러 종류의 수준별 영어 강좌 배치에 대한 결정에 이용될 수 있다(Xi & Mollaun, 2006: 1). TAST는 기존의 TSE와 여러 측면에서 다른데, 기존의 테이프 녹음기 대신 컴퓨터를 활용하고 학문적 맥락에서의 의사소통능력을 측정하기 위해 과제 구성과 채점 방식을 바꾼 것이 그 핵심이다.

TAST는 모두 여섯 과제로 구성되어 있으며, 그 중 두 개는 말하기 기능 독립형으로서 수험자의 개인적인 정보나 친숙한 화제에 대한 의견을 말하는 것이며, 나머지 네 개는 기능 통합형으로서 듣기와 읽기가 통합된 과제가 두 개, 그리고 나머지 두 개는 읽기와 듣기, 말하기를 모두 통합한 과제이다. 기능 통합적인 과제의 내용은

기본적인 대학 생활의 경험을 나타낼 만한 것들이나 교양 수준의 학문적인 지식과 관련된 것들이다.

TAST는 개정된 TSE와 같이 총괄적인 채점 방식을 사용하지만, 기존의 포괄적인 척도를 전반적인 과제 수행(general description)을 중심으로 전달(delivery), 언어 사용(language use), 화제 전개(topic development)에 따라 각 수준별로 채점하도록 한다는 면에서 큰 변화가 있었으며, 이는 분석적 채점과 총괄적 채점의 장점을 절충한 형태라고 볼 수 있다. TAST는 채점 기준을 기능 독립형 과제와 통합형 과제에 따라 각각 구성하였다.

## 2. 아이엘츠(IELTS)<sup>6)</sup>

아이엘츠(the International English Language Testing System, IELTS<sup>TM</sup>)는 영국 케임브리지 대학의 평가 연구 기관인 UCLES(University of Cambridge Local Examinations Syndicate)가 개발하고, 영국 문화원(the British Council)이 시행 및 관리를 담당하는 시험이다. 아이엘츠는 1989년부터 지금까지 토플과 함께 대표적인 영어 능력 평가 도구로 널리 사용되고 있다. 아이엘츠의 말하기 시험은 수험자가 면접관(interlocutor)을 마주하며 제시된 과제를 수행하는 직접식 말하기 평가라는 점이 특징이다.

---

6) 아이엘츠의 말하기 시험은 학문 목적 모듈과 일반 목적 모듈 모두 응시해야 하기 때문에 대학에서의 언어 사용 맥락을 반영한 토플 TAST에 비해 과제 맥락이 일반적이지만, 독화에서부터 토론 등 다양한 범위와 수준을 포함하고 있어 대학 입학 가능성 확인을 목적으로 하는 학문 목적 응시자의 요구를 반영하고 있다. 기타 유럽의 학문 목적 외국어 능력 시험(DALF, TestDaF 등)은 토플과 아이엘츠에 비해 개발 역사가 짧고, 주로 전문적인 학술 능력에 초점을 두고 있어 본 연구의 논의에서 제외하였다.

## 1) 아이엘츠의 말하기 시험 개발 과정

아이엘츠의 말하기 시험은 영국의 초기 영어 능력 평가 도구였던 EPTB(English Proficiency Test Battery)와 그 뒤를 이은 ELTS(English Language Test Service)의 개발 및 타당화 연구를 통해 그 개발 양상을 살펴볼 수 있다(Davies, 2008).

영연방에서의 유학생 선발과 관련하여 공식적으로 처음 개발된 영어 평가 도구인 EPTB는 영국의 고등교육기관으로 유학을 오는 비영어권 학생이 급증한 1960년대에 도입되었다. 영국 문화원과 버밍햄 대학(University of Birmingham)은 1963년부터 65년까지 공동으로 EPTB를 개발하였으며, 이후 1980년까지 영국 문화원을 통해 국제적으로 사용되었다.

EPTB는 200개 이상의 듣기와 읽기 문제로 구성이 되었으며, 말하기나 쓰기에 대한 평가가 직접적으로 이루어지지 않았다. 대신 말하기 능력에 대한 평가는 듣기 평가에서의 발음 평가(음소구별, 억양, 강세 등)로 대체하였다.<sup>7)</sup>

EPTB에 대한 반성적 연구를 바탕으로 개발된 ELTS는 영국 문화원과 UCLES가 협의체를 구성하여 함께 개발하였다. ELTS는 1975년부터 준비가 시작되어 1980년에 첫 선을 보였는데, 특수 목적 영어(ESP)의 성격을 가지고 있지만, 개발 과정에서 기존의 EPTB가 학문적인 측면의 영어 기술(skill)에 초점을 둔 것을 확장하여 일반적인 영어 기술에 대한 요구를 포함하였다는 것이 특징이다. ELTS는 일반적 모듈(General module)과 학문적 모듈(Manual module)로

7) EPTB는 미국의 초기 토플과 비슷한 시기에 시행된 시험이면서도 평가 도구 구성 측면에서 언어적 지식만이 아닌 언어 활동과 관련된 평가 과제를 포함하고 있었다.(Davies, 2008: 20)



구성되었는데, 학문적 모듈은 먼비(1978)를 바탕으로 학습자의 요구를 반영하여 주요 선호 전공인 생활 과학(Life Science), 사회 과학(Social Science), 물리학(Physical Science), 공학(Technology), 의학(Medicine), 교양 학문(general academic)으로 선택적으로 구성하였다. ELTS 말하기 시험은 인터뷰 방식으로 이루어졌으며, 수험자가 지금 하는 일이나 미래 학업 계획에 대한 내용과 자신이 선택한 특정 전공 영역에 대한 읽기 자료(booklet) 중 면접관이 선택한 화제에 대한 토의(discussion)로 이루어졌다.

ELTS는 개발 당시 점수의 해석과 관련된 타당도 문제와 이전 시험(EPTB)에 비하여 실제적으로 향상된 것이 불분명하고, 수험자가 선택하는 전공 영역의 적합성 문제 등이 한계로 지적되면서 타당화 연구가 이루어졌다(Davies, 2008: 40). ELTS의 타당화 관련 연구는 예측타당도 연구와 공인타당도 연구, 그리고 구인타당도 연구와 유사 검사 도구의 수준 비교 연구로 구성되었다.

예측타당도 연구에서는 ELTS 응시자의 영어 수준이 실제 해당 전공 영역의 학업 성적에 중요한 영향을 미칠 것이라는 가정 하에 연구를 진행하였는데, 분석 결과에서 신입생의 경우에는 ELTS와 학업 성적의 상관성이 유의하지 않은 것으로 나타났다. 그러나 대학원생의 경우 예측타당도가 있는 것으로 나타났으며, 특정 평가 모듈(일반 모듈과 인터뷰)의 경우에는 전반적인 시험의 성취도를 예측할 수 있는 것으로 나타났다. ELTS와 EPTB, ELBA(English Language Battery)를 비교하는 공인타당도 연구에서는 ELTS가 전문적인 영역의 지식과 언어적인 능력을 동시에 평가한다는 점에서 가치가 있지만, 이러한 방식의 타당성에 대한 고려가 제대로 이루어지지 않았고, 문항의 수나 배열, 시간 등의 문제가 결합되어 실용적인 측면에서도 문제가 있다는 점을 확인하였다. 햄프-라이언즈

(Hamp-Lyons, 1988: 13)는 이러한 현상이 평가 도구로서 선발과 진단의 목적을 동시에 가지고 있기 때문에 나타나는 혼란이라고 보았으며, 헤닝(1988)은 ELTS의 문항이 언어 사용보다 시험 보는 기술이나 상식으로 해결 가능한 면이 있어 안면 타당도와 예측타당도가 확인된 말하기 시험을 제외한 나머지 영역에 대한 수정이 필요하다고 하였다.

80년대에 들어 ELTS 관리 위원회는 ELTS의 타당화 연구와 기타 관련 연구들을 바탕으로 ELTS에 대한 상세한 검토를 실시하고, ELTS 수요 집단에 대한 적합성을 검토하여 내외적 타당성을 확인하였다. 또한 특수 학문 목적 영어 능력이라 할 수 있는 전공별 모듈 방식의 타당성을 확인하기 위해 전공 선택에 따른 응시자 집단의 결과를  $t$ -검정을 통해 분석하여 모듈별 시험의 한계를 확인하였다(Alderson & Urquhart, 1983, 1985).

ELTS 위원회는 1987년부터 도구 개선 작업을 시작하며 ELTS와 관련된 수요자들로부터 ELTS에 대한 인식과 구조의 문제, 그리고 시행 지역 확대에 관한 의견을 수집하였다. 그리고 조사 결과를 바탕으로 새로운 ELTS에 대한 전반적인 계획을 수립하였다. 먼저 말하기 시험을 학문적 모듈에서 일반 모듈로 이동시켜 전공이나 관심 분야의 차이로 인해 인터뷰가 제대로 이루어지지 못하는 문제를 해소하고, 모든 응시자의 말하기 능력을 측정하도록 하였다. 새로운 ELTS의 개발진은 평가 요소에 대한 점검을 위해 전문가들을 대상으로 평가하고자 하는 학문적 능력의 구인을 탐색하였는데, 전공 영역에 따라 텍스트 선정이 쉽지 않고, 진점수(true score)의 해석이 평가 내용에 대한 친숙성의 영향을 받고 있어 판단이 어렵다는 점을 확인하였다.

새로운 ELTS의 말하기 평가 모듈은 ‘도입(introduction)-담화 확

장(extended discourse)-발화 유도(elicitation)-추측과 태도(speculation and attitudes)-결론(conclusion)'의 5단계로 구성되어 시간 통제 하에 진행하였다. 모든 인터뷰 내용은 녹음하였으며, 평가자가 응시자의 정보를 사전에 보고 인터뷰에 참고하였다.

80년대 말 이루어진 ELTS 개정을 바탕으로 개발된 아이엘츠는 89년부터 사용되기 시작하였는데, 효율성을 고려하여 모듈을 학문적 모듈과 일반적 모듈로 간소화 하여 예측타당도를 높이고자 하였다. 아이엘츠는 89년부터 93년까지 초기 시행을 한 후에, 아이엘츠 자문 위원회(IELTS Advisory Committee)의 수정 지침을 통해 최신 측정 이론을 반영하여 평가 내용과 과정이 정교화되었으며, 95년부터 개정되어 시행되었다. 개정된 아이엘츠는 듣기, 읽기, 언어 사용, 쓰기, 말하기로 구성되었으며, 학문 목적 모듈(IELTS Academic)과 일반 목적 모듈(IELTS General Training)에 따라 특수하게 구성된 읽기와 쓰기 시험을 보고, 듣기와 말하기 시험은 공통된 시험에 응시하도록 하고 있다.

## 2) 아이엘츠 말하기 시험의 구성과 특징

아이엘츠 말하기 시험은 수험자 관련 질문과 선택한 주제에 대한 면담, 추상적인 개념이나 문제에 대한 의견 말하기의 세 부분으로 구성되며, 시험시간은 약 11~14분이 소요된다.<sup>8)</sup> <표 1>은 현재 시행되고 있는 아이엘츠 말하기 시험의 평가 기능과 과제 유형, 평가 목표, 시간에 따라 정리한 것이다.

8) 2016년 현재 시행되고 있는 아이엘츠 시험에 관한 자세한 사항은 아이엘츠 홈페이지 [www.ielts.org](http://www.ielts.org)에서 확인할 수 있다.

&lt;표 1&gt; 아이엘츠의 말하기 시험 구성

	Part1	Part2	Part3
기능	소개 및 인터뷰 (introduction and interview)	긴 말차례 (long turn)	토의 (discussion)
과제 유형	집, 가족, 하는 일, 공부, 관심사 등 익숙한 화제 에 관한 일상적인 질문 에 답하기	특정 화제에 대한 개인 적인 경험이나 생각에 관한 구체적으로 설명 하기	특정 화제에 대한 일반 적이고 추상화된 문제 에 관해 의견을 말하기
평가 목표	일상적인 화제와 평범 한 경험이나 상황에 대 한 의견이나 정보 의사 소통 능력	주어진 화제에 대한 담 화를 구성하고, 이를 적절한 표현을 바탕으 로 조직하여 길게 말하 는 능력	주어진 문제에 대한 분 석과 논의, 추측을 통 해 의견을 표현하고 정 당화하는 능력
시간	4-5분	3-4분	4-5분

아이엘츠 말하기 시험의 채점은 유창성과 응집성(fluency and coherence), 발음(pronunciation), 어휘(lexical resource), 문법(grammar range and accuracy)에 대한 0~9 수준으로 구성된 채점 기준표에 따라 이루어지며, 토플과 마찬가지로 전문적인 훈련과 인증을 받은 시험관들이 담당한다. 채점 결과는 각 채점 기준에 대한 등급 점수(band score)로 부여되는데, 이 때 점수는 0~9 등급을 .5로 각각 나누어 부여하며, 네 채점 영역 점수의 평균으로 최종 말하기 점수가 결정된다.

### 3. 학문 목적 말하기 평가 도구 개발의 시시점

이상에서 제시한 토플과 아이엘츠에서의 말하기 평가 개선 및 타당화 과정을 바탕으로 학문 목적 말하기 평가 도구 개발과 관련한

시사점을 제시하면 다음과 같다.

첫째, 학문 목적 말하기 평가 도구는 최신의 언어 능력 및 평가 이론을 바탕으로 설계 및 개발되며, 계속되는 이론적인 발전을 바탕으로 개선 및 타당화하여야 한다. 토플과 아이엘츠의 개발과 개정 및 타당화 과정을 지탱한 것은 그 당시에 제기된 언어 교육 및 평가 관련 이론들이었다. 토플과 아이엘츠의 말하기 시험은 초기에는 지나치게 언어 지식 중심으로만 평가하거나 선다형 문제로 정확한 발음을 알고 있는지를 측정하려고 하는 등 지식 및 문법 중심의 언어 능력관을 반영하고 있었다. 그러나 현재는 의사소통능력을 중심으로 문제 해결을 유도하는 과제를 통해 실제적이고 직접적인 말하기 능력 평가를 실시할 수 있게 되었다. 또한 이러한 접근의 배경에는 심리측정학적인 이론의 발전으로 다국면 라쉬 모형이나 일반화 가능성도, 구조방정식 등의 검사 도구의 안정성과 타당성을 확인할 수 있는 최신 이론이 반영될 수 있다.

둘째, 학문 목적 말하기 평가 도구 개발은 평가 도구의 실제적인 사용을 고려하여 효율성과 경제성을 확보하는 방향으로 이루어져야 한다. TSE가 처음 개발된 후 평가 과제의 수가 줄어들고, 채점 기준이 압축되고, 적도를 조정하였던 이유는 평가 도구의 효율성과 경제성을 높이고자 하는 것이었다. 과제의 숫자가 필요 이상으로 많을 때, 응시자에게는 피로도 및 시험 준비로 인한 부담이 더해질 수 있으며, 관리자도 긴 시간 동안 시험장을 관리해야 하고, 채점의 소요 시간도 증가하는 문제가 발생한다. 아이엘츠 말하기 시험도 문항의 수나 배열, 시간 배분 등의 측면에서 시험의 실용성과 효율성을 확보하기 위해 간소화하는 과정을 거쳤는데, 이러한 변화는 기술의 발전으로 자료 수집과 보관, 결과 분석 등이 수월하고 정교해지면서 가능해진 것이었음을 주목해야 한다.

셋째, 학문 목적 말하기 평가 도구의 개발은 실제성 확보를 위한 안면타당도 및 예측타당도 연구가 뒷받침되어야 한다. 특수 목적 언어 평가로서 학문 목적 말하기 평가가 갖추어야 할 가장 기본적인 점은 학문 목적 학습자와 교사들로부터 확보되는 안면 타당도이다. 단순하게 평가 도구의 난이도나 편의성 측면만이 아니라 실제 측정하고자 하는 학문 목적 말하기 능력이 무엇이며, 학습 현장에서 이와 같은 능력이 실제로 요구되며, 교사의 입장에서는 어떤 인식을 갖고 있는지를 종합적으로 살펴서 평가 도구의 효용성을 높여야 한다. 또한 토플과 아이엘츠 개발 과정에서 평가 관련 주체들을 참여시켰던 것과 같이 예측타당도 연구를 통해 평가 도구의 예측 가능성을 실제적으로 확인하여야 한다.

넷째, 대규모 학문 목적 말하기 평가의 개발 과정에서는 평가 결과 활용을 고려하여 이를 위한 명확한 증거가 수집될 수 있도록 구인타당도와 공인타당도 연구가 이루어져야 한다. 토플과 아이엘츠 연구들에서 강조된 점은 의사소통의 실제적인 특성을 반영하고 있는가의 문제이다. 구인타당도는 말하기 능력의 구인에 대한 이론적 검토와 실제적인 언어 사용에 대한 고려를 통해 확인되며, 평가 도구 구성의 핵심인 과제와 채점 기준 등과 직접적으로 관련된 정보이다. 또한 공인타당도는 학문 목적 말하기 평가의 응시자가 다른 평가 도구 응시 결과와 비교하였을 때의 설명력의 차이를 알아보는 것으로서, 검사 결과 활용 시 중요한 정보가 된다.

다섯째, 학문 목적 말하기 평가 도구 개발은 실용적인 측면에서 동시대적인 기술을 적용한 평가 도구 개발 및 활용에 대한 접근이 필요하다. 토플과 아이엘츠의 말하기 시험은 개발 과정에서 종이 사용의 유무, 녹음기 사용, 컴퓨터 및 멀티미디어 활용 등의 면에서 기술 발전을 적용하려 하고 있었다. 말하기 평가는 장르의 특성으로

인해 종이나 녹음기 같이 고정된 매체를 활용한 평가에 한계가 있다. 실용적이고 경제적인 측면을 고려할 뿐만 아니라 수험자의 입장에서 동시대적인 기술이 적용된 도구를 활용한 말하기 평가를 실시하는 것은 자연스럽고 당연한 일이다.

국외 학문 목적 말하기 평가 도구 개발 과정에서 나타난 실용성·경제성·실제성 및 이론과 기술의 동시대성은 학문 목적 한국어 말하기 평가 도구 개발에서도 중요하게 고려되어야 할 사항이다.

## IV. 학문 목적 한국어 말하기 평가 도구 개발의 쟁점

### 1. 학문 목적 말하기 평가의 개념과 구인

언어 평가를 통해 평가하는 언어가 무엇인지를 결정하는 것은 평가의 목적과 관련되어 있으며, 이는 평가 도구 개발의 핵심적인 문제이다. 학문 목적 한국어 말하기 평가의 목적은 응시자 또는 평가 결과 사용자의 필요와 관련되어 있다. 학문 목적 말하기 능력은 유학생의 입학 시점을 기준으로 입학 전의 입학 자격 조건으로서 갖추어야 할 학문 목적 말하기 능력과 입학 후에 실제 학문 목적 한국어 말하기 과제 해결 능력과 관련된 것으로 나누어 볼 수 있다. 지금까지의 학문 목적 말하기 평가 연구들에서는 학문 목적 말하기 평가 능력이 일반 목적 한국어 능력의 상위 수준과 강의 시간에 해결해야 하는 학문적 문제 해결 능력이 혼재되어 제시되어 왔다. 이는 학문 목적 한국어 말하기 평가 대상에게 혼란을 일으킬 수 있다.

&lt;표 2&gt; 학문 목적 말하기 평가의 구인과 대상

연구자	구인(평가 범주, 채점 기준)	대상
한상미 (2009)	언어적 능력(음운-담화, 화용적 능력, 사회언어적 능력)	한국 대학 진학 목표 유학생
박광진 (2010)	내용 구성, 과제수행력, 어휘-문법, 발음과 억양, 상호작용	대학 수학 과정에서 요구되는 말하기 활동 과정 수행 역량 확인
권수현 (2013)	과제 수행 능력, 논리-사고력, 언어적 능력, 사회언어학적 능력, 담화 능력, 전략 사용 능력	진학 목적 학습자 및 학업 부진 유학생
심재경 (2014)	문법적 능력, 담화적 능력, 사회언어학적 능력, 전략적 능력, 과제 수행 능력, 논리-사고력	한국의 대학(원)의 생활 및 학업과 관련된 한국어
김상경 (2015)	문법적 능력, 담화적 능력, 전략적 능력, 다회 언어학적 능력, 과제 수행 능력	유학생 및 유학 목적 한국어 학습자
박선운 (2015)	언어적 능력(발음, 어휘, 문법, 담화적 요소), 화용적 능력(담화 구성 능력, 사회-문화적 지식), 인지적 능력(논리-사고력, 과제수행력)	대학 진학 목적 학습자
권순용 (2015)	내용 구성, 과제수행력, 어휘-문법, 발음과 억양, 상호작용	진학 목적 학문 목적 한국어 말하기 수업

용어로 인한 혼란은 구인에 대한 논의에서도 나타나는데, 구인은 채점이나 평가 등급 기술과 관련된 영역으로서 다루어지고 있다. 학문 목적 한국어 말하기 평가의 개념과 구인을 파악하기 위해서는 포괄적 의미에서의 학문 목적 말하기 능력과 구체적인 목적에 따라 선발과 진학을 위한 학문 목적 말하기 능력, 학업 수행의 수월성을 위한 학문 목적 말하기 능력으로 구분하여야 하며, 각각의 맥락에서 요구되는 언어를 상정하여 구인을 기술해야 한다. 기존의 학문 목적 한국어 말하기 평가 연구에서 제시된 구인과 평가 대상은 <표 2>와 같다.

학문 목적 한국어 말하기 평가 연구에서 제시한 구인 중 논리-사고력과 같이 학문 목적 평가의 특성이 두드러진 구인이 제시되기도



하였으나, 대체로는 기존의 말하기 평가 연구에서 제시하는 구인과 유사하며, 평가 대상별로 두드러진 특징을 체계적으로 제시하지 않았다.

구인은 본질적으로 평가할 대상 및 내용을 결정하는 기준이 되는 것으로서, 시험 점수를 통해 추론되는 수험자의 능력에 대한 이론적 근거가 되는 것이다. 펄처(Fulcher, 2003: 18)는 조작적인 측면에서 구인이 측정 가능하며, 점수화할 수 있는 대상과 관련된 것이라고 하였다. 학문 목적 말하기 평가는 응시자와 결과 사용자의 목적에 따라 구성되며, 이 때 구인은 평가 도구를 통해 측정 가능한 것으로서 평가 대상자의 과제 수행을 통해 확인할 수 있으며 점수화할 수 있는 것이어야 한다.

## 2. 학문 목적 말하기 평가 도구의 개발과 과제

학문 목적 한국어 말하기 평가에 관한 연구들에서 과제 선정 및 도구 개발에 관한 논의가 주로 이루어졌던 배경에는 공인된 한국어 말하기 능력 평가 도구가 없으며, 또한 유학생의 선발 및 수학 가능성과 관련하여 통용되는 TOPIK에 말하기 시험이 포함되어 있지 않았기 때문인 것으로 보인다.

루오마(2004)는 평가 도구 개발은 평가 과정 전반에서 하나의 과정에 해당하는 것이며, 선후 관계에 따라 연결된 평가 대상의 필요성(목적) 확인이나 품질 관리 과정을 거친 후에 평가 시행과 채점, 그리고 평가 도구 개선 등에 이르는 과정과의 연관성을 고려해야 한다고 하였다. 또한 바흐만과 파머(2010: 33)는 언어 평가에서 가장 우선되는 목적이 수험자에 대한 의사 결정을 돕는 정보를 수집하는 것이라고 하였는데, 이는 평가의 계획 시 평가의 마무리 단계에서 성취해야 하는 목표와 관련된 정보가 수집되도록 해야 한다는 의미

이다.9) 평가 개발과 관련하여 학문 목적 한국어 말하기 평가 연구를 살펴보면, 주로 평가 도구의 작동에 관심을 두고 실제 평가 도구와 과제 개발에 집중한다는 것을 알 수 있다

지금까지 학문 목적 말하기 평가 연구에서 제시된 과제의 유형은 <표 3>과 같다.

<표 3> 학문 목적 말하기 평가의 과제

연구자	과제(기술, 평가 항목)
한상미 (2009)	초급-구술면접 반기, 학교 시설물 이용하기 중급-부탁하기, 상담하기, 공식적인 공지하기 고급-강의 수강하기(내용 확인, 보충 설명 요청, 토론하기(요약, 동의, 반박, 이견 말하기))
박광진 (2010)	사진 설명·묘사하기 / 신문 기사 요약하기 / 찬반 의견 제시하기
권수현 (2013)	특정 주제에 대해 의견을 피력하기 특정 목적 수행 상황에서 알맞은 발화하기 특정 상황에 대해 이야기 구성 및 발화하기 특정 상황 가정 후 역할 수행하기
심재경 (2014)	개인적인 경험에 대한 질문 답하기 상황이나 자료에 대한 의견 표현 및 주장하기 상황이나 문제에 대한 의견 표현 및 주장하기 글을 읽거나 대화를 요약하여 설명하기
김상경 (2015)	낭독하기 / 강의 듣고 말하기 / 토론에서 의견 말하기 / 자료보고 발표하기
박선윤 (2015)	대학 생활 이야기하기 / 요약하여 설명하기 / 자신의 의견 말하기 / 발표하기
권순용 (2015)	인터뷰 / 역할놀이 / 토의하기 / 발표하기 / 그림이나 자료 활용하여 말하기

9) 바흐만과 파머(2010: 143-145)는 평가 개발과 사용 과정을 ‘초기 계획(initial planning)-설계(design)-조작화(operationalization)-예비 검사(trialing)-평가 사용(assessment use)’의 연속적인 다섯 단계로 나타냈다. 여기서 첫 단계인 초기 계획은 평가를 통해 의도하는 결과와 의사 결정, 결정을 내리게 하는 정보, 정보를 얻기 위한 평가, 도구 개발의 필요성과 관련 자원 검토 등으로 이루어진다.

학문 목적 한국어 말하기 평가에서 과제는 과제 혹은 평가 항목, 기능 등으로 제시되어 왔으며, 강의 상황을 중심으로 한 말하기 과제가 강조되어 나타났다. 이는 학문 목적 한국어 말하기 과제의 맥락으로 대학(원)의 강의에 주목한 것인데, 대학 생활 경험이 없거나, 유학을 준비하는 과정의 학습자에게는 과제의 비친숙성으로 인하여 난이도가 높은 과제로 인식될 가능성이 크다.

학문 목적 한국어 말하기 평가의 과제는 실제 상황 맥락에서 요구되는 과제와 관련된 것이어야 하며, 응시 목적에 따라 과제의 세부적인 요구나 조건이 달라질 수 있다. 언어 평가에서 과제를 특정 상황에서 특정한 목적이나 목표를 성취하기 위해 언어를 사용하는 활동으로 정의한 바흐만과 파머(1996: 44)는 ‘특정’이라는 점에 주목하여 개별 평가 문제들의 비맥락성과 실제성 문제가 과제를 통해 해소할 수 있다고 보았다. 학문 목적 한국어 말하기 평가에서 과제를 문항으로 구성하기 위해서는 과제로서의 구체적인 특성을 잘 갖추어야 하며, 이를 위해 과제 작성 과정에서 과제의 특성을 구체적으로 기술하고 실제 문항으로 구성해야 한다. 펠처(2003: 57)는 말하기 평가의 과제 틀로서 과제 특성(task orientation), 상호 작용성(interactional relationship), 목적 지향성(goal orientation), 대화자 지위와 친숙성(interlocutor status and familiarity), 화제(topics), 상황(situation)을 제시하였다. 이러한 말하기 과제 기술 틀은 과제에 대한 정보를 구체적으로 확인할 수 있게 한다는 점에서 평가 목적에 따라 과제 선정 과정에서 유용하게 활용될 수 있는 정보들을 담을 수 있다.

### 3. 학문 목적 말하기 평가 방법과 타당화 연구

평가 도구에 대한 타당화는 평가 도구가 평가 대상을 타당하게 평가하는지에 관한 증거들을 수집하는 것이며, 내용타당도와 구인 타당도가 말하기 평가 도구 개발의 기초이다. 평가의 타당화와 관련하여 기존의 연구들은 이론적 타당화에 치중하며, 주로 국외 사례를 통한 접근을 해 왔으며, 김상경·박동호(2014)와 김상경(2015)에서 시도한 타당화 시도가 유일하다. 타당화는 개발 과정에서부터 이를 고려하여 관련된 영역들과의 관계 속에서 타당도 논증(validity argument)<sup>10)</sup> 과정으로 이루어지며, 이는 평가의 세부 실행 계획 등을 담은 평가 설명서 작성에도 반영된다.

말하기 평가는 평가 방법에 따라 일대일 인터뷰나 컴퓨터(인터넷) 기반 평가로 이루어지는 것이 보통인데, 한국어교육에서 가장 일반적인 것은 일대일 인터뷰이다. 지금까지 학문 목적 한국어 평가 연구에서도 컴퓨터 기반 말하기 평가를 실시한 김상경(2015)을 제외하고는 모두 일대일 인터뷰 평가를 전제하고 있었다. 말하기 평가 방법은 직접식(direct method)과 간접식(indirect method), 그리고 반직접식(semi-direct method)으로 나뉘며, 직접식의 가장 대표적인 형태가 일대일 인터뷰이다. 간접식은 초기 말하기 평가가 어휘 검사나 쓰기 검사로 이루어졌던 것과 관련되며, 반직접식은 평가자와 피평가자가 분리된 환경에서 테이프나 컴퓨터를 통해 녹음된 것을 듣고 평가하는 것으로서, iBT TOEFL이 대표적이다. 현재 보편화된 직접식과 반직접식 말하기 평가는 각각 장단점을 갖고 있는데, 직접

10) 크론바흐(Cronbach, 1988: 3)는 타당화 논거에 대해 타당화 연구(validation research)를 대신하는 것으로서 최종적인 평가(evaluation)에 대한 논거(argument)의 논리성과 관련된 개념으로 이를 제시하였다.

식의 경우에는 피평가자에 대한 안면타당도(face validity)를 확보할 수 있지만, 피평가자에게 균일하지 않은 평가 조건이 제공될 수 있다. 또한 평가자는 대화자(interlocutor)와 채점자(rater)의 역할을 동시에 수행하는 부담이 있으며, 이는 평가의 신뢰도에도 부정적인 영향을 미칠 수 있으며, 비용이 많이 든다는 문제도 있다. 반직접식 평가의 경우에는 피평가자에게 균일한 평가 환경을 제공할 수 있고, 평가자 영향이 없으며, 대규모로 시행 시에 비용이 절약된다는 장점이 있지만, 녹음기나 컴퓨터 등의 기기에 익숙하지 않은 수험자들에게 불리할 수 있으며, 실제 사람에게 이야기 하지 않기 때문에 상호작용 능력을 평가하기 어렵다는 한계가 있다. 학문 목적 한국어 말하기 평가에서는 먼저 평가 대상 언어의 특징에 따라 평가 방법을 결정하여야 하며, 평가 참여자의 규모와 평가 시행 환경 등도 고려하여야 한다. 또한 학문 목적 말하기 평가 도구의 타당한 개발을 위해서는 평가 도구의 영향력과 실효성을 고려하여 이론타당도 확보와 내용타당도와 구인타당도, 그리고 공인타당도(concurrent validity)와 예측타당도(predictive validity), 채점타당도(scoring validity) 등에 관한 연구도 이루어져야 할 것이다.

#### 4. 채점과 척도 개발

학문 목적 한국어 말하기 평가 연구에서 채점(rating/scoring)과 채점을 위한 척도(rating scale) 연구는 기존 평가 도구의 모형을 바탕으로 내용을 재구성하여 제시하거나 요구 조사의 내용을 반영하여 채점 기준표를 제시하는 것으로 이루어졌다. 주로 학문 목적 말하기 과제의 특성을 고려하고, 평가 기준과 점수 배점(가중치)을 달리하는 모습이 여러 연구에서 나타났다(한상미, 2009; 박광진, 2010;

김상경, 2015; 박선운, 2015). 김상경(2015)에서는 이론적으로 구성한 채점 척도를 활용하여 실제 평가를 시행하고, 채점자 및 문항 양호도를 분석하였다. 한국어 말하기 평가 연구에서는 이항(2012, 2013)이 한국어 발음 채점에 관한 양적 연구를 수행하였는데, 음성 자료 수집과 분석이 어렵고, 자료 해석을 위한 관련 연구의 부족으로 그 수가 매우 적다.

직접적으로 학문 목적 한국어 평가 척도를 개발한 최정순(2006)에서는 학문 목적 언어 능력 평가의 척도가 보편적인 적용을 위해 언어권별로 다를 수 없으며, 유럽공통참조기준(CEFR)의 사례를 바탕으로 학문 목적 한국어에 대한 평가 등급을 구성하였다.<sup>11)</sup> 한상미(2009)에서는 평가 범주로 제시한 언어적·화용적·사회언어적 능력을 기준으로 각각의 세부 범주에 대한 등급별 숙달도 수준을 작성하고, 이를 바탕으로 5점 척도에 따라 채점을 하게 하였는데, 박선운(2015)과 함께 문항의 수준에 따라 평가 범주의 비중을 달리하였다는 것이 특징이었다. 이러한 논의에서 제시한 채점 기준 들은 자료 분석과 연결되지 않고 있어 아쉬움이 있다.

박광진(2010)과 김상경(2015)은 과제별로 채점 기준이 다른 혼합형 채점 방식을 제안하였는데, 김상경(2015)은 실제 시험을 실시하고 전문가 채점을 실시하였다는 것이 특징이다. 이 연구에서 채점은 채점자 5인이 75명의 응시자를 부분적으로 공유하는 부분 연결 방식으로 이루어졌으며, 그 결과는 다국면 라쉬 모형을 적용한 FACETS을 통해 분석되었다.<sup>12)</sup> 평가에서 채점은 평가의 결과를 직

11) 최정순(2006)은 언어 수준에 따라 '기본 한국어 학업능력평가'와 '한국어 학업능력평가'로 분리하고, 평가 등급을 '우수 합격-합격-근소 불합격-불합격'의 4단계로 나누며, '우수 합격'은 장학금 지급 기준으로, '근소 불합격'은 조건부 입학이나 전공과목 이수 기준으로 활용하는 제안을 하였다.

접적으로 결정하는 행위로서 사람을 통해 이뤄지는 과정이기 때문에 이에 대한 전문적인 훈련과 체계적인 설계가 뒷받침된 말하기 평가 도구 설계가 필요하다.

## 5. 교육적 환류와 영향력

학문 목적 한국어 말하기 평가를 통해 얻어진 결과는 도구 개발의 목적과 관련되어 교육 및 기타 여러 측면에 영향력(impact)을 나타낼 수 있다. 특히 교육적 맥락에서는 평가의 결과가 수험자가 점수를 얻는 것뿐만 아니라 교육 전반의 개선에도 영향을 미치는 환류(wash-back)를 일으킬 수 있다. 학문 목적 한국어 말하기 평가에서 교육적 환류와 관련된 논의는 드물지만 김상경(2015)에서 부분적으로 찾아볼 수 있다. 이 연구에서는 채점 결과를 바탕으로 대규모 평가 도구로서의 타당성을 확보하기 위해 일반화가능도 연구를 실시하여 평가의 표준화된 사용을 위한 채점 조건을 제시하였다는 점에서 평가 도구 사용과 결과를 활용에 대한 잠재적이지만 유용한 정보를 제공하고 있다.

학문 목적 한국어 말하기 평가의 결과가 교육적인 의미를 고려하여 활용되고, 평가에 관련된 주체가 이를 신뢰하고, 이를 여러 환경에서 긍정적인 방향으로 활용할 수 있을 때 평가 도구의 가치를 나타내는 책무성(accountability)과 공정성(fairness)이 강화될 수 있다. 발콤과 코자(Balcom & Kozar, 1994)는 유학생 배치 고사인

---

12) FACETS은 수험자들의 전체적인 분포와 시험 유형별 난이도, 채점자들의 엄격성(severity), 과제별 난이도와 점수 분포 등에 관한 상대적 거리(logit) 정보를 제공하는데, 이러한 심리측정학적 정보는 평가 결과 보고와 해석, 채점 및 평가 도구 개선 등과 관련하여 유용한 정보이다.

OTESL(the Ontario Test of English as a Second Language)에 대해 소개하고 있는데, 입학 여부를 결정할 때 사용되는 토플과 달리 이 시험은 석·박사 과정의 대학원 유학생에게 학문 목적 프로그램의 이수를 결정한다. 이 평가를 통해 제공되는 학습 프로그램을 결정하는 진단적 정보는 직접적으로 교수·학습 내용 개발에 있어서 실제적인 학습자 담화와 관련된 정보를 제공할 수 있기 때문에 교육적 환류 측면에서 활용도가 높을 수밖에 없다. 김혜영(2010: 79)은 한국의 대학들이 학문 목적 학습자들을 위한 한국어 수업을 따로 마련하기보다 일상생활의 의사소통에 초점을 둔 일반 목적 한국어 교육을 제공하고 있어 한국어 고급을 마치고도 유학생들이 많은 어려움을 느낄 수밖에 없고, 이러한 원인으로 주제와 내용의 차별성과 전문 용어의 문제를 지적한 바 있다. 아직까지 공인된 학문 목적 한국어 말하기 평가 도구가 없기 때문에 교육적 환류와 영향에 대한 논의가 잘 이루어지지 않고 있지만, 앞으로의 말하기 평가 도구 개발 과정에서는 평가 시행 및 결과 활용과 관련하여 평가의 정의(justice)가 이루어지도록 영향력을 고려하여야 할 것이다.

## V. 학문 목적 한국어 말하기 평가 도구 개발의 과제

### 1. 평가 계획 단계: 응시 목적에 따른 요구 분석

학문 목적 한국어 말하기 평가를 개발하기 위해서는 평가 계획 단계에서 학문 목적 말하기 평가의 시행을 요구하는 이들의 목적을 고려하여야 하며, 이 때 목적에 따라 평가 대상 언어의 특성이 규명



될 수 있다. 채이프(Chafe, 1986: 26-27)는 학문적인 말하기(academic speaking)가 대화적인 말하기(conversational speaking)와 학문적 쓰기(academic writing)의 연속선상에 놓여 있으며, 학문적 말하기의 화자는 주로 독화(monologue) 중심의 교훈적이고 계획된 발화를 하고, 기본적인 말하기 장르적인 특성과 학문적 말하기에 대한 경험 및 자신의 성격과 정체성, 학습 목적 등에 영향을 받는다고 하였다. 학문 목적 한국어 말하기 평가에서 기초적인 학문적 말하기 능력을 검사하여 입학에 위한 자격 조건으로 사용할지 아니면 입학을 한 후에 구체적인 학문적 말하기 능력을 교육시킨 후에 학문적 과제 수행을 준비시키기 위해서 사용할지가 결정되고, 그에 따라 구인과 평가 과제, 도구 개발이 이루어져야 한다. 평가 계획을 위한 요구(needs)를 알아보기 위해서는 처음 학문 목적 한국어 말하기와 관련하여 시간에 따른 인식의 차이를 주목하여 볼 필요가 있으며, 이에 따라 교육 대상과 시점에 따른 평가의 목적을 상세화하여야 한다. 또한 현재 학문 목적 한국어 말하기 평가의 이해 관계자를 교실 말하기 평가 상황으로 한정하지 말고, 평가 규모에 따라 보이지 않는 이해 관계자가 될 수 있는 평가 도구 개발자 및 학교 관계자, 정책 입안자 등으로 확대하여 이들이 갖고 있는 기대와 인식을 조사하고, 이를 평가 목적 구성과 도구 개발에 반영하여 결과타당도(consequence validity)도 계획 단계부터 고려해야 할 것이다.

## 2. 평가 도구 개발 1단계: 구인타당도 확보

평가 도구의 핵심은 평가 대상을 관찰 및 측정할 수 있는 구인으로 파악해 내는 것이며, 말하기 평가 도구 개발에서도 구인 선정 및 타당성 확보가 실제적인 시작이다. 기존의 말하기 평가 구인 관련

연구들은 전문가 델파이 조사를 통해 구인을 탐색하거나(김정숙, 2014), 특정 구인만을 탐색하는 연구(지현숙, 2005; 강현주, 2013)가 이루어졌었다. 국내 영어교육에서는 김수경(2012)이 구인타당도 연구를 수행한 바 있는데, 낭독 과제 수행에서 나타난 수험자의 언어적 지식과 지문의 난이도 및 다른 평가 과제 점수와의 상관을 분석하여 구인 타당도를 알아보았다.

학문 목적 한국어 말하기 평가에서 구인타당도를 확보하기 위해서는 평가 목적에 따른 평가 대상 언어의 특징과 구체적인 능력 요소인 구인이 밝혀져야 한다. 민병곤 외(2014: 111)에서는 일반 성인 대상의 국어 말하기 평가 구인으로 내용 구성(내용 선정, 조직)과 표현 및 전달(표현, 전달)의 두 차원으로 나누고 중요도를 고려하여 그 비중을 60 대 40으로 제시하였다. 학문 목적 한국어 말하기 평가의 구인타당도를 확보하기 위해서는 기존의 말하기 평가 구인 관련 연구들의 결과와 실제 평가 수행 담화에 대한 분석, 그리고 해당 결과가 의미하는 바에 대한 다차원적인 분석이 이루어져야 한다. 타당도가 확보된 구인은 학문 목적 말하기 평가 구인으로서 채점 항목 또는 채점 기준표(scoring rubric) 기술의 핵심이 될 것이다.

### 3. 평가 도구 개발 2단계: 평가 설명서 진술

학문 목적 한국어 말하기 평가는 평가가 기본적으로 갖추어야 할 객관성을 지향해야 하며, 이를 위해 평가 도구의 개발 과정이 한 개인의 역할로 이루어지기 보다는 공동 개발과 전문가 검토 과정을 요한다. 평가를 실행할 때에도 수험자 영향과 검사 환경을 고려하여 균질한 조건에서 평가가 시행되도록 해야 한다. 말하기 평가는 다른 기능의 평가와 달리 직접적인 수행을 평가하기 때문에 검사 환경에

따라 여러 가지 고려해야 할 요소가 발생한다. 말하기 평가의 어려움에 관하여 루오마(2004: 1)는 평가자가 갖게 되는 인상(평가)에 여러 가지 영향이 미치는 상황에서, 평가 점수는 정확하고 적절하기를 기대하기 때문이라고 하였다. 따라서 말하기 평가가 결과를 사용자들이 기대하는 것과 같이 신뢰할 만한 평가로서 기능하기 위해서는 평가의 설명서가 제시되어야 한다.

평가 설명서는 시험의 전체적인 형식과 각종 틀, 구인, 과제, 예상 응답, 채점 등을 기술한 것으로서 바흐만과 파머(1996)는 이를 ‘청사진’(blueprints)이라고 하였다. 펠처(2003: 117-118)는 컴퓨터 기반 검사에서 기록해야 하는 것으로 평가 도구 개발자의 자질과 자격, 응시 대상의 특성, 평가의 사용 목적, 구인에 대한 정의, 과제 유형 설계, 예비 검사 결과, 평가 척도와 등급 기술, 평가 관리 정보, 과제 유형 설계 관련 참고 사항, 평가의 제한점을 제시하였다. 이와 같은 정보들은 평가 설명서 작성의 주요 내용으로서 평가 도구가 객관적으로 준비하고 실행하는 데 요구되는 것들이다.

#### 4. 평가 시행 및 결과 분석 단계: 채점 및 채점자 분석

말하기 평가를 실제로 계획하고 시행하는 근본적인 목적은 평가 결과를 통해 의사 결정을 내리는 것이며, 따라서 의사 결정을 타당하게 내리기 위해서는 평가의 결과인 채점이 타당하게 이루어져야 한다. 말하기 평가에서 채점에 관한 연구는 채점에 영향을 미치는 과제와 수험자 담화 분석을 비롯하여 점수를 부여하는 채점자와 과제 수행 및 채점 과정의 영향 요인 분석, 채점자의 인지적인 특성 등에 이르기까지 매우 폭넓고 다양하다.

말하기 평가의 결과는 숫자나 몇 줄의 문장 이상의 평가 도구와

수험자, 시험관, 검사 환경, 채점자 등과의 상호작용 결과물이다. 특히 평가자(rater)는 수험자의 발화를 듣고 등급이나 채점 기준에 따른 점수를 결정하는 역할을 한다. 학문 목적 한국어 말하기 평가에서 채점은 시험 방법에 따라 다른데, 직접식 검사에서는 대화자인 시험관이 직접 하거나, 반직접식 검사의 경우에는 전문 채점자가 녹음된 것을 듣고 채점을 하게 된다. 직접식 검사를 통해서만 확인할 수 있는 과제의 경우에는 면대면 평가로 진행해야 하지만, 비용과 피로도의 문제, 평가자가 순간적인 기억과 판단에 의존해야 한다는 한계를 고려해야 한다. 이를 해결할 수 있는 반직접식 검사는 발화 상황이 비실제적이지만, 균일한 조건에서 검사가 이루어질 수 있고, 직접식 채점의 물리적 한계를 극복할 수 있다는 점에서 대규모 평가에서 고려될만한 채점 방식이다.

학문 목적 한국어 말하기 평가의 타당한 채점을 위해서는 채점 과정에서 검사 특성을 고려하되, 채점자 훈련이나 연습이 제공되어야 한다. 채점 시에는 평가 기준표의 해석과 채점자의 언어적 배경, 채점 경험 등이 영향을 미칠 수 있으며, 이와 관련한 채점자 특성 연구가 함께 이루어져야 채점의 타당도가 확보될 수 있다. 채점자의 언어적 배경은 모국어에 대한 채점을 하는지 아니면 외국어에 대한 채점을 하는지에 관한 문제와 수험자의 모어에 대한 채점자의 능력과 관련이 된 것이며, 채점 경험은 숙련된 채점자와 숙련되지 않은 채점자에 대한 측면을 연구하는 것이다. 이와 관련한 채점자의 성별, 나이, 전공, 교육 경험 등의 변인에 따라 분석하여 채점의 공정성을 확보하고, 채점자 훈련의 내용을 구성할 수 있다.

## 5. 평가 결과 활용 단계: 예측타당도 및 공인타당도 확보

학문 목적 한국어 말하기 평가의 결과는 점수와 능력에 대한 정보를 제공할 뿐 아니라, 평가 결과를 통해 교과 과정에서의 전공별 수학 가능도나 일반적인 한국어 능력 평가 결과들과의 관계가 검증될 때 유용도가 높은 평가가 이루어졌다고 말할 수 있을 것이다. 윤희원·민병곤(2011)에서 제시한 프랑스 대학의 사례에서는 유학생의 재학 수준에 따라 다양한 학문 목적 언어 교육 프로그램을 지원하며, 유럽공통참조기준에 따라 공인된 평가 결과를 활용하여 학문 목적 교육 프로그램 배치를 결정하거나, 단기 집중 연수 과정을 개설하여 학업 상황에 따른 학습 기회를 제공하는 등 유학생의 학업 과정을 고려한 학문 목적 언어 능력 지원을 하고 있음을 확인할 수 있었다. 학문 목적 한국어 말하기 평가가 학습자의 학업 능력 수행 가능성을 어느 정도로 예측할 수 있는지는 예측타당도 연구를 통해 확인할 수 있다. 예측타당도 연구들은 주로 평가 결과와 학업 성취도를 직접적으로 비교하기 쉬우나, 학업 성취도에 미치는 영향 요인을 통제하는 것이 불가능하기 때문에 측정 오차가 크거나 결과의 확대 또는 축소 해석의 오류가 발생할 수 있다. 이러한 오류를 피하기 위해서는 영향 요인을 최소화해야 하며, 지속적인 관찰을 통해 수험자의 말하기 능력에 대한 타당한 근거를 구축해야 하는데, 안면 타당도를 가진 유학생 지도 교수의 평정이나 인식을 통해서 알아보는 것도 한 방법이 될 수 있다. 말하기 평가의 결과가 유학생의 한국어 능력과는 어떠한 관계인지 알아보기 위한 공인타당도 연구도 이루어질 수 있는데, 이는 TOPIK에서 받은 등급과의 상관관을 통해 확인할 수 있으며, 이러한 결과는 검사 결과의 동등화(equating) 연구로도 이어질 수 있다. 이때에는 각 검사 도구의 난이도 및 변별도,

추측도 등에 관한 문항 정보가 제공되어야 하며, 평가 도구가 안정화되도록 연구 결과를 바탕으로 개선 작업을 지속적으로 진행하여야 할 것이다.\*

## 참고 문헌

- 강현주(2013), 「한국어 말하기 평가의 구인으로서 상호작용능력 연구」, 고려대학교 박사학위논문.
- 교육부(2016), 『2016년 교육기본통계』, 교육부.
- 권순용(2015), 「학문 목적 한국어 학습자의 말하기 평가 도구 개발 연구」, 부경대학교 석사학위논문.
- 김상경(2015), 「학문 목적 한국어 학습자의 말하기 능력 평가 방안 연구」, 경희대학교 박사학위논문.
- 김상경·박동호(2014), 「한국어교육학: 학문 목적 한국어 말하기 평가 과제 유형 개발 연구 - 다국면 라쉬 모형과 일반화 가능성도 이론 적용을 중심으로-」, 『새국어교육』 100, pp.115-141.
- 김수경(2011), 「소리내어 읽기 말하기 평가과제의 구인타당도 연구」, 이화여자대학교 석사학위논문.
- 김정숙·이정희·장은아·이준호(2010), 『한국어능력시험 체제 개선 연구』, 국립국제교육원.
- 김정숙(2000), 「학문적 목적의 한국어 교육과정 설계를 위한 기초 연구: 대학 진학생을 위한 교육과정을 중심으로」, 『한국어교육』 11-2, 국제한국어교육학회, pp.1-19.
- 나임순(2006), 「외국인 유학생의 문화적응 스트레스와 생활 스트레스에 미치는 영향」, 『한국비영리연구』 5-2, 한국비영리학회, pp.159-197.

---

\* 이 논문은 2016. 11. 15. 투고되었으며, 2016. 11. 22. 심사가 시작되어 2016. 12. 5. 심사가 완료되었고, 2016. 12. 6. 편집위원회 심의를 거쳐 게재가 확정되었음.

- 민병곤·정재미·박재현(2014), 「성인의 국어 말하기 능력 평가 결과 분석」, 『화법연구』 24, 한국화법학회, pp.101-136.
- 민진영(2014), 「외국인 유학생의 한국어 능력 부족에 따른 대학원 생활 경험 연구」, 『다문화교육연구』 7-2, 한국다문화교육학회, pp.159-181.
- 박광진(2010), 「학문 목적 학습자 대상 교실 말하기 평가 방안 연구」, 배재대학교 석사학위논문.
- 박선옥(2009), 「학문 목적 말하기 교재 분석을 통한 교재 개발 방향 연구」, 『어문연구』 37-3, 한국어문교육연구회, pp.381-409.
- 박선윤(2015), 「학문 목적 한국어 말하기 평가 도구 개발 방안」, 중앙대학교 석사학위논문.
- 박자영(2016), 「한국어 말하기 평가 설계를 위한 시험 입력물 연구」, 배재대학교 석사학위논문.
- 서동선(2015), 「외국인 유학생을 위한 대학 수학 목적 한국어능력 평가 방안 연구」, 경희사이버대학교 석사학위논문.
- 송지현(2005), 「학문 목적 한국어 교육을 위한 과제 중심 요구 분석」, 이화여자대학교 석사학위논문.
- 심재경(2014), 「학문 목적 한국어능력시험 개발을 위한 기초 연구」, 고려대학교 석사학위논문.
- 육건화(2014), 「과업 기반 학문목적 한국어 말하기 교육의 설계 연구」, 인천대학교 석사학위논문.
- 유해준(2015), 「외국인 유학생을 위한 대학 교양 한국어 교재 개발 방안」, 『사고외표현』 8-1, 한국사고외표현학회, pp.73-106.
- 윤신애(2008), 「학문 목적 듣기 말하기 필요 분석 연구: 외국인 유학생과 한국어 교사를 대상으로」, 이화여자대학교 석사학위논문.
- 윤희원·민병곤(2011), 「해외 대학의 비모어 대학(원)생을 위한 언어 지원 사례 연구」, 『국어교육연구』 29, 서울대학교 국어교육연구소, pp.389-413.
- 윤희원·서혁(2010), 『한국어능력시험(TOPIK) 개선 방안 연구』, 한국교육과정평가원.
- 이덕희(2003), 「요구 분석을 통한 학문 목적의 한국어 교육과정 설계 연구」,

연세대학교 석사학위논문.

- 이혜영(2004), 「학문 목적 한국어 교과과정 설계 연구」, 『한국어교육』 15-1, 국제한국어교육학회, pp.137-164.
- 지현숙(2005), 「인터뷰 시험 담화 분석을 통한 한국어 구어 능력 평가」, 『국어교육연구』 16, 서울대학교 국어교육연구소, pp.79-104.
- 최정순(2006), 「학문 목적 한국어 교육의 교육과정과 평가」, 『이중언어학』 31, 이중언어학회, pp.277-313.
- 최정순·윤지원(2012), 「연구 동향 분석을 통해 본 학문 목적 한국어교육 연구의 실태와 제언」, 『어문연구』 74, 어문연구학회, pp.131-156.
- 한상미(2009), 「학문 목적 한국어 말하기 평가 연구 - 대학 입학 전 과정을 중심으로-」, 『한국어교육』 20-1, 국제한국어교육학회, pp.207-238.
- 황현주(2005), 「학문 목적 한국어 교육과정 개발을 위한 과제 단위 요구 분석: 중국인 유학생을 대상으로」, 연세대학교 석사학위논문.

Alderson, J. C. & Urquhart, A. H.(1983), "The effect of student background discipline on comprehension: a pilot study", In Hughes, A. and Porter, D.(Eds), *Current Developments in Language Testing*, London: Academic Press, pp.121-127.

Alderson, J. C. & Urquhart, A. H. (1985), "The effect of students' academic discipline on their performance on ESP reading tests", *Language Testing* 2-2, pp.192-204.

Bachman, L. F.(1990), *Fundamental considerations in language testing*, 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.

Balcom, P. & Kozar, S.(1994), "An ESP speaking course of international graduate students", *TESL Canada Journal* 12-1, pp.58-68.

Boldt, R. F. & Oltman, P. K.(1993), *Multimethod construct validation of the Test*



- of Spoken English*, ETS Research Report Series, Princeton, NJ: Educational Testing Service.
- Butler, F. A., Eignor, D., Jones, S., McNamara, T. & Suomi, B. K.(2000), *TOEFL 2000 speaking framework*, ETS Research Report Series, Princeton, NJ: Educational Testing Service.
- Canale, M. & Swain, M.(1980), "Theoretical bases of communicative approaches to second language teaching and testing", *Applied Linguistics* 1, pp.1-47.
- Chafe, W.(1986), "Academic Speaking", *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, pp.26-40.
- Clark, J. L. & Swinton, S. S.(1980), *The Test of Spoken English as a measure of communicative ability in English medium instructional settings*, ETS Research Report Series, Educational Testing Service.
- Cronbach, L. J.(1988), "Five perspectives on validity argument", In H. Wainer & H. Braun(Eds.), *Test Validity*, Hillsdale, NJ: Erlbaum, pp.3-17.
- Davies, A.(2001), "The Logic of Testing Languages for Specific Purposes", *Language Testing* 18-2, pp.133-147.
- Davies, A.(2008), *Assessing Academic English: Testing English Proficiency, 1950-1989: the IELTS Solution*, Cambridge: Cambridge University Press.
- Douglas, D.(2000), *Assessing Languages for Specific Purposes*, Cambridge: Cambridge University Press.
- Douglas, D. & Smith, J.(1997), *Theoretical underpinnings of the Test of Spoken English revision project*, TOEFL Monograph Series, Princeton, NJ: Educational Testing Service.
- Enright, M.(2008), *TOEFL Program History*, TOEFL iBT Research Insight, Princeton, NJ: Educational Testing Service.
- Enright, M. & Tyson, E.(2008), *Validity evidence supporting the interpretation and use of TOEFL iBT scores*, TOEFL iBT Research Insight, Princeton, NJ: Educational Testing Service.
- Fulcher, G.(2003), *Testing Second Language Speaking*, Harlow: Pearson

Education.

Henning, G., Schedl, M. & Suomi, B. K.(1995), *Analysis of proposed revisions of the Test of Spoken English*, ETS Research Report Series, Educational Testing Service.

Lazaraton, A., & Wagner, S.(1996), *The revised test of spoken English(TSE): discourse analysis of native speaker and nonnative speaker data*, ETS Research Report Series, Educational Testing Service.

Linacre, J. M.(2002), *Institute for Objective Measurement*, retrieved from <http://www.rasch.org/rmt/rmt162f.htm>

Luoma, S.(2001), "Test Review: Test of Spoken English", *Language Testing* 18-2, pp.225-234.

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

Munby, J.(1978), *Communicative Syllabus Design*, Cambridge: Cambridge University Press.

Seedhouse, P. & Harris, A.(2011), "Topic development in the IELTS Speaking Test". *IELTS Research Reports* 12, pp.69-124.

Skehan, P.(1984), "Issues in the testing of English for specific purposes", *Language Testing* 1-2, pp.202-220.

<웹 사이트>

국제영어능력시험(International English Language Test System, IELTS) 홈페이지: [www.ielts.org](http://www.ielts.org)

미국 교육평가원(Educational Testing Service, ETS) 홈페이지: [www.ets.org](http://www.ets.org)

영국 케임브리지 대학 영어 평가원(Cambridge English Language Assessment) 홈페이지: <http://www.cambridgeenglish.org>

한국 영국문화원 홈페이지: [www.britishcouncil.kr](http://www.britishcouncil.kr)

## ■ 국문초록

# 학문 목적 한국어 말하기 평가 도구 개발의 쟁점과 과제

민병곤 · 이성준

본 연구는 학문 목적 한국어 말하기 평가의 도구의 개발과 관련한 주요 쟁점과 과제를 제시하는 것을 목적으로 한다. 이를 위해 먼저 학문 목적 말하기 평가를 구성하는 이론적 배경으로서 특수 목적 언어 평가 연구와 학문 목적 한국어 말하기 교육, 그리고 학문 목적 한국어 말하기 평가에 관한 연구들을 검토하였다. 그리고 대표적인 국외 학문 목적 말하기 시험을 실시하고 있는 TOEFL과 IELTS에서 말하기 시험이 개발되어 현재의 모습으로 사용되기까지의 과정을 살펴보고, 이를 통해 학문 목적 한국어 말하기 평가 도구 개발에 관한 시사점을 도출하였다. 이상의 논의를 바탕으로 학문 목적 한국어 말하기 평가 도구 개발의 쟁점으로 제시한 것은 평가의 개념 정립과 구인 설정, 평가 과제 개발, 평가 방법 선택 및 타당화, 채점과 척도 개발, 교육적 환류와 영향력이다. 끝으로 학문 목적 한국어 말하기 평가 도구의 개발을 위한 앞으로의 과제로서 개발 과정에서 응시 목적을 고려하여 요구 분석을 실시하고, 구인타당도 확보와 면밀한 평가 계획서 작성, 그리고 채점자 분석 연구의 필요성을 강조하고, 예측타당도와 공인 타당도 연구를 통해 학문 목적 한국어 말하기 평가 도구의 안정화와 품질 개선 작업이 지속적으로 이루어져야 함을 제안하였다.

[주제어] 말하기 평가, 학문 목적 언어 평가, 언어 평가 도구 개발, 타당화 연구, 한국어교육

■ Abstract

## Issues in the development of Korean speaking test for academic purposes

Min, Byeong-gon · Lee, Sung-jun

This paper examines recent approaches to developing speaking tests and in particular the testing of Korean speaking for academic purposes. The first approach is concerned with testing theory of the language for specific purposes (LSP) and Korean for academic purposes (KAP) as a means of verifying the claims for speaking test development. The second approach is analyzing the validation procedures of TOEFL and IELTS as representative academic speaking tests and relating them to the development of Korean speaking tests for academic purposes. The article argues that the development of Korean speaking tests for academic purposes is founded on the conceptual framework of speaking tests, construct setting, item writing, connecting testing method and validation, scoring rubric and rating scale development, and consideration of educational washback and impact. From this view point, developing Korean speaking tests for academic purposes can be justified in terms of practical needs analysis, construct validity projects, circumstantial specification, and rater effect analysis, and it is suggested that continuous concurrent and predictive validation research be conducted to improve speaking tests.

[Key words] speaking assessment, language testing for academic purposes, test development, validation project, Korean language education