한국어 능력 시험 평가 문항의 내용타당도 분석*

－제12회 일반 한국어(S-TOPIK)의 쓰기·듣기·읽기 영역을 중심으로－

전은주**

＜차 례＞

1. 머리말

이 논문은 한국어 능력 시험(TOPIK : Test of Proficiency in Korean)[1]의 평가 문항이 한국어 학습자의 숙달도를 평가함에 있어서 어느 정도의 내용타당도를 지니는가를 분석하는 데 목적이 있다. 한국어 능력 시험은 1997년 처음 시작되어 2007년 한국교육과정평가원이 주관하여 제12회 시험을 시행하였다.[2] http://www.topik.or.kr은 한국어 능력 시험을 운영하기 위하여

* 이 논문은 2007년도 부산대학교 인문사회연구기금에 의하여 연구되었음.
이 논문은 국어교육학회 제38회 학술발표대회(2007. 12. 15)에서 발표한 내용을 수정한 것이다.

** 부산대학교

1) 한국어 능력 시험의 영어 명칭은 KPT(Korean Proficiency Test)에서 2005년부터 TOPIK (Test of Proficiency in Korean)으로 개칭되었다.

2) 한국어능력평가 제1, 2회(1997년, 1998년)는 한국학술진흥재단에서 주관하였다.

개설한 홈페이지로 여기에는 한국어 능력 시험의 목적을 다음과 같이 밝히고 있다.

> 한국어 능력 시험의 목적은 한국어를 모국어로 하지 않는 외국인 및 재외동포들에게 한국어 학습 방향을 제시하고, 한국어 보급을 확대하며, 그들의 한국어 사용능력을 측정하여 그 결과를 유학, 취업 등에 활용케 하는 데 있다.

이 목적 이외에 한국교육과정평가원(2005)에서는 "한국어 교육 기관의 연수, 교육과정 및 교육평가 방법을 표준화하는 것"을 추가하고 있다. TOPIK 홈페이지와 한국교육과정평가원(2005)에 한국어 능력 시험의 목적을 위와 같이 여러 가지로 진술하고 있지만 이 시험의 주된 목적은 응시자의 한국어 사용 능력을 평가하는 것이다. 한국어 능력 시험은 한국어 학습자와 한국어 교육 종사자에게 국가에서 공인하는 한국어 숙달도 평가라는 점에서 큰 의미를 지니므로 언어 능력 평가 도구로서 갖추어야 할 요건을 충족한 것이어야 한다.

한국어 능력 시험에 대한 연구는 여러 측면에서 이루어져 왔다. 김정숙 외(2005ㄱ, ㄴ)에서는 한국어 능력 시험의 문제점을 개선하기 위한 방안을 제시한 바 있다. 김정숙 외(2005ㄱ)에서는 한국어 능력 시험의 실용도를 높이기 위하여 등급 부여 방식을 6종 6등급 체계에서 3종 6등급 체계의 시험으로 개선할 것을 제안하였으며, 김정숙 외(2005ㄴ)에서는 평가의 신뢰도를 높이기 위한 방안으로 평가 문항 유형을 제안한 바 있다.[3] 이러한 연구 결과의 반영으로 제10회 한국어 능력평가에서는 3종 6등급 체계를 채택하고 있다. 김유정(2006)은 제6~8회 한국어 능력 시험의 등급별, 영역별 난이도를 분포를 분석한 논문으로, 세 차례의 시험에서 난이도 분포가 동일하지 않으며, 문항 난이도 하의 문항수와 비례하여 합격률이 높아지는 문제점을 지적하고 있다. 이해영(2005)에서는 제8회 한국어 능력 시험

3) 이 논문에서는 어휘・문법, 듣기, 읽기 영역에서 100% 객관식 문항 출제와 쓰기 영역에서 10개의 객관식 문항과 4~6개의 주관식 완성형문항과 1개의 주관식 작문 문항을 출제할 것을 제안하였다.

에 사용된 듣기와 읽기 영역의 텍스트와 문항 유형을 분석하여 이해 능력 평가를 위한 문항 작성 지침을 제시하고 있다. 이 외에도 민병곤(2005)에서는 한국어 능력 시험의 지원, 시행 관리, 출제 관리 등 운영 전반에 걸친 현황과 과제를 살피고 있으며, 조항록(2006)에서는 국내외에서 시행되고 있는 한국어 숙달도 평가들에 대하여 살펴보면서 한국어 능력 시험과 다른 시험의 평가 체계에 대하여 비교하고 있다. 한국어 능력 시험을 온전히 평가하기 위해서는 평가 목표와 내용, 영역, 문항 수, 배점, 문항 유형, 점수체제 등 시험 체제의 측면, 문항의 타당도, 신뢰도, 객관도 실용도 등 평가 도구의 측면, 출제, 시행, 채점, 결과 활동 등 시험 운영·관리의 측면을 전반적으로 살펴보아야 한다(민병곤, 2005 : 139). 한국어 능력 시험이 국가에서 공인한 명실상부한 평가도구로서 제 역할을 다하기 위해서는 한국어 능력 시험에 대한 다각도의 연구를 통하여 그 문제점을 개선해나가야 한다.

그러나 앞서 살펴본 바와 같이 한국어 능력 시험에 대한 선행연구들은 실용도, 신뢰도, 난이도, 평가 체제, 문항 유형과 텍스트 유형 분석 등의 분야에서 주로 이루어졌으며, 한국어 능력 시험이 시험의 주된 목적인 한국어 사용 능력을 제대로 평가하고 있는가 하는 타당도에 대한 검토는 제대로 이루어지지 않았다. 어떤 종류의 측정도구이든지 그것의 타당성을 판단하기 위해 제작 초기에 필수적으로 겪어야 하는 과정이 내용타당도의 확인이다(황정규, 2002 : 99). 내용타당도(content validity)란 한 검사가 측정하려는 타당성의 준거를 그 측정 도구의 내용, 즉 내적 준거에 비추어 보는 타당도이다(황정규, 2002 : 98). 따라서 이 글에서는 타당도의 범위를 제한하여, 한국어 숙달도 평가인 한국어 능력 시험이 학습자의 한국어 능력을 평가하는 도구로서 내용타당도(content validity)를 충족하고 있는가에 대하여 살펴보고자 한다.[4] TOPIK 홈페이지에는 이 시험의 평가 기준을 등급별로

4) 타당도란 한 검사 혹은 평가도구가 '측정하려고 의도하는 것'을 어느 정도로 충실히 측정하고 있느냐로 타당도의 개념 속에는 반드시 준거(criteria)의 개념이 수반된다(황정규, 2002 : 96). 즉 타당도란 평가도구가 설정하고 있는 준거에 비추어보았을 때 평가가 가능한 개념이며, 평가도구가 가지고 있는 맥락을 떠나 이루어진 논의는 의미가 없다. 일

제시하고 있다. 이것은 한국어 능력 시험에서 평가하고자 한 것이 무엇인가 그 기준을 제시한 것으로, 이 평가도구에서 측정하고자 하는 내용에 대한 대표성을 가진 것이므로 한국어 능력 시험의 내적 준거로 볼 수 있다.[5] 그러므로 이 논문에서는 2007년 하반기에 실시된 제12회 한국어 능력 시험의 일반 한국어(S-TOPIK)를 자료로 하여, 한국어 능력 시험의 평가 문항들이 평가 기준에 규정된 내용을 어느 정도로 적절하게 측정하고 있는가를 분석해 봄으로써 한국어 능력 시험의 내용타당도를 살펴보고자 한다.

2. 한국어 능력 시험의 내용타당도 분석 기준

2.1. 한국어 능력 시험의 평가 기준

대부분의 평가 도구는 평가하고자 하는 특성의 전체를 측정하는 것이 아니라 그 특성을 대표하는 일부분만을 측정한 뒤 그 결과를 전체 해석에 적용하게 된다. 그러므로 평가 도구가 측정하고자 하는 내용을 적절하게 평가하고 있을 때 내용타당도가 확보되는 것이며, 내용타당도가 있어야 평가 결과를 전체 내용에 대하여 확대 해석할 수 있다. 언어 평가에서 내용타당도 검증의 가장 큰 문제점은 언어능력의 구성 내용이 아직까지 완전히 규정되지 못하여 검증의 기준이 주관적이라는 점이다(정동빈 외, 1991 : 284). 한국어 능력 시험에서 평가하고자 하는 한국어 사용 능력 역시 추상적인

반적으로 타당도는 내용타당도, 예언타당도, 공인타당도, 구인타당도로 나누고 예언타당도와 공인타당도를 준거 관련 타당도라고 한다.

5) 한국어 능력 시험이 응시자의 한국어 사용 능력을 측정하기 위한 목적을 제대로 달성하기 위해서는 이 평가도구가 측정하고자 하는 한국어 사용 능력이 무엇인가 하는 평가 기준을 타당하게 설정하는 작업이 먼저 이루어져야 한다. 그러나 본고는 한국어 능력 시험 평가 문항의 내용타당도 분석이 목적이므로 이 평가 기준 자체의 타당성에 대한 본격적인 논의는 하지 않겠다.

개념이므로 그 구성 요소를 객관화하기는 매우 어렵다. 그러나 한국어 사용 능력을 평가하기 위해서는 이 추상적인 개념의 특성을 대표하는 일부를 정하고 이것을 평가의 기준으로 설정한 뒤, 이 평가 기준을 평가의 목표로 두고 이에 대하여 측정하는 방식을 취할 수밖에 없다. 그러므로 이 논문에서는 앞서 언급한 바와 같이 한국어 능력 시험 요강에 평가 기준으로 제시되어 있는 것을 한국어 능력 시험의 내용타당도를 분석하는 준거로 삼고자 한다.

앞서 언급한 바와 같이 한국어 능력 시험은 과거 6종의 시험을 통하여 1~6등급으로 등급을 판정하던 것을 2006년부터 초, 중, 고급 3종의 시험을 실시하고 각 등급 내에서 점수에 따라 상, 하를 나누어 총 6등급을 부여하는 방식으로 바뀌었다.[6] 이는 기존의 등급제를 등급제와 점수제를 결합한 절충식 방식으로 볼 수 있다. 한국어 능력 시험의 문항 구성은 다음과 같다.

<표 1> 한국어 능력 시험의 문항 구성

영 역	어휘·문법	쓰 기		듣 기	읽 기
유 형	객관식	주관식	객관식	객관식	객관식
문항수	30	5~7	10	30	30
배 점	100	60	40	100	100

2006년 이전 등급제 방식 시절에는 등급별로 어휘·문법, 쓰기, 듣기, 읽기 영역 각각에 대한 평가 기준을 제시하던 것을 2006년부터는 등급별로 영역에 대한 구분 없이 아래와 같이 평가 기준을 제시하고 있다.

6) 한국어 능력 시험은 평가 영역별로 과락 점수가 없고, 전 영역 평균성적이 합격 점수에 해당할 경우 그에 맞는 평가 등급을 부여한다. 평가등급별 합격점수와 과락점수는 다음과 같다.

시험구분	초 급		중 급		고 급	
평가등급	1급	2급	3급	4급	5급	6급
합격점수	50점 이상	70점 초과	50점 이상	70점 초과	50점 이상	70점 초과
과락점수	40점 미만	50점 미만	40점 미만	50점 미만	40점 미만	50점 미만

<표 2>일반 한국어 능력 시험의 평가 기준(http://www.topik.or.kr)

등 급		평가 기준
초급	1급	• '자기 소개하기, 물건사기, 음식 주문하기' 등 생존에 필요한 기초적인 언어 기능을 수행할 수 있으며 '자기 자신, 가족, 취미, 날씨' 등 매우 사적이고 친숙한 화제에 관련된 내용을 이해하고 표현할 수 있다. • 약 800개의 기초 어휘와 기본 문법에 대한 이해를 바탕으로 간단한 문장을 생성할 수 있다. • 간단한 생활문과 실용문을 이해하고, 구성할 수 있다.
	2급	• '전화하기, 부탁하기' 등의 일상생활에 필요한 기능과 '우체국, 은행' 등의 공공시설 이용에 필요한 기능을 수행할 수 있다. • 약 1,500~2,000개의 어휘를 이용하여 사적이고 친숙한 화제에 관해 문단 단위로 이해하고 사용할 수 있다. • 공식적 상황과 비공식적 상황에서의 언어를 구분해 사용할 수 있다.
중급	3급	• 일상생활을 영위하는 데 별 어려움을 느끼지 않으며, 다양한 공공시설의 이용과 사회적 관계 유지에 필요한 기초적 언어 기능을 수행할 수 있다. • 친숙하고 구체적인 소재는 물론, 자신에게 친숙한 사회적 소재를 문단 단위 로 표현하거나 이해할 수 있다. • 문어와 구어의 기본적인 특성을 구분해서 이해하고 사용할 수 있다.
	4급	• 공공시설 이용과 사회적 관계 유지에 필요한 언어 기능을 수행할 수 있으며, 일반적인 업무수행에 필요한 기능을 어느 정도 수행할 수 있다. • 또한 '뉴스, 신문 기사' 중 평이한 내용을 이해할 수 있다. 일반적인 사회적, 추상적 소재를 비교적 정확하고 유창하게 이해하고, 사용할 수 있다. • 자주 사용되는 관용적 표현과 대표적인 한국 문화에 대한 이해를 바탕으로 사회, 문화적인 내용을 이해하고, 사용할 수 있다.
고급	5급	• 전문 분야에서의 연구나 업무 수행에 필요한 언어 기능을 어느 정도 수행할 수 있다. • '정치, 경제, 사회, 문화' 전반에 걸쳐 친숙하지 않은 소재에 관해서도 이해하고 사용할 수 있다. • 공식적, 비공식적 맥락과 구어적, 문어적 맥락에 따라 언어를 적절히 구분해 사용할 수 있다.
	6급	• 전문 분야에서의 연구나 업무 수행에 필요한 언어 기능을 비교적 정확하고 유창하게 수행할 수 있다. • '정치, 경제, 사회, 문화' 전반에 걸쳐 친숙하지 않은 주제에 관해서도 이용하고 사용할 수 있다. • 원어민 화자의 수준에는 이르지 못하나 기능 수행이나 의미 표현에는 어려움을 겪지 않는다.

한국어 능력 시험이 내용타당성을 만족하는 평가도구라면 이 평가 기

준에서 제시한 내용이 등급별 평가 문항에 제대로 반영되어야 한다. 그러나 위 평가 기준에는 각 등급별로 한국어 사용 능력을 평가할 때 무엇에 대하여 평가할 것인가 하는 평가 범주가 세분화되어 있지 않고 포괄적으로 기준이 제시되어 있다. 뿐만 아니라 평가의 목적이 한국어 사용 능력의 측정이다 보니 평가 기준이 실제 표현과 이해의 의사소통 상황에서의 수행 정도에 초점을 두고 있어 어휘나 문법 영역에 대한 것은 그리 뚜렷하지 않다. 또 각 등급별로 동일한 속성에 대하여 평가 기준이 지속적으로 제시되어 있지 않아 전후 등급에 기술된 내용을 통하여 추론해 보아야 하는 문제점을 가진다.

2.2. 내용타당도 분석을 위한 범주 설정

한국어 능력 시험의 내용타당도를 <표 2>의 평가 기준에 나타난 공통적 속성을 중심으로 범주화하고 이것이 등급별로 문항에 제대로 실현되어 있는가를 살펴본다면 분석 항목이 뚜렷해진다는 장점이 있다. 따라서 이 글에서는 평가 기준에 여러 등급에 걸쳐 나타나는 공통적인 속성을 범주화하고 이를 내용타당도 분석의 범주로 삼고자 한다. 앞서 살펴 본 <표 2>의 평가 기준은, 한국어 숙달도가 높을수록 동일한 내용 항목에 대하여 수준의 차이가 어떻게 나타나야 하는가에 대한 것이 분명히 드러나지 않는 부분이 있어 체계적인 기술이라고 보기는 어렵다. 그러나 여러 등급에 걸쳐, 평가 기준에 공통적으로 비슷한 성격의 것들이 포함되어 있다. 예를 들자면 모든 등급에는 다음과 같이 '언어 기능(function)'에 대한 내용이 있다.

- 1급 : 기초적인 언어 기능을 수행할 수 있다.
- 2급 : 일상생활에 필요한 기능과 공공시설 이용에 필요한 기능을 수행할 수 있다.
- 3급 : 다양한 공공시설의 이용과 사회적 관계 유지에 필요한 기초적 언어 기능을 수행할 수 있다.
- 4급 : 공공시설 이용, 사회적 관계 유지에 필요한 언어 기능, 일반적인

업무수행에 필요한 기능을 어느 정도 수행할 수 있다.
- 5급 : 전문 분야에서의 연구나 업무 수행에 필요한 언어 기능을 어느 정도 수행할 수 있다.
- 6급 : 전문 분야에서의 연구나 업무 수행에 필요한 언어 기능을 비교적 정확하고 유창하게 수행할 수 있다.

이 밖에 언어 사용의 맥락, 내용, 텍스트 혹은 담화 형태[7] 등에 대한 항목이 공통적으로 포함되어 있으며, 4급과 6급에는 정확성과 유창성에 대한 항목도 포함되어 있다. 이들 항목 중 정확성과 유창성은 실제 한국어 능력 시험에서 난이도 조정에 의하여 평가 점수로 변별이 되기 때문에 내용타당도 분석을 위한 평가 범주로 포함하지 않았다. 그러므로 평가 기준에 제시된 사항을 내용타당도 분석을 위한 범주에 따라 다음과 같이 정리할 수 있다.

<표 3> 한국어 능력 시험의 범주별 평가 기준

등 급		평가 기준			
		기 능	맥 락	내 용	텍스트 형태
초급	1급	• 기초적인 언어 기능	• 사적 • 생활문, 실용문	• 친숙한 화제	• 간단한 문장
	2급	• 일상생활 언어 기능 • 공공시설 이용 기능	• 사적 • 공식적 상황과 비공식적 상황의 구분	• 친숙한 화제	• 문단 단위

7) 음성언어 의사소통에서 표현되거나 이해되는 언어 사용의 결과물인 담화텍스트의 형식을 이글에서는 편의상 담화 형태라고 부르겠다.

8) ACTFL(American Council on the Teaching of Foreign Languages)의 숙달도(proficiency) 평가에서는 평가 등급을 초급(Novice)－중급(Intermediate)－상급(Advanced)－최상급(Superior)으로 두고 과제/기능, 맥락, 내용, 정확성, 텍스트 형태 등을 평가 범주로 하여 기준을 제시하고 있다. 이 글에서 한국어 능력 시험의 평가 기준을 분석해 본 결과 이들 다섯 평가 범주에 대한 기술이 포함되어 있었으나 각 등급에 규칙적으로 적용되어 기술되지는 않았다. 언어 능력의 숙달도를 평가한다는 측면에서 평가 대상 언어가 한국어이든 영어이든, 평가 범주는 대동소이할 수 있다고 생각하나 숙달도에 따라 각각의 평가 범주의 특성이 어떻게 달라지는가에 대한 체계적 진술은 필요하다.

등 급		평가 기준			
		기 능	맥 락	내 용	텍스트 형태
중급	3급	• 공공시설 이용 가능 • 사회적 관계 유지에 필요한 기초적인 언어 가능	• 문어와 구어의 기본적 특성 구분	• 친숙한 사회적 소재	• 문단 단위
	4급	• 공공시설 이용 가능 • 사회적 관계 유지에 필요한 기능 • 일반적 업무수행가능	• 뉴스, 신문기사	• 뉴스, 신문기사 중 평이한 내용 • 사회적, 추상적 소재 • 사회, 문화적인 내용	
고급	5급	• 전문 분야 연구 가능 • 전문 분야 업무 수행 가능	• 공식적 비공식적 맥락 • 문어적 구어적 맥락	• 정치, 경제, 사회, 문화 전반에 걸쳐 친숙하지 않은 소재	
	6급	• 전문 분야 연구 가능 • 전문 분야 업무 수행 가능		• 정치, 경제, 사회, 문화 전반에 걸쳐 친숙하지 않은 소재	

위 표에서 살필 수 있듯이 '기능'은 응시자가 언어를 사용하여 기능을 수행할 수 있는 능력과 관련되며, '맥락'은 텍스트가 표현되거나 이해되는 환경 또는 조건 등에 대한 것이다. '내용'은 텍스트의 소재나 화제, 주제 등에 대한 것이며, '텍스트 형태'는 표현되거나 이해되는 텍스트의 양에 대한 것이다. 위 표에 빈칸으로 나타나는 부분은 바로 위 등급의 수준이 지속되는 것으로 해석할 수 있으며, 초급, 중급, 고급 내에서 상, 하로 등급을 구분하는 기준으로 정확성과 유창성의 개념이 적용되어 있다고 볼 수 있다.

이 글에서는 이 평가 기준을, 한국어 능력 시험의 내용타당도를 분석하는 준거로 삼아 한국어 능력 시험의 문항에 이러한 평가 범주의 내용 특성이 등급별로 적절히 평가되었는가를 살펴보고자 한다. 그러나 앞서 언급한 바와 같이 이 평가 기준에 어휘·문법' 영역을 위한 평가 기준이 제대로 제시되어 있지 않으므로 한국어 능력 시험의 '쓰기', '듣기', '읽기' 영역의 평가 문항으로 자료를 제한하겠다. 또, 평가 범주에 대한 내용타당도 분석을 위하여 각 범주를 세분할 내용을 설정할 때에 한국어 능력 평

가의 평가 기준에 제시된 것을 중심으로, 평가 기준에 기술된 표현을 최대한 수용하여 사용하고자 한다. 이 경우 세부 내용이 범주의 특성을 분석하는 체계에 다소 맞지 않더라도 이 글이 평가 기준을 준거로 한국어 능력 시험의 내용타당도를 분석하고자 한 것이므로 평가 기준의 표현을 최대한 수용하는 것이 혼란의 여지를 줄일 수 있는 방법이기 때문이다.

3. 기능과 맥락 범주의 내용타당도

3.1. 기능 범주

한국어 능력 시험은 숙달도에 따라 수행할 수 있는 언어 기능에 차이가 있음을 변별할 수 있어야 한다. 즉 1급이라는 숙달도를 판정받은 사람보다는 2급이라는 숙달도를 판정받은 사람이 언어를 사용하여 수행할 수 있는 기능이, 난이도가 더 높거나 많아야 한다. 현재 한국어 능력 시험에서는 숙달도가 높을수록 수행할 수 있는 언어 기능이 다음과 같이 변별되어야 한다고 평가 기준으로 제시하고 있다.

<표 4> 기능 범주의 등급별 수행 정도

언어 기능	초 급		중 급		고 급	
	1급	2급	3급	4급	5급	6급
기초적인 언어 기능	✓					
일상생활 언어 기능		✓				
공공시설 이용 기능		✓	✓✓	✓✓		
사회적 관계 유지에 필요한 언어 기능			✓	✓✓		
일반적 업무수행기능				✓		
전문 분야 연구 기능					✓	✓✓
전문 분야 업무 수행 기능					✓	✓✓

9) ✓는 보통 수준, ✓✓는 그 항목의 ✓보다 심화된 수준을 의미함.

<표 4>에서 살필 수 있듯이 1급에서 생존에 필요한 기초적인 언어 기능을 수행할 수 있는 것에서 출발하여 숙달도가 높아질수록 사회적 관계, 일반 업무 수행, 전문 분야 연구와 업무 수행 등으로 기능이 심화된다. 평가 기준에는 기초적인 언어 기능의 예로 자기 소개하기, 물건사기, 음식 주문하기 등을, 일상생활에 필요한 기능의 예로는 전화하기, 부탁하기 등을 들고 있고, 이 외의 언어 기능의 예에 대해서는 구체적으로 언급을 하고 있지 않다. 한국어 능력 시험의 네 평가 영역 중 의사소통의 맥락을 가지고 언어를 사용하여 표현하거나 이해하는 기능이 표출될 수 있는 영역은 쓰기, 듣기, 읽기 영역이다. 그러므로 이들 영역에 대하여 제12회 일반 한국어 시험에서 등급이 높아짐에 따라 수행할 수 있는 기능이 어떻게 변화되도록 문항을 출제하였는가를 살펴보았다.

<표 5> 평가 문항에 나타난 언어 기능의 빈도(%)

언어 기능	초급			중급			고급		
	쓰기	듣기	읽기	쓰기	듣기	읽기	쓰기	듣기	읽기
기초적인 언어 기능	35.3	50.0	26.7	0.0	0.0	0.0	0.0	0.0	0.0
일상생활 언어 기능	58.8	43.3	66.7	93.8	80.0	93.3	46.7	56.6	36.7
공공시설 이용 기능	5.9	6.7	6.7	0.0	0.0	0.0	0.0	0.0	0.0
사회적 관계 유지에 필요한 언어 기능	0.0	0.0	0.0	0.0	13.3	0.0	0.0	3.3	0.0
일반적 업무수행기능	0.0	0.0	0.0	6.2	0.0	0.0	0.0	6.7	0.0
전문 분야 연구 기능	0.0	0.0	0.0	0.0	6.7	6.7	40.0	26.7	63.3
전문 분야 업무 수행 기능	0.0	0.0	0.0	0.0	0.0	0.0	13.3	6.7	0.0
합 계	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

평가 문항에 나타난 언어 기능의 빈도를 나타낸 <표 5>를 통하여 다음과 같은 분석 결과를 얻을 수 있다.

① 초급 문항에 나타난 언어 기능 범주는 쓰기, 듣기, 읽기 영역 모두에서 기초적인 언어 기능, 일상생활 언어 기능, 공공시설 이용 기능 등을 골고루 다루고 있으므로 평가 기준에 비추어 내용타당도가 매우 높다.

② 중급 문항에서 사회적 관계 유지에 필요한 기능과 일반적 업무 수행 기능에 대한 평가가 제대로 이루어지지 않고 있다. 평가 기준에 따르면 언어 기능과 관련하여 초급과 중급 문항의 가장 큰 차이점은 중급에서 이 두 기능에 대하여 측정한다는 점이다. 그러나 위 결과에서 나타나 있듯이 그 빈도가 매우 낮다.

③ 고급 문항은 주로 일상생활 언어 기능과 전문 분야 연구 기능에 대한 평가로 이루어져 있고, 평가 기준에 제시한 전문 분야 업무 수행 기능에 대한 평가 문항은 그 빈도가 매우 낮았다.

④ 평가 기준에 숙달도가 높아짐에 따라 다양한 언어 기능을 정확하고 유창하게 수행할 수 있어야 한다고 제시하고 있으나 실제 한국어 능력 시험의 평가 문항에서는 숙달도가 높아짐에 따라 일상생활 언어 기능을 수행하는 텍스트의 어휘, 문장 구조, 화제 등의 난이도를 높임으로써 등급 간의 차이를 꾀하는 방식을 취하고 있다.[10]

이상과 같이 한국어 능력 시험의 평가 문항은 평가 기준에서 해당 등급에서 수행할 수 있어야 하는 언어 기능 특성이라고 정해 놓은 것을 실

10) 듣기 능력 평가에서 난이도를 조정하는 방법을 박영순(2004 : 279~280)에서는 다음과 같이 설명하고 있다.
난이도는 지문의 문장 구조, 어휘, 문장의 길이 등에 의해서 일차적으로 정해질 수 있겠지만, 의미론적으로 언어적 해석, 화용론적 해석, 관용적 해석, 은유적 해석을 필요로 하느냐에 따라서도 결정될 수 있고, 직접적 표현이냐 간접적 표현이냐, 주제가 표면에 명시적으로 잘 드러나 있느냐, 함축적이고 암시적으로 숨어 있느냐에 따라서도 난이도가 달라질 것이다. 또한 청자의 상상력이나 추리력이 필요하냐, 단순한 이해력으로 해결 가능하느냐에 따라서도 듣기 문제의 수준이 달라지며, 어휘의 의미가 단순하냐 복잡하냐, 일상어휘냐 특수 전문 어휘냐에 따라서도 달라진다. 그러므로 듣기평가는 크게 어휘력, 문장해석력, 담화 해석력 면에서 평가가 이루어질 수 있고, 말의 요지, 줄거리, 강조점 등에 대한 이해, 분석, 종합력을 골고루 테스트할 수 있다.

제 평가에서는 대체적으로 제대로 측정하고 있지 않음을 살필 수 있었다. 그러므로 평가 기준에서 언어 기능 범주에 해당하는 부분에 대해서는 내용타당도를 높이는 문항 구성이 필요하다.

3.2. 맥락 범주

한국어 능력 시험의 평가 기준에는 언어를 사용하는 환경을 사적 맥락과 공식적 맥락으로 구분하고 있다. 사적 의사소통 상황에서 언어를 사용하여 내용을 이해하고 표현할 수 있다가 점차로 숙달도가 높아져 2급 이상이 되면 공식적 의사소통과 비공식적 의사소통 상황을 구분하여 언어를 사용할 수 있어야 한다는 것이다. 또 3급 이상이 되면 상황 맥락에 따라 문어와 구어의 기본 특성을 구분해서 이해하고 사용할 수 있는지를 평가 기준에 설정하고 있다. 즉 음성언어 의사소통의 경우라도 공식적 상황에다가 격식성까지 갖추어야 하는 경우는 문어의 특성을 가진 표현을 할 수 있어야 한다는 것이다. 반면에 공식적 상황이라도 격식성을 갖추지 않아도 되는 경우라면 의사소통 참여자에게 친밀감을 주기 위하여 구어의 특성을 가진 표현을 할 수 있어야 한다. 읽기와 쓰기와 같이 문자언어 의사소통의 상황이라면 상황 맥락에 따라 문어나 구어로 된 글을 읽고 쓸 수 있어야 한다. 아래 <표 6>은 한국어 능력 시험의 평가 기준에 나타난 맥락 범주에 대한 항목이 등급별로 어느 정도의 수행을 할 수 있어야 하는가를 정리한 것이다.

<표 6> 맥락 범주의 등급별 수행 정도

맥 락	초급		중급		고급	
	1급	2급	3급	4급	5급	6급
사적 상황	✓	✓✓				
공식적 상황과 비공식적 상황의 구분		✓			✓✓	
문어와 구어의 특성 구분			✓		✓✓	

한국어 능력 시험의 평가 문항에 언어를 사용하는 맥락이 등급에 따라 적절히 반영되었는가를 살피기 위해서는 다양한 측면에서 접근을 하여야 한다. 왜냐하면 맥락이 의사소통을 구성하고 있는 여러 요소에 영향을 미치기 때문이다. 우선 문항 유형 자체가 공식적이거나 비공식적인 상황에서, 문항에서 의도한 맥락에 따라 적절한 표현을 찾게 하는 것이 실제 어느 정도 출제되었는가를 살펴보는 방법을 생각할 수 있다. 이것은 맥락에 따라 적절한 표현을 할 수 있는가를 측정하는 직접적인 방법이다. 그러나 직접적 방법은 아니지만 언어 사용의 맥락을 살펴보는 다음과 같은 방법도 가능하다. 우선 평가 문항에 사용된 텍스트의 맥락을 분석해 보는 방법이 있다. 이는 평가 문항 유형이 어떤 것이든 그 문항에 정확히 반응하기 위해서는 텍스트가 안고 있는 맥락을 이해할 수 있어야 하기 때문이다. 또 어말어미의 사용 양상을 살펴봄으로써 그 텍스트가 사용된 맥락이 공식적 상황인지 비공식적 상황인지를 판단할 수 있다. 일반적으로 비공식적 상황에서는 '-아/어요'나 반말을 주로 쓰며 공식적인 상황에서는 '-(스)ㅂ니다'를 사용한다. 그러나 공식적 상황이라 하더라도 참여자 간의 관계, 발화자의 성별, 격식성의 정도 등에 따라 '-아/어요'를 쓰기도 하고 '-아/어요'와 '-(스)ㅂ니다'를 혼용해서 쓰기도 한다. 결국 어떤 어말어미를 선택하는가는 상황 맥락 내에서 결정되는 것이다. 이 밖에 담화 혹은 텍스트의 종류를 살펴보는 방법 역시 맥락을 판단할 수 있는 방법이 된다. 텍스트의 종류 역시 문어체를 써야 할 것인지 구어체를 써야 할 것인지 판단할 수 있는 근거가 된다.

제12회 일반 한국어의 평가 문항에서 텍스트의 맥락과 관련된 사항을 어떻게 측정하고 있는가를 위의 네 방법을 사용하여 분석해 보았다.

1) 직접적으로 평가한 문항

맥락에 따라 적절한 표현을 하는가는 쓰기 영역에서, 맥락을 충분히 이해하고 이에 이어지는 맞는 표현을 찾아낼 수 있는가는 듣기, 읽기 영역에서 평가가 가능하다. 그러나 이처럼 맥락 사용에 대한 평가를 직접적

으로 하는 문항은 없었다. 다만 쓰기 초급과 고급에 이러한 평가의 가능성이 엿보이는 문항이 2문항씩 있었다. 이들은 글을 읽고 ( ) 안에 알맞은 말을 쓰게 하는 유형의 문항이다.

<표 7> 맥락에 대하여 평가 가능한 문항의 예

제12회 S-topik 초급 쓰기 46번
46. 다음 글을 읽고 ( )에 알맞은 말을 쓰십시오. (5점) 저희 할머니께서는 자주 편찮으셨습니다. 할머니께서 편찮으실 때마다 마음이 많이 아팠습니다. 그래서 저는 어릴 때부터 ( ). 지금은 의사가 되려고 대학교에서 열심히 공부하고 있습니다. 공부하는 것은 힘들지만 마음은 즐겁습니다. 꼭 좋은 의사가 되어서 아픈 사람들을 도와주겠습니다.
채점 기준 5점 : 의사가 되고 싶었습니다/ 싶었어요. 3점 : 미미한 형태적 오류(싶었씁니다 등) 1점 : ① 시제 사용의 오류(의사가 되고 싶습니다/ 싶어요) ② 3점 항목의 오류가 두 가지 중복되는 경우

그러나 이 주관식 문항에 대한 채점 기준을 보면 텍스트의 사용 맥락이 공식적인지 비공식적인지를 판단하여 표현할 수 있는 어말어미에 대한 것은 고려하고 있지 못하다. '의사가 되고 싶었습니다'라고 답을 쓴 경우는 텍스트 전체에 사용된 어말어미를 고려하여 이 텍스트가 공식적인 맥락을 가진다고 판단하고 이에 맞게 어말어미를 선택한 경우이다. 그러나 '의사가 되고 싶었어요'라고 쓴 경우는 의미적으로는 맞는 표현이나 이 텍스트의 맥락이 공식적이며 전후에 온 문장의 어말어미가 '-(스)ㅂ니다'로 끝났다는 것을 고려하지 못한 표현이다. 이 문항이 맥락 사용에 대한 것을 평가하고자 하는 의도가 있었다면 채점 기준에서 이 두 표현에 차등을 두었을 것이다. 만일 응시자 중 '의사 되고 싶었어요'라고 답을 쓴 경우가 있다면 이것은 어떻게 처리해야 하는가? 위 채점 기준대로라면 형태적인 오류나 시제 사용의 오류가 없으므로 5점을 부과해야 한다. 그러나

비격식적인 구어체의 표현에서는 조사의 생략이 빈번하게 이루어지지만 격식성을 갖춘 구어체나 문어체의 표현에서는 조사가 생략되면 문장이 어색해진다는 것을 이해하는 피험자라면 이와 같은 표현을 쓰지는 않을 것이다. 그러나 조사를 생략하는 경우의 답안에 대하여 채점 기준 제시가 없는 것으로 보아 맥락 사용에 대한 평가를 고려하지 않았다고 볼 수 있다. 이외 3문항 역시 채점 기준을 분석해 본 결과 이와 같은 문제가 반복되었다. 결국 평가 기준에서 제시한, 문맥 사용과 관련된 직접적인 평가 문항은 전혀 출제 되지 않았다고 볼 수 있다.

2) 간접적으로 평가한 문항

평가 문항에 나타난 텍스트에 어떤 맥락이 주로 사용되었는가를 분석해보면 이 시험의 출제 과정에 텍스트의 맥락에 대하여 어떤 고려를 하였는지 간접적으로 살펴볼 수 있다. 평가 문항의 텍스트가 사적 맥락을 가지는가, 공식적인 맥락을 가지는가를 살펴보면 다음과 같다.

<표 8> 평가 문항의 텍스트에 나타난 맥락의 빈도(%)

맥 락	초 급			중 급			고 급		
	쓰기	듣기	읽기	쓰기	듣기	읽기	쓰기	듣기	읽기
사적 맥락	82.3	83.3	56.7	50.0	53.3	13.3	40.0	26.7	16.7
공식적 맥락	11.8	10.0	43.3	43.7	46.7	86.7	53.3	73.3	83.3
정보 결여[11]	5.9	6.7	0.0	6.3	0.0	0.0	6.7	0.0	0.0
합 계	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

위 <표 8>의 결과를 통하여 제12회 TOPIK 평가 문항에 나타난 텍스트의 맥락과 관련하여 다음과 같이 해석할 수 있다.

11) 쓰기 영역의 주관식 문항의 경우, 쓰기 주제만 주어질 뿐 과제에 대한 맥락이 제공되어 있지 않다. 같은 주제로 글을 쓰더라도 어떤 목적으로, 누구(독자)를 대상으로, 어떤 형식의 글로 써야 하는가 등에 따라 실제 글쓰기의 모습은 달라진다.

① 초급에서 고급으로 갈수록 공식적 맥락 텍스트의 사용 비율이 높아졌다. 초급의 경우 쓰기, 듣기, 읽기 세 영역 모두 비공식적 맥락의 텍스트가 주로 사용되었다. 초급의 경우 공식적 맥락의 텍스트가 쓰기 영역에서 11.8%, 듣기 영역에서 10.0%, 읽기 영역에서 43.3%로 나타났다. 고급의 경우 쓰기, 듣기, 읽기 세 영역 모두 공식적 맥락의 텍스트가 주로 사용되었다. 비공식적 맥락의 텍스트가 쓰기 영역에서는 40.0%, 듣기 영역에서는 26.7%, 읽기 영역에서는 16.7%로 나타났다.

② 쓰기 영역에 맥락 정보가 결여된 문항이 나타났다. 이들은 모두 주관식 문항으로 아래 예와 같은 형태로 출제되었다.

<표 9> 맥락 정보가 결여된 쓰기 영역 주관식 문항의 예

<초급 쓰기>
* [47] 다음을 읽고 150~300자로 글을 쓰십시오.(30점)

47. 좋아하는 친구가 있습니까? 왜 그 친구를 좋아합니까? 좋아하는 친구에 대해서 소개해 보십시오.

* 쓰기 예

	저	는		따	뜻	한		우	유	를		마	십	니	다	.	그	러	면
몸	이		따	뜻	해	져	서		잠	이		잘		옵	니	다	.	여	러

이 주제로 친구에게 편지를 쓸 때나 일기를 쓸 때, 학교 과제로 친구를 소개하기 위한 글을 쓸 때 등은 구체적인 글의 형식과 문장의 형태는 차이가 있다. 그 이유는 독자와 글의 종류, 글을 쓰는 목적 등이 다르기 때문이다. 한국어 능력 시험 개요에 쓰기 영역 주관식 문항을 평가하기 위한 범주로 내용 및 과제 수행, 글의 전개 구조, 언어 사용, 사회언어학적 격식 등이 제시되어 있다. 이 중 사회언어학적 격식 범주의 평가 내용은 '작문의 장르적 특성 등에 맞추어 격식(register)의 사용이 적절한가를 평가'한다고 제시되어 있다. 그러

나 문항에 맥락과 관련된 정보가 결여되어 있으므로 이 부분을 정확하게 평가하기는 곤란하다. 피험자가 맥락에 따른 글쓰기를 할 수 있는가를 평가하고자 한다면 맥락을 구체적으로 제시해 주고 그 맥락에 비추어 올바른 표현을 하고 있는가를 평가해야 하는 것이 바람직할 것이다.

③ 같은 등급 내에서도 읽기 영역이 공식적 맥락 사용 비율이 높게 나타났다. 문자로 표현된 글이라고 해서 반드시 공식적 맥락만 있는 것은 아니다. 편지글, 메모, 수필, 일기 등 문자언어 의사소통에서도 얼마든지 비공식적인 맥락의 글이 있을 수 있다. 그러나 위 표에서와 같이 중급, 고급의 경우는 공식적 맥락의 빈도가 80% 이상으로 나타난 것은 비공식적 맥락의 다양성을 간과한 것으로 볼 수 있다.

다음으로 평가문항의 텍스트에 나타난 어말어미의 유형 빈도를 살펴보면 다음과 같다.

<표 10> 평가 문항의 텍스트에 나타난 어말어미의 유형 빈도(%)

어말어미		초 급			중 급			고 급		
		쓰기	듣기	읽기	쓰기	듣기	읽기	쓰기	듣기	읽기
경어체	-(스)ㅂ니다	41.1	6.7	100.0	12.5	43.3	6.7	0.0	46.7	0.0
	-아/어요	47.1	90.0	0.0	37.5	36.7	0.0	13.3	20.0	0.0
	-(스)ㅂ니다/-아/어요	5.9	3.3	0.0	0.0	10.0	0.0	20.0	23.3	0.0
반 말		0.0	0.0	0.0	0.0	10.0	0.0	0.0	10.0	0.0
평어체[12]		0.0	0.0	0.0	37.5	0.0	93.3	46.7	0.0	100.0
판단 불가		5.9	0.0	0.0	12.5	0.0	0.0	20.0	0.0	0.0
합 계		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

12) 이 글에서는 '-이다(plain style)'와 같이 상대를 높이지 않고 객관적으로 어떤 사건이나 사물을 진술할 때 어말어미를 사용하는 방식을 평어체라고 칭한다.

위 <표 10>의 결과는 다음과 같은 사항을 의미한다.

① 초급 읽기의 텍스트에는 경어체의 어말어미만을 사용하고 있고 고급 읽기 텍스트에는 평어체의 어말어미만을 사용하고 있다. 경어체든 평어체든 하나의 유형에 대한 어말어미만 텍스트에 나타나는 것은 구어와 문어를 맥락에 따라 구분하여 사용할 수 있는가를 살펴보고자 한 의도가 없음을 의미한다.
② 초급 듣기의 텍스트 중 경어체 어말어미인 '-아/어요'가 사용된 것이 90.0%라는 것은 초급 듣기 텍스트를 비공식적 맥락을 중심으로 구성하였다는 것을 의미한다.
③ 한 문항의 텍스트 내에서 '-(스)ㅂ니다'와 '-아/어요'가 함께 사용된 것은 맥락에 따라 격식성이나 유대, 효과 등을 고려하여 어말어미를 달리 사용할 수 있음을 보여준다.
④ 실제 대화 참여자의 관계에 따라 '-게', '-오' 화계가 다양하게 나타날 수 있으나 이를 반영하지 못했다.
⑤ 초급 텍스트에는 공적 사적 맥락에 따라 어말어미를 적절히 사용할 수 있는가를 측정하고자 하는 고려가 크게 없었다고 볼 수 있다.

이 밖에 텍스트와 담화의 종류가 평가 문항에 어떻게 나타나는 것을 살펴봄으로써 맥락 사용 여부에 대한 평가자의 의도를 간접적으로 알 수 있다. 평가 기준에는 '간단한 생활문과 실용문을 이해하고 구성할 수 있다(1급)'와 '뉴스, 신문 기사 중 형이한 내용을 이해할 수 있다(4급)'처럼 텍스트와 담화의 종류를 통하여 맥락을 제시한 부분이 있다. 즉 사적 맥락의 간단한 생활문과 실용문이 초급의 평가 문항에, 공식적 맥락의 뉴스와 신문 기사 등이 중급 이상의 평가 문항에 사용되는 것이 마땅하다. 초급에서 고급으로 갈수록 사적 맥락에서 공식적 맥락의 텍스트와 담화가 나타나야 하며 이들의 종류가 다양해지는 것이 바람직하다. 담화 혹은 텍스트의 종류가 어떤 경향을 보이는가를 분석해 보면 다음과 같다.

<표 11> 문항에 나타난 텍스트와 담화 종류의 빈도(%)

쓰 기				듣 기				읽 기			
텍스트 종류	초급	중급	고급	담화 종류	초급	중급	고급	텍스트 종류	초급	중급	고급
대화문	47.1	37.5	26.7	대화	86.6	53.3	43.3	설명문	3.3	43.3	56.7
*문장류[13]	17.6	12.5	13.3	인터뷰	0.0	10.0	13.3	수필	30.0	20.0	3.3
공고문, 메모	5.9	12.5	0.0	강의, 강연	0.0	6.7	13.3	표, 영수증 등	6.7	3.3	0.0
설명문	11.8	31.2	13.3	방송, 광고	6.7	13.3	6.7	안내문	16.7	10.0	0.0
감상문	0.0	0.0	13.3	보고	0.0	3.3	0.0	기사문	0.0	6.7	10.0
기사문	0.0	0.0	6.7	토론	0.0	0.0	6.7	논설문	0.0	10.0	0.0
안내문	0.0	0.0	6.7	뉴스	0.0	13.3	16.7	감상문	0.0	0.0	3.3
보고문	0.0	0.0	6.7	기타	6.7	0	0.0	광고문	10.0	6.7	0.0
논설문	0.0	0.0	13.3					옛날 이야기	0.0	0.0	6.7
기 타	17.6	6.3	0.0					신문 기사 제목	0.0	0.0	6.7
								규약문	0.0	0.0	3.3
								전기문	0.0	0.0	6.7
								대담문	0.0	0.0	3.3
								기타 (어휘, 구, 문장류)	33.3	0.0	0.0
합 계	100.0	100.0	100.0	합 계	100.0	100.0	100.0	합 계	100.0	100.0	100.0

위 <표 11>의 결과를 통하여 다음과 같은 사실을 알 수 있다.

① 초급의 경우 쓰기 영역에서는 대화문, 듣기 영역에서는 대화, 읽기 영역에서는 수필 등이 다른 종류보다 많이 사용된 것은 결국 1급

13) '문장류'는 텍스트의 종류가 될 수 없어서 별도의 처리를 하였다. 이들은 단문이나 단문 두 개를 연결하도록 문항이 구성된 것이다. 이 경우 텍스트의 종류를 판단하기 어려우나 분류를 위하여 일단 '문장류'로 표시해 두었다.

의 지표인 간단한 생활문과 실용문을 이해하고 구성할 수 있는가를 측정하기 위한 것으로 판단된다. 또 고급으로 갈수록 이들 텍스트와 담화 종류에 대한 빈도는 낮아지고 상대적으로 공식적인 맥락을 가진 다양한 텍스트와 담화 종류가 많이 사용되는 것을 볼 수 있는데 이 역시 평가 기준에 따라 숙달도에 따른 맥락 사용 능력을 다양하게 측정하고 있다는 측면에서 바람직하다고 볼 수 있다.

② 각 영역마다 기타로 분류된 것이 존재한다. 쓰기 영역에 텍스트의 종류를 분류할 수 없는 문항은 앞서 쓰기 영역에서 맥락에 대한 정보를 제공하지 않았던 문항과 동일하다. 또 듣기 영역의 경우 음운 식별적 듣기 능력을 측정하는 아래 문항의 경우, 주어진 한 문장을 통하여 담화의 종류를 규정짓기는 어렵다.

<표 12> 담화의 종류 파악이 어려운 듣기 영역 문항의 예

<초급 듣기>
[1~2] 다음을 듣고 <보기>와 같이 ( )에 알맞은 것을 고르십시오.
<보기> 생략
1. (3점) 이게 ( )이에요.
① 밤 ② 밥 ③ 발 ④ 방
2. (4점) ( )이 있어요.
① 달 ② 돌 ③ 둘 ④ 들

이 문항을 사적 맥락의 대화에서 음운을 정확하게 식별하고 이에 제대로 반응하는 것을 측정하도록 하는 것이 실제 피험자의 한국어 의사소통 능력을 평가하는 방법으로 타당하다. 맥락 정보가 결여되어 있고 담화의 종류조차 파악이 되지 않는 텍스트는 결국 의사소통을 전제로 한 것이라고 보기 어려우며, 이러한 텍스트가 문항에 사용된다는 것은 궁극적으로 한국어 능력 시험이 평가 기준에 따라 한국어 의사소통 능력을 평가해야 한다는 점에서 본다면 내용타당도를 떨어뜨리는 요인이 될 수 있다. 하나의 담화로서 맥락을 가진 듣기 문항으로 기술하기 위해서 위의 듣기 문항 1은 다음과 같이 수정할 필요가 있다.

1-1. 다음 물음에 맞는 대답을 고르십시오.
(3점) 가 : 이게 ( )이에요. [밤]
나: _____
① 참 맛있겠어요. ② 뜨거워 보여요.
③ 어두워요. ④ 참 넓어 보여요.

위 1-1 문항의 경우는 음운 식별적 듣기 능력을 측정하고자 하는 평가 목표는 같지만 이를 의사소통 맥락 속에서 평가할 수 있도록 하고 있다. 정확하게 음운 식별을 한 경우라면 전체 대화 상황에 맞게 적절한 반응을 하게 된다. 또 읽기 영역의 초급에 기타로 분류된 것은 어휘, 구, 문장 두 개 등을 읽기 문항에 사용한 경우이다. 이들 역시 맥락 사용에 대한 정보를 제공하고 있지 못하다는 점에서 문제가 된다. 또 어휘의 의미를 파악하는 문항은 '어휘·문법' 영역에서 측정하는 것이 마땅하다.

③ 쓰기 영역의 경우 문장류의 형태의 문항이 초, 중, 고급에 걸쳐 나타난다. 이들은 단문 두 개를 연결하게 하거나 제시된 표현을 이용하여 한 문장을 쓰게 하는 문항으로 아래와 같은 것들이다. 그러나 이들 문항은 쓰기 맥락이 제대로 제공이 되어 있지 않은 상태에서 문장 쓰기를 하도록 기술되어 있다는 점에서 맥락 사용과 관련된 쓰기 능력에 대하여 평가하고자 하지는 않았음을 알 수 있다.

<표 13> 문장류 형태의 쓰기 문항의 예

<초급 쓰기 영역>
* 보기와 같이 두 문장을 바르게 연결한 것을 고르십시오. <보기 생략>
36. 목이 마릅니다. 물 좀 주십시오.
① 목이 마른데 물 좀 주십시오. ② 목이 마르거나 물 좀 주십시오.
③ 목이 마르지만 물 좀 주십시오. ④ 목이 마르려고 물 좀 주십시오.

<중급 쓰기 영역>
[41~42] 제시된 표현을 모두 사용해 한 문장으로 쓰십시오.(각6점)
41. 한 달이 넘다/연락이 안 되다/걱정이다

<고급 쓰기 영역>
[41~42] 다음 글감에 대해서 제시된 표현을 사용해 한 문장으로 만드십시오.(40자~50자)
41. <태양열 사용의 효율성> 태양열/에너지 절약/자연보호/일석이조

두 단문의 의미에 맞게 연결어미를 사용할 수 있는가 하는 것은 현 한국어 능력 시험의 체제에서는 '어휘·문법' 영역에서 얼마든지 평가가 가능하므로 쓰기 영역에서는 실제 텍스트를 구성할 수 있는가를 측정하는 것이 바람직하다.

이상과 같이 한국어 능력 시험의 평가 문항에 나타난 텍스트의 맥락과 관련하여 네 측면에서 분석해 본 결과, 평가 기준에 제시하였던 것처럼 공식적 상황과 비공식적 상황에서의 언어 구분 능력을 직접적으로 평가하고 있지 않으며 등급별로 제시되었던 맥락에 대한 기준이 평가 문항에 제대로 반영되어 있지 않았다. 이를 통하여 한국어 능력 시험의 평가 문항이 평가 기준에 제시된 맥락 관련 내용을 적절히 반영하였다고 보기는 어려우므로 이 부분과 관련된 내용타당도는 낮다고 볼 수 있다.

4. 내용과 텍스트 형태 범주의 내용타당도

4.1. 내용 범주

텍스트의 내용(content) 범주에서는 평가 문항에 사용된 텍스트의 화제나 주제 등이 등급에 따라 어떤 경향을 띠는가를 살필 수 있다. 동일한 구조로 된 문장이라 하더라도 어떤 화제나 주제에 대하여 표현하는가에 따라 그 문장을 구성하는 어휘가 달라진다. 뿐만 아니라 텍스트의 화제나 주제는 그 언어에 내재되어 있는 고유한 문화와 가치관이 담겨져 있으므로 어

느 정도의 화제나 주제에 대하여 이해하고 표현할 수 있는가를 통하여 목표언어의 문화에 대한 이해 정도를 가늠할 수 있기도 하다. 한국어 능력 시험의 등급별 평가 기준에 내용과 관련된 부분을 정리하면 다음과 같다.

<표 14> 내용 범주의 등급별 수행 정도

내 용	초 급		중 급		고 급	
	1급	2급	3급	4급	5급	6급
친숙한 (일상적) 화제	✓	✓				
친숙한 사회적, 구체적 소재			✓			
사회적, 추상적 소재[14]				✓		
사회, 문화적인 내용				✓	✓	
정치, 경제, 사회, 문화 전반에 걸쳐 친숙하지 않은 소재					✓	✓✓

위 표에 나타나듯이 초급에서 친숙한 일반 화제에 대하여 이해하고 표현할 수 있어야 하며 숙달도가 높아질수록 사회적, 추상적 소재를 다룰 수 있어야 한다. 중급에서는 사회, 문화적인 내용을, 고급의 경우는 정치, 경제 분야의 내용을 이해하고 표현할 수 있어야 한다. 또 같은 내용 속성을 가진 소재도 친숙성과 구체성의 정도에 따라 등급에 차이를 두고 있다. 즉 친숙한 소재보다는 친숙하지 않은 소재, 구체적인 소재보다는 추상적인 소재 등을 다루는 것을 더 숙달도가 높은 것으로 정해두고 있다. 텍스트에 나타난 내용을 살피기 위하여, 내용 범주의 특성을 아래 <표 15>와 같이 세분하였다. 이는 같은 등급의 평가 기준에 속한 내용이더라도 화제나 소재가 영역 구분이 명확한 것은 나누는 것이 문항에 나타난 텍스트의 경향을 파악하는 데 도움이 되기 때문이다.

14) 4급의 평가 기준에 '뉴스, 신문 기사 중 평이한 내용을 이해할 수 있다'는 표현은 결국 '사회적 소재에 대하여 이해할 수 있다'와 같은 의미이므로 이 항목에 포괄하였다.

<표 15> 평가 문항에 나타난 텍스트 내용의 등급별 빈도(%)

텍스트 내용	초 급			중 급			고 급		
	쓰기	듣기	읽기	쓰기	듣기	읽기	쓰기	듣기	읽기
친숙한 일상적 화제	94.1	86.6	90.0	81.3	33.3	13.3	6.7	16.7	6.7
친숙한 사회적, 구체적 소재	0.0	6.7	0.0	18.7	13.3	20.0	6.7	6.7	0.0
문 화	0.0	0.0	0.0	0.0	16.7	20.0	20.0	10.0	50.0
정 치	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0
경 제	0.0	0.0	0.0	0.0	6.7	0.0	6.7	10.0	20.0
기 타	0.0	0.0	0.0	0.0	6.7	6.7	13.3	13.3	10.0
판단 불가	5.9	6.7	10.0	0.0	0.0	0.0	0.0	0.0	0.0
합 계	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

한국어 능력 시험 평가문항에 나타난 텍스트의 내용을 등급별로 분석한 <표 15>를 통하여 다음과 같은 결론을 얻을 수 있다.

① 초급 듣기 텍스트에 '친숙한 사회적, 구체적 소재'가, 중급 듣기 문항에 경제에 대한 소재가 각각 6.7%씩 사용되었다. 평가 기준에 초급의 경우 사적이고 친숙한 화제를, 중급의 경우 친숙한 구체적인 소재는 물론 사회적 소재를 이해하고 표현할 수 있다고 제시되어 있다. 그러므로 초급 듣기 텍스트에 '사회적, 구체적 소재'가 나타난 것은 내용타당도에 문제가 있음을 의미한다.

② 초급 쓰기, 듣기, 읽기 영역의 문항에 텍스트의 내용에 대하여 판단하기 어려운 것이 있었다. 쓰기 영역의 경우 주어진 단어를 가지고 어순을 맞추어 단문을 만들게 하는 문항이, 듣기 영역의 경우 앞서 언급되었던 음운 식별적 듣기 능력을 측정하는 문항이, 읽기 영역의 경우 어휘, 구, 단문의 의미에 맞는 픽토그램이나 문장을 찾게 하는 문항 등이다. 이들 문항의 특징은 주로 단문 이하의 단위로 비의사소통적 맥락 하에 어휘나 구, 단문의 의미를 이해하고 표현하게 한다는 것이다. 초급이라 하더라도 '사적이고 친숙한 화

제에 관해 문장(1급)이나 문단(2급) 단위로 이해하고 사용할 수 있어야 한다'는 평가 기준에 비추어 본다면 이러한 문항은 이 시험의 내용타당도를 높이기 위해서라도 지양되어야 한다.

③ 중급 쓰기 영역의 평가문항에 '친숙한 일상적 화제'를 다룬 텍스트가 81.3%나 사용되었다는 것이다. 초급과 중급이 내용 범주에서 변별되는 부분은 사회적, 추상적 소재를 이해하고 표현할 수 있다는 것이다. 그러나 이처럼 중급 쓰기 영역의 텍스트 내용에 정작 중급의 특성을 나타내는 화제나 소재는 18.7% 정도밖에 사용되지 않은 것은 문제가 많다.

④ 정치에 대한 화제나 소재를 다룬 것은 고급 듣기 평가 문항에 3.3% 사용된 것 이외에 없었다. 고급이 중급과 비교하였을 때 내용 범주에서 변별적인 부분 중의 하나가 정치, 경제적 주제나 소재가 다루어진다는 것이다. 그러나 고급 듣기 영역에서 전체 텍스트 중 겨우 3.3%만 정치와 관련된 내용이 나타났으며 쓰기와 읽기 영역에서는 아예 다루어지지 않았다는 것은 등급별로 텍스트에 어떠한 내용 특성이 나타나야 한다는 것에 대하여 별 고려가 없었음을 의미한다.

⑤ 중급 듣기와 읽기 영역과 고급 쓰기, 듣기, 읽기 영역에 걸쳐 '기타'라고 분류된 텍스트의 내용은 과학 분야와 관련된 것이었다. 중급과 고급 전체 영역에 과학 분야 관련 내용의 텍스트가 사용된 빈도가 정치, 경제 분야 관련 내용 텍스트가 사용된 빈도보다 높다는 것은 평가 기준에 제시된 텍스트의 내용 설정에 대하여 근본적인 재고가 필요하다는 것을 의미한다.

이상의 해석들을 통하여 살펴보면 내용 범주에 대해서는 한국어 능력 시험의 평가 문항이 평가 기준에 비추어 내용타당도가 현저히 낮다고 볼 수 있다.

4.2. 텍스트 형태 범주

텍스트 형태 범주에서는 한국어 능력 시험의 문항이 한국어 사용에서 표현되거나 이해되는 텍스트의 양에 대한 것을 적절히 평가하고 있는가를 살필 수 있다. 텍스트의 양은 내용을 생성하고 이해하는 데 얼마나 유창성한가와도 관련이 되므로 피험자의 숙달도를 평가하는 중요한 요인이 될 수 있다. 현재 한국어 능력 시험의 평가 기준에서는 텍스트의 형태에 대하여 등급별로 다음과 같이 제시하고 있다.

<표 16> 텍스트 형태 범주의 등급별 수행 정도

텍스트 형태	초 급		중 급		고 급	
	1급	2급	3급	4급	5급	6급
간단한 문장	✓					
문단 단위		✓	✓			

위 <표 16>에 나타나듯이 1급에서는 간단한 문장을 이해하고 생성할 수 있어야 하며 2급부터는 문단 단위의 이해와 표현이 가능한가를 평가 기준으로 삼고 있다. 평가 문항에 나타난 텍스트가 어느 정도의 양을 가지는지를 살펴보기 위하여 텍스트 형태를 어휘・구, 한 문장, 두 문장, 세 문장, 한 문단, 두 문단 이상 등과 같이 세분하여 분석하였다. 이는 평가 기준에 제시되어 있는 '간단한 문장'과 '문단 단위'라는 표현으로는 텍스트가 일반적으로 어느 정도의 양을 가지는지를 비교해 보는 데 한계를 가지기 때문이다. 평가 문항에 나타난 텍스트의 형태 빈도를 살펴보면 다음과 같다.

<표 17> 텍스트의 형태 빈도(%)

쓰 기				듣 기				읽 기			
텍스트 형태	초급	중급	고급	담화 형태	초급	중급	고급	텍스트 형태	초급	중급	고급
어휘, 구	0.0	0.0	0.0	어휘, 구	0.0	0.0	0.0	어휘, 구	13.3	0.0	3.3
한 문장	11.8	12.5	13.3	한 문장	6.7	0.0	0.0	한 문장	6.7	0.0	3.3

쓰 기				듣 기				읽 기			
텍스트 형태	초급	중급	고급	담화 형태	초급	중급	고급	텍스트 형태	초급	중급	고급
두 문장	41.2	37.5	26.7	두 문장	40.0	6.7	10.0	두 문장	23.3	0.0	0.0
세 문장	35.2	0.0	13.3	세 문장	46.6	16.7	10.0	세 문장	30.0	6.7	3.3
한 문단	5.9	43.7	33.4	한 문단	6.7	76.6	80.0	한 문단	26.7	93.3	66.7
두 문단 이상	5.9	6.3	13.3	두 문단 이상	0.0	0.0	0.0	두 문단 이상	0.0	0.0	23.4
합계	100.0	100.0	100.0	합계	100.0	100.0	100.0	합계	100.0	100.0	100.0

위 <표 17>의 결과를 바탕으로 텍스트의 형태 범주에 대한 내용타당도의 정도를 분석하면 다음과 같다.

① 쓰기 영역의 경우 한 문장 형태의 텍스트가 중, 고급에 나타나는 것은 문제가 있다. 이들은 앞서 맥락 사용과 관련하여 <표 11>에 '문장류 형태의 쓰기 문항'에 중, 고급 쓰기 영역에 해당하는 내용이다. 즉 중, 고급 쓰기 영역에 제시된 표현을 이용하여 한 문장을 쓰게 하는 문항으로, 2급부터 문단 단위 이상의 이해와 표현 여부를 평가 기준으로 제시하고 있다는 것을 보면 중급과 고급에 이 같은 평가 문항은 내용타당성이 결여되었다고 볼 수 있다.

② 듣기 영역의 경우 거의 대부분의 담화 형태가 두 문장 이상의 크기를 가진다. 듣기 영역에서 두 사람 이상의 대화참여자를 가지는 담화의 경우는 최소대응쌍이 나타나야 하는 구조라서 두 문장 이상의 담화 형태로 주어질 수밖에 없다. 이런 측면에서 보면 듣기 영역 초급에 나타난 한 문장의 담화 형태는 의사소통을 전제로 하지 않았다는 점에서 재고가 필요하다. 이들 문항은 앞서 언급한 바 있는 음운의 식별적 듣기를 평가하기 위한 문항으로, 간단한 문장(단문)이 대응쌍으로 오는 담화 형태로 수정되는 것이 바람직하다.

③ 읽기 영역의 초급, 고급에 어휘, 구 형태의 텍스트가 나타났다. 평가 기준에서 사용하는 텍스트의 최소 단위가 초급 단계에 나타나는 '간단한 문장'이므로 어휘·구 형태의 텍스트로 문항을 출제하

는 것은 현재의 평가 기준에 비추어 본다면 내용타당도를 떨어뜨리는 요인이 된다. 그리고 어휘·구 형태의 텍스트가 읽기 영역의 고급 문항에 나타나는 것은 이미 2급 이후에 문단 단위의 텍스트 형태는 거의 사용되지 않는 것을 기준으로 보면 이 역시 내용타당도에 부정적 영향을 미친다. 그러나 실제 이 문항을 분석해 보면 관용 표현이나 은유 표현이 나타난 신문 기사의 제목을 보고 내용을 추측해 보게 하는 것으로 고급의 읽기 기능을 살펴보는 방법으로 타당하다. 결국 평가 기준을 준거로 본다면 문제가 되는 부분이 실제 그 영역의 기능 수행 측면에서 보았을 때는 문제가 되지 않는다는 것은 한국어 능력 시험의 평가 기준 자체의 타당성에 대한 전면적인 재고가 필요함을 의미한다.

④ 듣기 영역의 경우 고급에서도 두 문단 이상의 긴 텍스트 형태가 주어지지 않는다. 쓰기 영역의 경우 주관식 문항에서 긴 텍스트를 구성할 수 있는지 평가하며 읽기 영역의 경우도 고급에 두 문단 이상의 긴 텍스트가 주어져서 이에 대한 이해력을 평가하고 있다. 그러나 듣기 영역의 경우는 긴 담화텍스트를 듣고 이해하는 문항이 없다. 강의나 강연의 경우도 한 문단 분량의 텍스트로 정보를 제한하고 있다. 그러나 한국어 능력 시험의 결과를 한국 대학을 입학할 때 평가 자료로 사용하고자 하는 목적을 가지고 있으며, 평가 기준에 고급에 전문 분야의 연구 기능을 가지고 있어야 함을 제시한 것으로 본다면 강의나 강연에서 정보의 양이 많은 긴 담화를 듣고 이해하는 기능을 다루는 것이 바람직하다.

이처럼 한국어 능력 시험의 문항에 텍스트의 형태에 대하여 어떻게 측정하고 있는가를 살펴보면 쓰기 영역의 경우 숙달도에 맞지 않는 텍스트의 형태가 중, 고급에 사용된 것, 듣기 영역의 경우 초급에 단문 형태의 일방적 말하기 텍스트가 사용된 것, 읽기 영역의 초급, 고급에 어휘, 구 형태의 텍스트가 사용된 것 등은 평가 기준에 비추어 보았을 때 내용타당도를 떨어뜨리는 부분이다.

5. 맺음말

이상과 같이 한국어 능력 시험의 평가문항이 어느 정도의 내용타당도를 가지는가를 한국어 능력 시험의 요강에 제시되어 있는 평가 기준을 준거로 기능, 맥락, 내용, 텍스트 형태 등의 범주로 나누어 살펴보았다. 이 글에서는 이해와 표현의 의사소통 상황에서 한국어가 실제 사용되는 것을 통하여 한국어 사용 능력을 평가한다는 측면에서 전체 평가 영역 중 '어휘·문법' 영역을 제외하고 '쓰기', '듣기', '읽기' 세 영역에 대한 내용타당도를 살펴 본 결과, 이들 네 범주에 대하여 평가 기준에 제시되었던 것이 문항에 제대로 반영되지 않았음을 확인할 수 있었다. 즉, 한국어 능력 시험의 평가 문항이 이 평가를 주관하는 기관에서 측정하고자 제시하였던 평가 기준에 대하여 대표성을 띤다고 보기 어렵다.

이같이 한국어 능력 시험의 평가 문항이 내용타당도가 높지 않은 원인으로 다음 두 가지를 생각할 수 있다. 첫째, 평가 문항을 작성할 때 출제자가 평가 기준을 충분히 고려하지 않았다는 것이다. 한국어 능력 시험의 핵심은 숙달도에 따라 등급화가 이루어져야 하므로 난이도가 뚜렷이 구분되어야 한다는 것이다. 언어 능력 시험에서 문항 간에 난이도의 차이를 둘 수 있는 방법은 여러 가지가 될 수 있으나 한국어 능력 시험이 평가 도구로서의 내용타당도를 높이기 위해서는 이 시험이 공인한 평가 기준에 따라 등급별 난이도가 조정되어야 한다. 그러므로 평가 문항을 작성할 때는 그 평가 도구가 어떠한 평가 기준을 가지고 있으며 그 내용이 어떻게 구체화될 수 있는가에 대한 세밀한 분석이 선행되어야 한다. 둘째, 한국어 능력 시험의 평가 기준에 설정된 내용이 타당하지 않을 수 있다는 것이다. 이 같이 평가 기준 자체의 타당성에 대하여 의문을 가지는 이유는 앞서 언급을 하였던 것처럼 이 평가 기준이 평가 기준으로서 구체적이지 않으며, 체계적이지도 않다는 점 등이다. 또 과학 분야의 텍스트가 중급과 고급 평가 문항에 걸쳐 나타나는 것처럼 해당 등급의 한국어 사용 능력을 평가하기 위하여 널리 다루고 있고 상식적으로도 수긍이 가는 부

분도 실제 평가 기준에는 제시가 되어 있지 않는 것이다.

일반적으로 내용타당도 연구가 안고 있는 가장 큰 단점은 내용타당도 분석에 연구자의 주관이 개입하기가 쉬우므로 연구 결과가 객관성을 확보하기 어렵다는 것이다. 그럼에도 불구하고 이 연구에서 한국어 능력 시험의 내용타당도 분석을 시도한 것은 평가도구에 대한 여러 측면의 분석 중 평가에서 원래 측정하고자 하였던 것을 정말 제대로 측정하고 있는가에 대한 연구가 기반이 되어야 이 평가 도구와 관련된 다른 연구 결과에 의의를 부여할지 말지가 결정되기 때문이다. 어떤 평가 도구가 신뢰도[15], 실용도, 난이도 등에 문제가 없다 하더라도 내용타당도가 낮다면 그 평가 도구는 재고되어야 한다. 한국어 능력 평가가 공인된 평가도구로서 내용타당도를 높이기 위해서는 위 두 가능성 모두를 열어두고 이들 문제점을 개선할 수 있는 연구가 필요하다.*

15) 신뢰도는 타당도의 필수조건이지 충분조건이 아니다. 예를 들어 한국어 학습자의 한국어 사용 능력을 측정하기 위하여 어휘와 문법에 대해서 평가를 한 뒤 이 평가도구의 신뢰도가 0.99라고 하여도, 이 평가도구는 신뢰도는 높지만 타당도가 낮기 때문에 이를 통하여 측정된 결과를 학습자의 한국어 사용 능력이라고 하는 것은 부당하다.

* 본 논문은 2008. 2. 28. 투고되었으며, 2008. 3. 6. 심사가 시작되어 2008. 3. 24. 심사가 종료되었음.

▣ 참고문헌

김영아(1996), 외국어로서의 한국어 능력 평가 연구, 고려대학교 박사학위 논문.

김유정(2006), 한국어 능력 시험의 난이도 분석 연구, 한국어교육17-1, 국제한국어교육학회.

김정숙 외(2005ㄱ), 한국어 능력 시험의 개선 방안 연구(I)－등급 부여 방식을 중심으로, 한국어교육16-1, 국제한국어교육학회.

김정숙 외(2005ㄴ), 한국어 능력 시험의 개선 방안 연구(II)－평가 문항 유형을 중심으로, 한국어교육16-2, 국제한국어교육학회.

김하수 외(1999), "범용 한국어 교육 교재(초급)의 개발" 사업 보고서, 한국어세계화재단.

민병곤(2005), 한국어 능력 시험의 운영 현황 및 과제, 한국어교육16-3, 국제한국어교육학회.

박영순 편(2002), 21세기 한국어교육학의 현황과 과제, 한국문화사.

박영순(2004), 외국어로서의 한국어 교육론(개고판), 월인.

성태제(2002), 현대교육평가, 학지사.

양태식 외(2000), 2000년도 한국어 세계화 추진 기반 구축 사업 보고서－한국어 중급 교수요목 개발 분과, 한국어세계화재단.

이해영(2005), 한국어 이해 능력 평가의 원리 및 실제, 한국어교육16-3, 국제한국어교육학회.

정동빈 외(1991), 영어교육론, 한신문화사.

조항록(2006), 한국어 능력 평가 체계의 현황과 과제, 한국어교육17-1, 국제한국어교육학회.

한국교육과정평가원(2005), 2005년도 제9회 한국어 능력 시험 요강.

한재영 외(2005), 한국어 교수법, 태학사.

황정규(2002), 학교학습과 교육평가, 교육과학사.

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

Brown, H, D.(2001), *Teaching by Principles: An Interactive Approach to Language Pedagogy*, Longman.

Cohen, A, D(1980), *Testing Language Ability in the Classroom*, 정명우・서천수(1990) 역, 새로운 언어능력 테스팅, 한신문화사.

<초록>

한국어 능력 시험 평가 문항의 내용타당도 분석

전은주

이 논문은 한국어 능력 시험(TOPIK : Test of Proficiency in Korean)의 평가 문항이 한국어 학습자의 숙달도를 평가함에 있어서 어느 정도의 내용타당도를 지니는가를 분석하였다. 이 논문에서는 의사소통 상황에서 한국어가 실제 사용되는 것을 통하여 한국어 사용 능력을 평가한다는 측면에서 한국어 능력 시험의 전체 평가 영역 중 '어휘·문법' 영역을 제외하고 '쓰기', '듣기', '읽기' 세 영역에 대한 내용타당도를 살펴보았다. 또 문항의 내용타당도를 분석하기 위한 범주를 한국어 능력 시험의 요강에 제시되어 있는 평가 기준을 준거로, 기능, 맥락, 내용, 텍스트 형태 등으로 설정하였다.

그 결과, 이들 네 범주에 대하여 평가 기준에 제시되었던 것이 각 영역별로 평가 문항에 제대로 반영되지 않았음을 확인할 수 있었다. 따라서 한국어 능력 시험의 평가 문항이 이 평가를 주관하는 기관에서 측정하고자 제시하였던 평가 기준에 대하여 대표성을 띤다고 보기 어렵다.

이같이 한국어 능력 시험의 평가 문항이 내용타당도가 높지 않은 원인으로 다음 두 가지 를 생각할 수 있다. 첫째, 평가 문항을 작성할 때 출제자가 평가 기준을 충분히 고려하지 않았다는 것이다. 둘째, 한국어 능력 시험의 평가 기준에 설정된 내용이 타당하지 않을 수 있다는 것이다. 어떤 평가 도구가 신뢰도 실용도, 난이도 등에 문제가 없다 하더라도 내용타당도가 낮다면 그 평가 도구는 재고되어야 한다. 한국어 능력 평가가 공인된 평가도구로서 내용타당도를 높이기 위해서는 위 두 가능성 모두를 열어두고 이들 문제점을 개선할 수 있는 연구가 필요하다.

【핵심어】 내용타당도, 한국어 능력 시험(TOPIK), 기능, 맥락, 내용, 텍스트 형태

<Abstract>

Analysis for Contents Validity of the TOPIK Questionnaires

Jeon, Eun-joo

This work analyzed the contents validity of the questionnaires of Test of Proficiency in Korean(TOPIK) for assessing the proficiency of the Korean learners. This article evaluated three parts of TOPIK-which are writing, listening, and reading-vocabulary and grammar excluded, on the premise that Korean proficiency is evaluated through actual usage of Korean in the communicative situation. We picked up four categories-function, context, contents and text form-for analyzing the contents validity, based on the assessment criteria presented at syllabus for TOPIK. In result, it was shown that the assessment criteria was not reflected by the questionnaires for each category. Therefore, it is not likely that the questionnaires for TOPIK are representative of the assessment criteria presented by the institute in charge of the test.

Two possible explanations exists for reasons why the questionnaires lack the contents validity. First, the provider of the questionnaire did not take account of the assessment criteria. Second, the assessment criteria, per se, are not appropriate in validity. If an assessment tool lacks contents validity, although it fulfills credibility, practicality and difficulty, then it should be reconsidered as a proper assessment tool. In order that the TOPIK could enhance the contents validity as an official assessment tool, a research is warranted for solving above issues with an open mind to those two possibilities.

【Key words】 contents validity, Test of Proficiency in Korean(TOPIK), function, context, contents, text form

【토론문】

“한국어 능력 시험 평가 문항의 내용타당도 분석”에 대한 토론문

지현숙(배재대)

어떤 언어 능력 시험이 ‘타당하다’는 것은, 시험 점수를 통해 예측할 수 있는 언어 능력의 근거가 확보되어 있다는 의미이다. 어떤 시험이든지 ‘근거’를 바탕으로 하여 타당성을 가짐으로써 시험을 보는 수험자도 그 시험을 믿고 볼 수 있고, 시험 결과를 사용하는 관련자(stakeholders)도 그 시험을 공신력 있게 받아들일 수가 있다. 하물며 한국어능력시험(TOPIK)과 같은 대규모의 고부담 평가 도구인 경우에는 튼튼한 근거가 뒷받침되는 타당성의 확보가 기본적이며 필수적이다.

우리가 어떤 시험에 대해 가지는 일차적인 관심은 그 ‘시험을 보는 사람의 능력을 어느 정도로 정확하게 반영하는가’일 것이다. 따라서 “시험 개발자는 시험의 타당성을 입증하기 위한 명확한 근거를 제공할 의무가 있다. 이는 법정에서 피의자의 변호인의 역할과 유사하다(Weir, 2005)”.

전은주 선생님께서는 제12회 TOPIK의 쓰기, 듣기, 읽기 영역의 문항 분석을 통해서 한국어능력시험의 내용 타당도를 분석하였다. 이 시험이 시행된 지 10년을 훌쩍 넘었고 그간 적지 않은 연구들에서 이 시험이 안고 있는 문제들을 다루고 지적해 왔으나 과연 ‘TOPIK은 타당한가?’를 묻는 논의는 아마도 이번이 처음일 것으로 생각이 되고 그런 점에서 인 연구는 충분히 논쟁적이라고 할 수 있겠다.

한국어 교육에서 평가 연구에 발전이 있기 위해서는 무엇보다도 시험의 근거가 되는 타당성에 늘 관심을 가져야 하고 타당하다는 증거를 확보하는 작업이 면밀하게 이루어져야 한다. 그런데 “타당성은 측정하고자 하는 시험의 ‘방법(method)’과 ‘구인(construct)’에 가장 큰 영향을 받는다

(Weir, 2005 : 7)". 전은주 선생님께서 수행하신 연구는 '내용 타당도'에 집중하여 TOPIK이 명시한 평가 목표에 평가 영역별 문항들이 얼마나 타당하게 출제되었는가를 분석하는 것이 주를 이루고 있다. 그러나 사실 그 이전에 TOPIK 듣기, 읽기, 쓰기, 어휘/문법 영역 각각의 구인과 방법은 무엇인가에 대한 분석과 검토, 문제 제기가 이루어져야 본고의 내용 타당성 논의가 더욱더 빛을 발할 수 있었다고 본다. 예컨대 쓰기 시험의 구인이나 평가 방식이 이미 적절치 못한데 내용 타당도가 높을 리는 만무하기 때문이다.[1)]

둘째, 발표자께서는 '기능, 맥락, 내용, 텍스트 형태' 등의 모든 범주에서 TOPIK이 내용 타당도가 낮음을 밝혀내셨다. 이에 대한 증거는 충분히 제시하였으므로 토론자는 전은주 선생님의 논의에 대해 충분히 공감이 간다. 그러나 선생님께서 "한국어 능력 시험이 평가 도구로서의 내용 타당도를 높이기 위해서는 공인한 평가 기준에 따라 등급별 난이도가 조정되어야 한다."고 결론을 내린 부분에 있어서, 일반적으로 내용 타당도란 "평가 도구가 측정하려는 목적에 맞는 내용을 얼마나 대표성 있게 또 광범위하게 포함하였는가(이완기, 2007 : 40)"에 관한 개념으로 이해한다면, TOPIK의 내용 타당도 부재가 어떻게 시험의 난이도 조정으로 연결이 되는지 선생님의 설명이 있었으면 한다. 또한 두 번째 결론에서 "TOPIK의 평가 기준이 구체적이지 않으며, 체계적이지도 않다"는 점을 들어 "한국어 능력 시험의 평가 기준에 설정된 내용이 타당하지 않을 수 있다"고 보는 논의는 어떻게 연결되는 것인지에 대해서도 조금 더 상세한 설명을 듣고 싶다.

셋째, 토론자는 언어 시험이 타당성을 가지기 위해서는 바람직한 시험 생산 시스템을 가져야 한다고 생각한다. 이 연구가 한국어 교육의 평가 연구에 실제적으로 보탬을 주기 위해서는 TOPIK의 생산 시스템을 혁신

---

1) 일각에서는 "내용 타당도는 부분적으로 평가 문항 내용의 다양성과 종류에 좌우되기 때문에 평가 문항의 균질성을 전제로 하는 평가 문항 신뢰도와는 상충하는 관계에 있다.(이완기, 2007 : 53)"고 보기도 한다. 그러나 토론자는 타당도를 신뢰도와 동등한 요건으로 보기보다는 '타당도는 신뢰도를 포함하는 상위 개념'이라고 보는 Weir(2005)의 견해에 동의한다.

할 수 있는 방안이 제안되어야 하지 않을까 싶다. 토론자의 조금은 과한 욕심일 수도 있겠으나, TOPIK이 타당성 자체가 의심이 되는 시험이라면, 타당성을 확보하기 위해서 어떻게 구인과 시험 방법이 구안되고 문항이 개발되어야 할 것인가가 결국 해결해야 할 당면 문제가 되기 때문이다.