개정 TEPS 구인타당도 검증

임의진[1†] · 전희성[1] · 윤정민[1] · 민선홍[1]
[1]서울대학교

Evidence of Construct Validity for New TEPS

Euijin Lim[1†], Heesung Jun[1], Jungmin Yun[1], and Sunhong Min[1]
[1]Seoul National University

ABSTRACT

The purpose of this study is to obtain evidence of construct validity for the revised TEPS. Multiple data sets were obtained from the operational TEPS administrations before and after the revision and four pilot tests during the revision process and used for correlation analysis and confirmatory factor analysis. Inter-section correlation coefficients and standardized factor loadings were compared across forms, and the model-data fit indices were examined for each form. Based on the results, it was found that the original TEPS and the revised TEPS measured very similar or virtually the same construct. Even though the revision introduced major changes to the TEPS, the underlying factor structure was the same for all the forms examined in this study.

Keywords: TEPS, test revision, construct validity, confirmatory factor analysis

1. 서 론

TEPS는 '한국인의 실질적 영어활용능력'을 측정하기 위해 개발되어 1999년에 처음 시행된 이래로 20년간 다양한 기관에서 활용되어왔다. 그 사이에 정보통신 기술의 발달로 인해 의사소통 방식과 영어 사용 상황에 큰 변화가 있었고 영어시험에 대한 수험자의 요구도 많이 변화했을 뿐만 아니라 영어교육 및 평가 분야에서도 여러 연구성과와 이론적 변화가 이루어졌다. 이러한 변화된 상황과 추세를 시험 설계와 제작에 반영하여 TEPS시험의 진정성과 타당도를 높여야 할 필요성이 꾸준히 제기되어왔고, 오랜 연구와 네 차례의 파일럿 검사 시행 끝에 2018년 5월 개정 TEPS가 처음 시행되었다. 개정 TEPS는 검사 길이, 하위점수의 구성, 문항 유형 등 여러 측면에서 기존 TEPS와 차이가 있다. 특히 개정 후에 문법과 어휘 영역의 시행 순서가 바뀌었고, 어휘와 문법 영역의 시험 시간을 통합하여 시행하게 되었으며, 청해와 독해 영역에 1지문 2문항 문제들이 도입된 점 등 상당한 변화가 있다는 점을 고려했을 때

† Corresponding author: ejlim.mail@gmail.com

https://doi.org/10.30961/lr.2019.55.S.37

개정 TEPS의 타당도를 검증하는 과정이 반드시 필요하다.

타당도란 검사가 측정하고자 하는 내재적 속성을 실제로 측정하고 있는 정도를 가리킨다. 즉, TEPS의 타당도 검증은 TEPS 점수가 '한국인의 실질적 영어활용능력'이라는 비가시적 속성을 나타낸다는 것을 뒷받침하는 증거를 수집하는 것이라 할 수 있다(TEPS Center, 2019). 전통적으로, 시험의 타당도를 뒷받침하는 증거는 크게 세 가지로 구분된다(Crocker & Algina, 1986). 먼저 내용에 대한 타당도 증거는 검사에 포함된 문항의 내용과 형식이 측정하고자 하는 내재적 속성을 포괄하는 정도를 가리킨다. 이는 사전에 제작된 시험 평가틀(test specification)이 측정하고자 하는 내용을 잘 반영하고 있는지, 그리고 실제 개발된 검사의 내용이 평가틀과 얼마나 일치하는지 점검하여 그 증거를 확보할 수 있으며, 내용 전문가의 경험과 지식을 바탕으로 판단하게 된다. 다음으로 외적 증거와 관련된 타당도 증거는 검사 점수가 측정하고자 하는 내재적 속성과 관련된 외부 변수와의 관계를 통해 확보된다. 이미 타당도가 인정된 다른 검사 점수, 혹은 시험을 치른 사람들이 추후에 보이는 수행 수준 등이 검사 점수의 타당도를 뒷받침할 준거가 된다. 예를 들어 동일한 수험자 집단에 TEPS와 해외에서 개발된 영어능력검사를 치르게 하여 그 점수 간 상관관계를 확인함으로써 TEPS 점수가 영어능력을 적절하게 평가하고 있음을 확인할 수 있다. 마지막으로 검사가 측정하고자 하는 내재적 속성의 구조가 실제 문항에 대한 반응과 얼마나 일치하는지 분석하여 타당도의 증거로 삼을 수 있다. 즉 수험자들의 응답 자료를 통계적으로 분석하여 그 구성이 사전에 정의한 내재적 속성의 요인 구조와 같은지 확인함으로써 검사의 구인타당도 증거를 확보하게 된다.

TEPS의 총점은 청해, 어휘, 문법, 독해의 하위 영역 점수로 구성되어 있다. 영어평가 분야의 전문가들은 외국어 능력이 여러 개의 상호 연관된 요소들(multicomponentiality)로 이루어져 있다는 점에 동의하고 있고, 수험자들은 총점뿐만 아니라 영역별로 보다 세부적인 평가 정보를 제공받기를 원한다. 요인분석은 구인타당도 증거 확보를 위해 사용되는 대표적인 연구 방법으로, 이를 통해 검사의 내적 구조를 파악하는 한편 수험자에게 영역별 점수와 총점을 제공하는 것이 적절한지 뒷받침할 수 있다(Sawaki, Stricker, & Oranje, 2009, p. 7).

본 연구에서는 네 차례에 걸쳐 시행된 파일럿 검사와 개정 전후 TEPS의 요인구조를 비교 분석하고, 개정 TEPS의 구인타당도 증거를 제시하고자 한다. 이를 통해 TEPS가 측정하고자 하는 '한국인의 실질적 영어활용능력'의 내재적 구조를 재확인하고 개정 후 그 요인구조에 변화가 있었는지 점검하고자 한다. 이를 위하여 실제 수험자들이 치른 검사 자료를 활용하여 상관분석 및 확인적 요인분석(confirmatory factor analysis)을 실시하였다.

2. 이론적 배경과 관련 연구

언어능력의 내재적 속성 구조에 대하여 이전에는 Oller(1979)가 주장한 바에 따라 언어능력의 내재적 구조가 단일한 일반요인(g-factor)이라고 설명하는 '단일 언어능력 가설(unitary competence hypothesis)'이 주목받았다. 그러나 Bachman과 Palmer(1982)는 이에 반박하며 이차요인모형(second order model)을 가정한 확인적 요인분석을 통해 일반요인 외에

명확히 구분되는 다른 내재적 특성들이 존재한다는 분리가설(divisibility hypothesis)의 근거를 제시하였다. Sang, Schmitz, Vollmer, Baumert, 그리고 Roeder(1986)의 연구에서는 3요인 상관모형을 가정하여 확인적 요인분석을 실시하였으며 이를 통해 일반요인의 존재에 대한 가정에 의문을 제기하였다.

이렇듯 언어평가 분야에서 확인적 요인분석을 포함한 구조방정식 모형(structural equation modeling)은 그 내재적 구조를 파악하고 검사 점수에 대한 적절한 해석과 사용을 뒷받침하기 위한 타당화 연구의 일환으로 널리 수행되어왔다(Kunnan, 1998). 이러한 연구를 통해 검사의 구인타당도 증거를 확보하는 것은 대규모 시험을 개발하고 시행하는데 반드시 필요한 절차라고 할 수 있다. 예를 들어 TOEFL의 경우 2005년에 인터넷 기반의 TOEFL iBT로 변화하면서 그 프로토타입(Stricker, Rock, & Lee, 2005)과 필드 테스트(Sawaki et al., 2009)의 요인구조를 제시하였다. 마찬가지로 TOEIC도 개정 전(Wilson, 2000)과 개정 후(In'nami & Koizumi, 2012; Yoo & Manna, 2017)의 요인구조를 분석하였다.

검사의 구체적인 요인구조는 연구 및 시험에 따라 다르게 나타난다. 예를 들어 TOEFL iBT의 경우 읽기, 듣기, 말하기, 쓰기로 구성된 4개의 1차 요인들(first-order factors) 위에 '외국어로서의 영어 능력'이라는 2차 요인(second-order factor)이 존재한다고 가정하는 이차요인모형을 적용한다(Sawaki et al., 2009, p. 6). 이러한 시험의 요인 구조는 읽기, 듣기, 말하기, 쓰기라는 네 영역의 점수와 총점을 수험자에게 함께 제공하는 근거가 된다.

TOEIC의 경우 In'nami와 Koizumi(2012)의 연구에서 4 개의 모형을 비교하였다. 비교된 모형들은 듣기와 읽기로 구성된 1차 요인들과 영어 이해력(receptive skill)이라고 하는 2차 요인을 가정하는 이차요인모형, 듣기와 읽기 요인 간 상관을 가정하는 2요인 상관모형, 듣기와 읽기 요인 간 상관을 가정하지 않는 2요인 모형, 그리고 듣기와 읽기를 구분하지 않고 영어 이해력이라는 일반요인만을 가정하는 1요인 모형이다. 확인적 요인분석을 실시한 결과, TOEIC의 내재적 구조를 가장 잘 설명하는 것은 듣기와 읽기 간 상관을 가정하는 2요인 모형으로 나타났고, 이는 TOEIC이 두 영역의 점수를 합산한 총점을 발표하는 것에 대한 증거가 된다고 하였다.

TEPS는 지금까지 실질적 영어활용능력이라는 하나의 일반요인 아래에 청해, 문법, 어휘, 독해의 네 변수가 존재하는 1요인 모형을 적용해 왔다(TEPS Center, 2016). 그러나 In'nami와 Koizumi(2012)는 단순한 1요인 구조로 외국어 실력을 설명할 수 없으며, 이차요인모형과 같이 더 세분화된 모형을 이용하여 구인을 설명할 필요가 있다고 언급하였다. TEPS의 요인구조를 이차요인모형으로 설정할 경우, 각 하위영역의 언어하위기능을 반영하는 세부능력별 점수(예를 들어 exchange, main idea, detail, inference)로 청해, 문법, 어휘, 독해로 구성된 일차 요인을 반영하고, 이러한 4개의 일차 요인이 단일한 상위요인, 즉 '한국인의 실질적 영어활용능력'에 수렴하게 된다.

3. 연구 방법

3.1. 자료

본 연구에서는 총 12개의 검사형에 대한 응답자료를 사용하였다. 회차마다 수험자 집단의 특성이 달라질 수 있으므로 여러 회차를 비교분석하여 결과의 안정성을 보고자 하였다. 구체적으로 분석에 포함된 검사형은 다음과 같다.

- 개정 이전에 시행된 TEPS 4개 회차(A, B, C, D형)
- 개정 과정에서 시행된 파일럿 검사 4개 회차(1-4차 파일럿)
- 개정 이후에 시행된 TEPS 4개 회차(E, F, G, H형)

이 중 개정 이전에 시행된 TEPS 4개 회차는 2018년 상반기에 시행되었고, 개정 이후에 시행된 TEPS 4개 회차는 2018년 하반기에 시행되었다. 파일럿 검사는 2017년 상반기부터 2018년 상반기까지 4개의 검사형을 활용하여 대학생, 영어학원 수강생, 외국어고등학교 학생 등 TEPS의 수험자 모집단의 특성을 반영하여 다양한 집단에서 시행되었고, 선행된 파일럿 검사의 분석 결과에 따라 이후의 파일럿 검사형에 개선점을 반영하였다. 파일럿 검사의 시행 일자와 수험자 집단은 다음과 같다.

- 1차 파일럿(2017년 2월 시행): 대학생 약 2,800명 응시
- 2차 파일럿(2017년 5-6월 시행): 대학생, 영어학원 수강생, 외국어고등학교 학생 등 약 600명 응시
- 3차 파일럿(2018년 2월 시행): 대학생 약 2,600명 응시
- 4차 파일럿(2018년 2월 시행): TEPS 센터 웹사이트를 통해 모집한 수험자 약 900명 응시

본 연구에서 사용된 12개 검사형의 원점수 총점에 대한 기술통계는 다음의 <표 1>과 같다.

표 1. 검사 자료에 대한 기술통계

구분	문항 수	검사형	수험자 수	평균	표준편차	왜도	첨도	최솟값	최댓값
개정 이전	200	A	약 6,900명	115.96	32.12	0.05	-0.67	0	198
		B	약 5,600명	128.93	31.38	-0.27	-0.63	18	197
		C	약 6,000명	116.71	28.50	0.09	-0.45	12	196
		D	약 6,700명	113.13	32.22	0.17	-0.69	28	200
파일럿 검사	135	1차	약 2,800명	95.81	19.47	-0.63	0.39	23	134
		2차	약 600명	74.26	22.70	0.41	-0.44	21	134
		3차	약 2,600명	85.14	21.27	0.00	-0.46	15	135
		4차	약 900명	79.87	22.04	-0.07	-0.63	9	133

개정 이후	135	E	약 6,200명	76.63	22.82	0.01	-0.75	1	134
		F	약 5,300명	72.59	21.86	0.21	-0.59	14	134
		G	약 4,500명	76.94	21.37	0.10	-0.62	1	135
		H	약 5,800명	75.74	20.86	0.14	-0.55	5	134

3.2. 분석 방법

먼저 TEPS의 하위영역, 즉 청해, 어휘, 문법, 독해의 네 개 영역이 의도한 검사의 내적 구조에 부합되는지 살펴보기 위해 영역별 점수간 상관계수를 산출하였다. 개정 전과 후 TEPS 하위영역의 문항 수와 검사 시간은 다음의 <표 2>와 같다.

표 2. 개정 전 후 TEPS 하위영역의 문항 수와 검사 시간

하위영역	개정 전		파일럿/개정 후	
	문항 수	검사 시간	문항 수	검사 시간
청해	60	55분	40	40분
어휘	50	15분	30	25분
문법	50	25분	30	
독해	40	45분	35	40분
합계	200	140분	135	105분

또한 각 영역이 포함하는 세부능력이 영역별 총점과 어떤 관계를 보이는지 상관계수를 통해 분석하였다. TEPS의 각 하위영역이 평가하는 세부능력은 다음과 같다. 먼저 청해는 대화의 흐름을 파악하는 능력(대화), 대화나 담화의 주제를 파악하는 능력(주제), 대화나 담화의 세부내용을 파악하는 능력(세부내용), 대화나 담화의 내용을 바탕으로 추론하는 능력(추론)을 포함한다. 어휘와 문법 영역은 각각 대화에서 쓰이는 어휘 혹은 문법 문항(구어), 담화에서 쓰이는 어휘 혹은 문법 문항(문어)으로 구성된다. 독해 영역은 글의 주제를 파악하는 능력(주제), 글의 세부내용을 파악하는 능력(세부내용), 글의 내용을 바탕으로 추론하는 능력(추론), 글의 흐름을 파악하는 능력(응집)을 포함한다. <표 3>은 개정 전과 후, 그리고 개정 과정에서 개발한 파일럿 검사에서 TEPS의 각 하위영역에 속한 세부능력별 문항 수와 각 세부능력별 문항 수가 해당 영역에서 차지하는 비율을 나타낸다. 개정 과정에서 청해-주제, 청해-세부내용, 독해-주제, 독해-세부내용에서 문항 수의 변화가 있었고, 3차 파일럿 검사부터 세부능력별 문항 수가 확정되었다. 확정된 개정 TEPS의 세부능력별 문항 비율을 개정 전과 비교하면, 청해에서 대화나 추론을 묻는 문항의 비율이 각각 50.0%, 10.0%로 유지되었다. 주제를 묻는 문항 비율은 다소 줄어들었고, 반면 세부내용을 묻는 문항 비율이 증가하였다. 어휘와 문법에서는 개정 전 구어와 문어 세부능력 비율이 서로 같았으나, 개정 후에는 문어를 묻는 문항 비율이 증가하였다. 독해 영역에서는 개정 전에 주제를 묻는 문항 비율이 50.0%로 전체 문항의 절반을 차지하였으나 개정 후에는 42.9%로 다소 축소되었고, 세부내용과 추론을 묻는 문항 비율이 소폭 상승하였다.

표 3. 개정 전 후 TEPS 하위영역의 세부능력별 문항 수와 비율

하위영역	세부능력	문항 수(비율)			
		A-D	1차 파일럿	2차 파일럿	3-4차 파일럿, E-H
청해	대화	30(50.0%)	20(50.0%)	20(50.0%)	20(50.0%)
	주제	14(23.3%)	7(17.5%)	7(17.5%)	6(15.0%)
	세부내용	10(16.7%)	9(22.5%)	9(22.5%)	10(25.0%)
	추론	6(10.0%)	4(10.0%)	4(10.0%)	4(10.0%)
어휘	구어	25(50.0%)	10(33.3%)	10(33.3%)	10(33.3%)
	문어	25(50.0%)	20(66.7%)	20(66.7%)	20(66.7%)
문법	구어	25(50.0%)	12(40.0%)	12(40.0%)	12(40.0%)
	문어	25(50.0%)	18(60.0%)	18(60.0%)	18(60.0%)
독해	주제	20(50.0%)	16(45.7%)	15(42.9%)	15(42.9%)
	세부내용	10(25.0%)	10(28.6%)	11(31.4%)	11(31.4%)
	추론	5(12.5%)	5(14.3%)	5(14.3%)	5(14.3%)
	응집	5(12.5%)	4(11.4%)	4(11.4%)	4(11.4%)

본 연구에서는 개정 TEPS가 전체 시험의 의도와 일관되게 기능하는지, 즉 시험이 표방하는 내적 구조가 실제 시행자료에 잘 반영되고 있는지를 보여주기 위하여 영역별 점수를 바탕으로 한 상관계수를 산출하고 개정 전과 어떻게 달라졌는지 확인하였다. 또한 각 세부능력별 점수를 측정단위로 하여 확인적 요인분석을 실시하고 개정 전후 표준화 요인계수(standardized factor loadings)를 비교였다. 요인계수는 요인이 측정변수에 미치는 직접적인 영향을 의미하며, 요인계수의 절댓값이 클수록 영향력이 강하다고 할 수 있다. 일반적으로 요인과 측정변수가 표준화된 상태에서 추정된 표준화 요인계수를 보고한다(Kim, 2016, p. 313). 본 연구에서는 이러한 결과를 바탕으로 개정에 따른 TEPS 측정구조의 변화가 있는지 확인하였다.

또한 개정 TEPS 평가틀의 모형 적합도(model-data fit)를 확인하여 TEPS의 구인타당도를 검증하고자 하였다. 모형 적합도는 연구자가 설정한 모형과 실제 자료가 얼마나 잘 부합되는지 평가하는 지표로, 모형에 대한 수용 또는 수정을 결정하는 기준이 된다. 모형 적합도 지수는 다양하게 개발되어 있으며 각각의 지수가 서로 다른 방식으로 작동하므로 연구 상황에 따라 적합한 지수를 선택할 필요가 있다(Hong, 2000). 본 연구에서는 모형 적합성 분석을 위해 CFI(Comparative Fit Index; Bentler, 1990), TLI(Tucker-Lewis Index; Tucker & Lewis, 1973; Bentler & Bonnett, 1980), RMSEA(Root Mean Square Error of Approximation; Steiger & Lind, 1980; Steiger, 1990; Browne & Cudeck, 1993), SRMR(Standardized Root Mean Square Residuals; Bentler, 1995) 지수를 산출하였

다. 각 지표에 대해 CFI와 TLI는 0.95 이상, RMSEA 0.08 이하, SRMR 0.05 이하일 때 모형이 적합하다고 판단하였다(Hong, 2000; Hu & Bentler, 1998, 1999).

4. 결과

4.1. 상관분석

<표 4>는 검사형별 하위영역간 상관계수를 나타낸다. 상관의 방향은 모두 정적이고 상관의 크기는 0.61-0.84의 범위로 적절한 수준의 상관을 보여주고 있다. 청해와 독해의 상관이 0.74-0.84 수준으로 높게 나타났고, 문법과 독해의 상관 또한 0.67-0.80 수준으로 높게 나타났다. 반면 청해와 어휘 간 상관은 0.61-0.74 수준으로 다른 하위영역간 상관보다 상대적으로 낮게 나타났다. 이러한 경향성은 모든 검사형에 공통적으로 나타나 TEPS가 측정하고자 하는 구인이 개정과정에 있어서 검사형마다 크게 달라지지 않았음을 보여준다.

표 4. TEPS 검사형별 하위영역간 상관계수

개정 전		청해	어휘	문법	독해	파일럿		청해	어휘	문법	독해	개정 후		청해	어휘	문법	독해
A	청해	1.00				1차 파일럿	청해	1.00				E	청해	1.00			
	어휘	0.68	1.00				어휘	0.71	1.00				어휘	0.68	1.00		
	문법	0.79	0.76	1.00			문법	0.72	0.71	1.00			문법	0.71	0.74	1.00	
	독해	0.80	0.75	0.79	1.00		독해	0.75	0.74	0.72	1.00		독해	0.79	0.72	0.74	1.00
B	청해	1.00				2차 파일럿	청해	1.00				F	청해	1.00			
	어휘	0.74	1.00				어휘	0.70	1.00				어휘	0.64	1.00		
	문법	0.78	0.78	1.00			문법	0.67	0.74	1.00			문법	0.69	0.70	1.00	
	독해	0.81	0.79	0.80	1.00		독해	0.76	0.72	0.76	1.00		독해	0.77	0.70	0.74	1.00
C	청해	1.00				3차 파일럿	청해	1.00				G	청해	1.00			
	어휘	0.72	1.00				어휘	0.61	1.00				어휘	0.64	1.00		
	문법	0.77	0.75	1.00			문법	0.72	0.68	1.00			문법	0.72	0.70	1.00	
	독해	0.78	0.77	0.75	1.00		독해	0.76	0.70	0.74	1.00		독해	0.78	0.68	0.74	1.00
D	청해	1.00				4차 파일럿	청해	1.00				H	청해	1.00			
	어휘	0.70	1.00				어휘	0.63	1.00				어휘	0.64	1.00		
	문법	0.77	0.76	1.00			문법	0.66	0.70	1.00			문법	0.69	0.69	1.00	
	독해	0.84	0.75	0.79	1.00		독해	0.74	0.65	0.67	1.00		독해	0.78	0.71	0.74	1.00

4.2. 확인적 요인분석

검사형별로 나타난 측정구조를 보다 면밀히 검증하기 위해 본 연구에서는 일반적인 영어능력과 각 하위영역에 대해 세부능력별 점수를 측정단위로 하여 이차요인모형을 가정하고 확인적 요인분석을 실시하였다. [그림 1]은 본 연구에서 가정한 이차요인모형을 나타낸다. 연구에 앞서 각 하위영역별 점수를 측정단위로 하고 일반적인 영어능력으로 설명하고자 한 1요인모형과 세부능력별 점수를 측정단위로 한 이중요인모형(bi-factor model)을 함께 비교하였으나, 이차요인모형의 모형적합도가 가장 높았을 뿐만 아니라 이차요인모형의 요인구조가 하위영역별 점수와 총점을 함께 제공하는 TEPS의 구조에 가장 부합한다고 판단되어 이를 가정하였다.

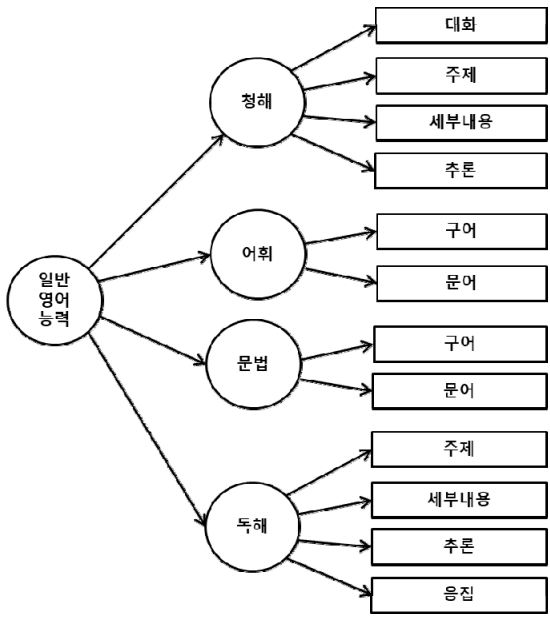

그림 1. TEPS 요인구조

<표 5>는 개정 전에 시행된 네 검사형에 대한 확인적 요인분석 결과를 나타낸다. 모형적합도 지표를 고려했을 때 모든 검사형에서 CFI와 TLI는 0.95 이상, RMSEA 0.08 이하, SRMR 0.05 이하로 나타나 본 연구에서 가정한 이차요인모형이 적절한 것으로 나타났다. 표준화 요인계수의 경우 청해는 세부능력 중 대화에 높은 요인적재치를 갖는 반면 추론에 상대적으로 낮은 적재치를 보였고, 어휘와 문법은 공통적으로 구어보다 문어 세부능력에 더 높은 요인적재치를 보였다. 독해의 경우 세부능력 중 주제에 요인적재치가 높게 산출되었고, 추론과 응집에서 상대적으로 낮게 나타났다. 이러한 경향은 A-D 검사형 전체에 걸쳐 비슷하게 나타났다. 일반적인 영어능력 요인과 각 하위영역 간에도 높은 요인적재치를 보였고, 특히 독해 영역에서 가장 높게 나타났다.

표 5. 개정 이전 TEPS 표준화 요인계수 및 모형 적합도

검사형		표준화 요인계수			
		A	B	C	D
청해	대화	0.872	0.873	0.888	0.900
	주제	0.855	0.847	0.855	0.866
	세부내용	0.810	0.774	0.748	0.833
	추론	0.612	0.759	0.651	0.633
어휘	구어	0.866	0.823	0.813	0.853
	문어	0.888	0.864	0.866	0.826
문법	구어	0.886	0.891	0.815	0.866
	문어	0.888	0.883	0.855	0.900
독해	주제	0.858	0.889	0.829	0.866
	세부내용	0.780	0.801	0.713	0.806
	추론	0.603	0.662	0.535	0.577
	응집	0.664	0.656	0.577	0.639
일반 영어 능력	청해	0.926	0.925	0.919	0.935
	어휘	0.881	0.947	0.940	0.902
	문법	0.955	0.949	0.955	0.930
	독해	0.978	0.974	0.985	0.997
모형 적합도	CFI	0.983	0.988	0.989	0.989
	TLI	0.978	0.984	0.985	0.985
	RMSEA	0.054	0.047	0.041	0.045
	SRMR	0.019	0.016	0.016	0.017

<표 6>은 개정 과정에서 시행된 파일럿 검사형에 대한 확인적 요인분석 결과를 나타낸다. 네 차례의 파일럿 검사에서 역시 모든 검사형에서 CFI와 TLI는 0.95 이상, RMSEA 0.08 이하, SRMR 0.05 이하로 나타나 본 연구에서 가정한 이차요인모형이 적절한 것으로 나타났다. 표준화 요인계수의 경우 청해와 독해는 개정 이전과 유사한 양상을 보였다. 즉, 청해는 대화 세부능력에 높은 요인적재치를 보였고 추론에 상대적으로 낮은 값을 가졌으며, 독해는 주제에 요인적재치가 높게 산출되었고, 추론과 응집에서 상대적으로 낮게 나타났다. 어휘와 문법 또한 전반적으로 개정 이전과 유사하게 구어보다 문어 세부능력에 더 높은 요인적재치를 보였으나, 2차 파일럿 문법에서는 구어 세부능력이 다소 높은 값을 갖는 것으로 나타났다. 개정 이전 검사형과 마찬가지로 일반적인 영어능력 요인과 각 하위영역 간 요인적재치는 매우 높게 나타났고, 다만 1차와 3차 파일럿에서 각각 어휘와 문법 영역의 요인적재치가 독해보다 미세하게 더 높은 값을 가졌다는 차이가 있었다.

표 6. TEPS 파일럿 검사 표준화 요인계수 및 모형 적합도

검사형		표준화 요인계수			
		1차 파일럿	2차 파일럿	3차 파일럿	4차 파일럿
청해	대화	0.858	0.835	0.881	0.847
	주제	0.811	0.744	0.714	0.761
	세부내용	0.753	0.702	0.805	0.779
	추론	0.614	0.450	0.618	0.692
어휘	구어	0.758	0.755	0.747	0.747
	문어	0.846	0.876	0.796	0.825
문법	구어	0.798	0.819	0.694	0.822
	문어	0.868	0.783	0.831	0.865
독해	주제	0.870	0.889	0.849	0.836
	세부내용	0.807	0.757	0.814	0.791
	추론	0.726	0.643	0.658	0.561
	응집	0.624	0.683	0.670	0.603
일반 영어 능력	청해	0.940	0.913	0.895	0.897
	어휘	0.968	0.936	0.898	0.902
	문법	0.955	0.940	0.977	0.893
	독해	0.956	0.962	0.965	0.928
모형 적합도	CFI	0.990	0.993	0.987	0.982
	TLI	0.987	0.990	0.983	0.976
	RMSEA	0.039	0.031	0.041	0.048
	SRMR	0.014	0.019	0.020	0.026

<표 7>은 개정 후에 시행된 네 검사형에 대한 확인적 요인분석 결과를 나타낸다. 각 검사형의 모형 적합도 지표는 CFI와 TLI는 0.95 이상, RMSEA 0.08 이하, SRMR 0.05 이하로, 이차요인모형을 가정하는 것이 적절했다는 것을 보였다. 표준화 요인계수를 살펴봤을 때 청해는 세부능력 중 대화에 높은 요인적재치를 갖고 추론에 상대적으로 낮은 적재치를 보인다는 것이 각 검사형 간, 그리고 개정 이전 검사형에서 산출된 결과와 동일하게 나타났다. 어휘와 문법은 구어보다 문어에서 요인적재치가 더 높게 나타난다는 점 또한 개정 이전 검사형과 동일하였다. 독해의 경우에도 개정 이전과 마찬가지로 세부능력 중 주제에서 요인적재치가 높게 산출되었고 추론과 응집에서 상대적으로 낮은 값을 보였다. 일반적인 영어능력 요인과 각 하위영역 간 요인적재치 또한 높게 나타났고, 개정 이전과 마찬가지로 독해 영역에서 가장 높게 나타났다.

표 7. 개정 이후 TEPS 표준화 요인계수 및 모형 적합도

검사형		표준화 요인계수			
		E	F	G	H
청해	대화	0.878	0.872	0.834	0.872
	주제	0.745	0.784	0.787	0.788
	세부내용	0.815	0.817	0.766	0.784
	추론	0.595	0.399	0.629	0.378
어휘	구어	0.778	0.732	0.723	0.781
	문어	0.861	0.840	0.769	0.808
문법	구어	0.822	0.720	0.819	0.719
	문어	0.859	0.895	0.841	0.824
독해	주제	0.864	0.817	0.818	0.855
	세부내용	0.788	0.785	0.781	0.775
	추론	0.586	0.546	0.660	0.541
	응집	0.624	0.654	0.664	0.633
일반 영어 능력	청해	0.918	0.898	0.926	0.911
	어휘	0.905	0.900	0.926	0.897
	문법	0.930	0.920	0.940	0.944
	독해	0.971	0.986	0.973	0.992
모형 적합도	CFI	0.985	0.992	0.988	0.988
	TLI	0.980	0.990	0.981	0.985
	RMSEA	0.047	0.031	0.044	0.039
	SRMR	0.020	0.015	0.019	0.016

<표 5>~<표 7>에 나타난 결과를 종합하면 개정 전, 개정 과정, 개정 후 시행된 검사형에서 모두 이차요인모형이 내재적 요인 구조를 적절하게 나타내는 것으로 나타났다. 각 세부능력에 대한 표준화 요인계수를 비교했을 때 청해는 대화에 높은 요인적재치를 갖고 추론에 낮은 값을 갖는다는 공통점을 보였다. 어휘와 문법은 구어보다 문어 세부능력에 요인적재치가 더 높게 나타나는 공통점이 있었으나 예외적으로 2차 파일럿 문법 영역에서 구어 세부능력이 문어보다 더 높은 값을 가졌다. 독해는 모든 검사형에서 주제 세부능력의 요인적재치가 높았고 추론과 응집에 해당하는 값이 낮게 나타났다. 이차요인인 일반 영어능력 요인과 일차요인인 각 영역 요인 간 요인적재치는 1차와 3차 파일럿을 제외한 모든 검사형에서 독해의 요인적재치가 가장 높게 나타났다. 1차와 3차 파일럿에서도 독해의 요인적재치가 두번째로 높은 값을 보여 다른 요인분석 결과와 크게 다르지 않았다.

5. 결론 및 제언

본 연구에서는 진정성과 타당도를 높이고자 변화를 거친 개정 TEPS의 내재적 속성 구조가 개정 이전과 동일하다고 할 수 있는지 살펴보고자 하였다. 이를 위해 개정 이전에 시행된 검사형과 파일럿 검사형, 개정 이후 시행된 검사형을 통해 수집된 자료를 활용하여 하위영역간 상관계수를 비교하였고, 검사형 별로 확인적 요인분석을 실시하였다.

검사가 측정하고자 하는 내재적 속성의 구조가 이전과 동일하게 유지되고 있는지 확인하는 것은 그 검사 점수를 이전과 같은 의미로 활용하기 위해 반드시 선행되어야 하는 절차이다. 본 연구를 통해 개정에 따른 변화에도 불구하고 TEPS의 내재적 속성 구조가 변화하지 않았음을 보여 구인타당도 증거를 확보했다고 할 수 있다. 또한 TEPS가 총점과 네 개의 하위영역 점수, 즉 '한국인의 실질적 영어활용능력'이라는 일반적인 영어 능력에 대한 정보와 각 하위영역에 해당하는 정보를 함께 제공하는 것이 실제 응답 자료에 나타난 요인 구조와 일치한다는 것을 실증적으로 검증했다는 의미가 있다. 다만 이러한 요인분석 결과를 해석함에 있어서 각 검사형을 치른 집단이 동일하지 않다는 점을 감안할 필요가 있다. A-D형과 E-H형의 시행 시점이 가깝고 수험자 집단이 크다는 점에서 집단의 특성이 크게 다르지 않을 것으로 생각되나 무선표집을 통해 동질성이 확보된 집단이라고 보기는 어렵고, 1-4차 파일럿 시험은 정기시험과는 다른 상황에서 시행되어 집단의 동질성을 담보할 수 없다. 따라서 추후 동일한 집단 혹은 무선표집을 통한 동등집단이 개정 전·후 검사형을 모두 치르게 하거나, 공통문항을 통해 척도를 연계하여 TEPS의 내재적 속성 구조를 다시 한 번 비교할 필요가 있다.

TEPS 개정은 영어교육 및 평가 분야에 축적된 여러 연구성과와 이론적 변화, 그리고 수험자의 요구를 반영하여 오랜 연구 끝에 이루어진 성과이다. 개정을 통해 TEPS의 진정성과 타당도를 높일 수 있었으나 변화에 적응하는 시점에서 그 점수 활용 및 해석에 있어서의 혼란은 피하기 어렵다. 본 연구가 이러한 혼란을 줄이는데 일정부분 기여할 수 있을 것으로 생각된다. 또한 본 연구를 바탕으로 개정 이후 시행되는 TEPS 검사형의 내재적 요인 구조를 지속적으로 추적 조사하여 안정성을 확보하는데 기여할 수 있을 것으로 기대된다.

References

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (1995). *EQS Structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structure. *Psychological Bulletin, 88*, 588-606.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equations models* (pp. 136-162). Newbury Park, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt.

Hong, S. (2000). The criteria for selecting appropriate fit indices in structural equation modeling and their rationales. *Korean Journal of Clinical Psychology, 19*, 161-177.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing, 29*, 131-152

Kim, S. Y. (2016). *Fundamentals and extensions of structural equation modeling*. Seoul: Hakjisa.

Kunnan, A. (1998). An introduction to structural equation modelling for language assessment research. *Language Testing, 15*, 295-332.

Oller. J. W. Jr. (1979). *Language tests at school*. London: Longman.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J., & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing, 3*, 54-79.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*, 5-30.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series MS-32). Princeton, NJ: Educational Testing Service.

TEPS Center (2016). *TEPS Technical Report: 2016 administration* (internal document). Seoul: TEPS Center.

TEPS Center (2019). *TEPS Technical Report: 2018 administration* (internal document). Seoul: TEPS Center.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Wilson, K. M. (2000). *An exporatory dimensionality assessment of the TOEIC test* (TOEIC Research Report No. RR-00-14). Princeton, NJ: Educational Testing Service.

Yoo, H., & Manna, V. F. (2017). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing, 34*, 101-126.