TEPS의 개정 배경과 기초연구

전희성[1†] · 권혁승[1] · 송미정[1] · 박찬호[2] · 이영미[1] · 이용원[1]
[1]서울대학교, [2]계명대학교

Background and Foundational Research for the Revision of the TEPS

Heesung Jun[1†] , Heokseung Kwon[1], Mi-Jeong Song[1], Chanho Park[2], Youngmi Lee[1], and Yong-Won Lee[1]
[1]Seoul National University, [2]Keimyung University

ABSTRACT

Since the first administration of the TEPS in 1999, there have been many important changes in the field of language teaching and assessment. The emergence of new information and communication technology has drastically changed the ways we communicate. This has made it necessary to re-conceptualize the construct of English language proficiency. To respond to such needs and to TEPS test takers' feedback that had accumulated for more than 15 years, a series of research projects was undertaken to revise the TEPS during the years of 2016-2018. Major changes included a reduction in the number of items and testing time and the addition of new item types (e.g., testlets) and more authentic passage formats (e.g., e-mails, instant messages, online newspaper articles). This paper introduces the background to how the decision to revise the TEPS came to be made and describes the rationale behind the revision of the TEPS test blueprint and specification.

Keywords: revised TEPS, test revision; needs analysis, theoretical justification

1. 서 론

TEPS(Test of English Proficiency developed by Seoul National University의 약자)는 1999년 첫 시행 이후 정부, 대학 및 기업체를 포함한 다양한 기관에서 선발, 임용, 승진 자료로 활용됨으로써 한국의 대표적인 일반 영어능력 시험으로 자리를 잡아왔다. TEPS가 첫 개발되어 시행을 시작한 이후 20여년 동안 정보통신 기술의 발달로 인해 의사소통 방식과 영어 사용 상황에 큰 변화가 있었고 영어시험에 대한 수험자의 요구도 많이 변화했을 뿐만 아니라

* 본 논문은 개정 TEPS 연구보고서(Kwon et al., 2018)의 한 장을 수정·보완한 논문임을 밝힘.

† Corresponding author: hsjun@snu.ac.kr

Copyright © 2019 Language Education Institute, Seoul National University.
This is an Open Access article under CC BY-NC License (http://creative-commons.org/licenses/by-nc/4.0).

https://doi.org/10.30961/lr.2019.55.S.17

영어교육 및 평가 분야에서도 여러 연구 성과와 이론적 발전이 있었다. 이러한 변화된 상황과 추세를 시험 설계와 제작에 반영하여 TEPS시험의 진정성과 타당도를 높여야 할 필요성이 꾸준히 제기되어왔다.

기존 시험을 개정하거나 새로운 시험을 개발할 때에는 상황에 따라 다소 차이가 있을 수 있지만 대개 다음과 같은 4개의 과정을 거치게 된다(Chapelle, Enright, & Jamieson, 2008). 첫째, 선행연구 종합, 이론에 대한 고찰, 전문가 자문, 그리고 수험자 요구 조사 등을 바탕으로 평가구인(assessment construct)을 정의하고 평가틀(test specification)을 제작하고 확립하는 과정, 둘째, 새로운 문항・과제의 원형(prototype)들을 개발하고 다듬는 과정, 셋째, 새로운 형태의 문항과 과제들을 활용해 제작한 시험세트를 파일럿 테스트(혹은 필드 테스트)하고 그 분석결과를 바탕으로 수정하고 다듬는 과정, 넷째, 시험의 내용을 결정하고 시험 준비용 자료와 함께 시험을 실제로 시행하는 과정이다.

본 소고의 목적은 TEPS 개정 작업을 추진하면서 주로 첫 번째 단계에서 이루어진 연구와 논의 그리고 주요 결정과정을 기술하는 데 있다. 특히, 기존의 평가틀을 개정하고 확정하면서 내려진 주요한 결정들이 어떤 배경 하에 어떤 논의과정을 거쳤고 또 어떤 근거를 토대로 해서 이루어졌는지를 밝히는 데 중점을 두어 기술하고자 한다.

2. TEPS의 개정 배경

TEPS의 개정 배경은 크게 이론적 배경과 수험자 요구라는 두 측면에서 논의해 볼 수 있다. 우선 이론적 배경 측면에서 보면 주로 그동안 언어평가 분야에서 일어난 여러 이론적 추세의 변화와 이러한 변화와 연계된 연구성과에 대해서 간략히 논의해 볼 수 있을 것이고, 수험자 요구 분석 측면에서는 지난 수년간 TEPS 시험 개선을 위해 수험자를 대상으로 이루어진 여러 차례의 설문 조사를 통해 확인된 요구 사항에 대해 논의를 해 볼 수 있을 것이다.

2.1. 이론적 배경

지난 20여년간 언어교육 및 평가 분야에서 많은 연구가 이루어졌고 새로운 이론적인 추세가 형성되었다. 그 중 가장 중요한 변화를 든다면 시험 타당도 검증의 틀(validity framework)에 대한 진전된 논의와 관점의 변화가 있었다는 점이다(Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008). 타당도 검증이 우선 통합적인 틀 안에서 시험의 설계와 제작, 시행, 채점, 그리고 결과의 활용에 이르는 평가의 전 과정에 걸쳐 단계마다 일관된 원칙 하에 이루어져야 하며, 단계별 검증의 구체적 내용은 타당도 논증(argument)의 형태로 기술되고 관련된 증거를 통해 그 타당성이 평가 받아야 한다는 점을 강조하고 있다. 실제로 교육심리 및 측정 분야에서는 이미 오래 전부터 전통적인 내용타당도(content validity), 준거관련타당도(criterion-related validity), 구인타당도(construct validity)를 별도로 수집하고 평가하기보다

는 이를 단일한 타당도 검증의 틀 안으로 통합하여 종합적으로 검증하려는 시도가 있었다(Kane, 1992, 2013; Messick, 1989).

이러한 통합적 타당도 검증의 틀을 만들고 적용하려는 이론적 시도가 언어평가 분야에서 활발하게 이루어지면서 여러 가지 새로운 흐름을 만들어 냈다. 타당도 검증을 시험 개발과 시행이 끝난 후 이루어지는 사후적 평가 조치로만 인식하기보다는 타당도 검증의 틀을 시험 설계, 제작, 시행, 평가의 전 과정을 안내하는 좀 더 사전적이고 선제적인(proactive) 지침으로 활용해야 한다는 점이 새로이 강조되고 있다. 최근 언어교육 및 평가 분야에서 많은 관심을 받고 있는 시험의 결과적 타당도(consequential validity)나 긍정적 환류효과(washback)와 같은 효과를 이끌어 내기 위해서는 시험설계나 문항제작 과정에서 시험의 진정성을 높이고 다양한 방안들을 좀 더 선제적으로 반영하는 것이 중요할 것이다.

이와 관련하여 TEPS가 그동안 시험 문항의 국부독립성(local dependence)을 유지하기 위해 모든 영역에서 한 지문에서 한 문제만 출제하는 원칙을 지켜 왔는데 이를 재고할 필요가 대두되었다. 문항국부독립성 원칙이란 한 문항의 정답 확률이 다른 문항의 정답 확률에 영향을 미쳐서는 안 된다는 출제의 원칙으로, 측정학적 지표인 신뢰도와 타당도가 높은 시험의 필수 조건이자 문항반응이론을 이용한 채점방식의 사용에 있어서 중요한 전제로 인식되어 왔다(Hambleton & Swaminathan, 1985; Lord, 1985; Lord & Novick, 1968). 하지만 좋은 시험의 또 다른 조건 중 하나는 진정성(authenticity)이다(Bachman & Palmer, 1996). 진정성은 시험 과제(test task)가 목표 언어 사용 과제(target language use task)와 얼마나 유사한지를 뜻한다. Choi(2011)와 Yi(2013)는 실생활에서 사람들이 긴 글을 읽어야 하는 경우가 많기 때문에 긴 지문에 대해 여러 개의 질문을 하는 것이 더 진정성 있는(authentic) 시험이라고 하였고, TEPS에 1지문 다문항 유형을 도입하는 것을 제안하였다. 1지문 다문항 유형은 대다수의 세계적인 영어 시험들(TOEFL, IELTS, GEPT, Cambridge 영어 시험, MELAB, CAEL, TOEIC 등)에서 사용하고 있다. 이러한 점들을 고려하여 신뢰도 유지에 중요한 1지문 1문항 유형을 다수 유지하면서 1지문 2문항을 일부 추가하여 시험의 진정성을 높이는 것을 TEPS의 개정방향 중 하나로 삼았다.

아울러 TEPS가 첫 시행되었던 1999년 이후 이메일, 스마트폰, 태블릿 컴퓨터, 인터넷 신문, 블로그 등 IT 기술의 발달로 새로운 형식의 의사소통 상황이 활발해졌고, 시험에서 변화된 의사소통 상황과 언어사용을 반영할 필요가 생겼다. 따라서 독해 지문에도 이와 같이 실제를 반영하는 디자인을 도입하여 진정성을 높이면서 실생활에서 사람들이 읽는 글이나 문서와 유사하게 보이도록 지문을 구성하였다.

2.2. 수험자 요구 분석

지난 수년간 여러 번 TEPS 수험자와 기업의 마케팅 담당 직원을 대상으로 한 설문 조사 연구를 통해 일관되게 확인한 내용은 수험자들이 기존 TEPS의 시험 길이와 문항 수에 대해 부담을 호소하고 있다는 것이었다. Lee et al.(2008)은 『TEPS의 브랜드 및 마케팅 전략 수립』에서 2008년 총 513명을 대상으로 TEPS 인지도 및 이용 현황에 대한 설문 조사를 실시하였다.

그 결과 TEPS가 학구적이고 어려운 이미지를 지니고 있으며(p. 27), 시험문항이 너무 많고 목표점수도달이 어렵다고 느낀 응시자가 많았다(p. 33). 또한 TEPS Council(2012)은 『TEPS 유형 및 난이도에 대한 외부의견』에서 응답자 9,218명을 대상으로 실시한 기획조사 설문에서 가장 많은 수험자들이(전체 응답자의 25.5%) 개선되어야 할 점으로 꼽은 사항은 '난이도 유지 및 조절'이었다. 특히 시험이 어렵고 독해영역의 배당된 시험시간이 짧다는 의견이 있었다. 마지막으로 2017년부터 약 1년간 TEPS의 마케팅을 담당했던 Multicampus(2017b)의 『구 TEPS에 대한 영업대표의 의견』에서는 (주)멀티캠퍼스 영업대표 20명을 대상으로 기존 TEPS에 대한 의견을 조사하였는데 시험 시간이 길다는 의견과 시험 난이도가 높다는 의견이 다수를 차지하였다. 시험 시간이 길기 때문에 수험자가 지루함을 느끼고 집중력이 떨어지며 체력적으로 힘들어 할 수 있다는 의견과 함께 이를 개선하기 위해 난이도와 문항 수를 조정하고 시험 시간은 짧게 조정할 필요가 있다는 의견도 있었다.

사실 시험의 문항 수와 신뢰도는 비례하기 때문에 문항 수가 많을수록 신뢰도 수치는 높게 나온다. 이러한 고려 때문에 기존 TEPS는 총 200문항으로 구성되어 있었고, 신뢰도 수치를 매우 높게 유지할 수 있었다. 하지만 시험의 길이가 길면 시험 피로도(test fatigue; Davis & Ferdous, 2005) 또는 주관적 인지적 피로도(subjective cognitive fatigue; Ackerman & Kanfer, 2009)가 높아진다. 이 경우 수험자가 피로나 부담으로 인해 자신의 능력을 제대로 발휘하지 못할 가능성이 있다. 그동안 TEPS 수험자로부터 시험 시간이 길어 너무 피곤하고 문항 수가 많아 시험 보는데 어려움을 겪는다는 의견이 있었다. 그래서 TEPS 시험의 신뢰도 및 다른 여러 측정학적 지표들을 계속 높은 수준에서 유지하면서도 문항 수와 시험 시간을 줄여 수험자의 심리적 부담을 감소시키는 방안을 찾는 것이 중요한 과제가 되었다.

3. TEPS 개정 연구의 역사

TEPS 시험 개정 노력의 시작은 실제 2005년경으로 거슬러 올라간다. 서울대학교 언어교육원은 기존의 TEPS 시험을 개선하기 위한 노력의 일환으로 Ryu et al.(2006) 『TEPS 개선방안 연구』, Song et al.(2007) 『새로운 TEPS 개발을 위한 연구』, Lee et al.(2008) 『TEPS 영역 재구성을 위한 심리측정학적 기초 연구』, Choi(2008) 『뉴텝스 개발 방향 검토』, Choi, Son, and Ahn(2008) 『New TEPS 문제유형에 대한 연구』, Lee et al.(2009) 『i-TEPS 모의시험의 제작/시행/분석과 본시험 실시 준비에 관한 연구』 등을 포함한 일련의 연구 프로젝트를 수행하였다. 이러한 연구 결과를 바탕으로 청해 40문항, 문법 30문항, 어휘 30문항, 독해 35문항, 총 135개의 4지선다형 객관식 문항에 말하기 및 쓰기 등 2개 영역을 추가한 컴퓨터 기반 영어 시험 형태의 i-TEPS시험을 개발하여 시행하였다. 하지만 처음부터 기존의 TEPS 지필 시험을 대체하기 보다는 병행하는 전략을 가지고 시행되었고 동시에 컴퓨터 기반 시험이었기 때문에 i-TEPS의 매 회차 수험자 수는 제한적인 규모에 그칠 수밖에 없었다(본 특별호의 Lee & Jun, 2019 참조).

지필시험으로서의 기존 시험의 전통을 이어가는 틀 내에서 기존 TEPS를 개정하기 위한

연구는 사실상 2012년 시작되었다고 볼 수 있다. Kim et al.(2012)은 『정기 TEPS 전면개정을 위한 기초연구』를 통해 기존 TEPS에서 개선이 필요한 부분과 개정의 방향을 논의하고 제시하였다. 우선 기존 TEPS에서 개정이 필요한 부분으로는 (1) 평가틀 및 시험 청사진, (2) 지문당 다문항 문제 출제의 필요성과 영역별 문항 수, (3) 점수체계와 성적표였다. 이 중에서 특히 1지문 1문항 원칙이 시험개발에 큰 부담으로 작용하고 있는데, 지문의 수가 많은 만큼 각 지문의 단어 수에 제한이 있어 각 지문의 주제를 깊이 있게 다루기 어려우며, 지문의 주제들이 겹치지 않도록 출제하는 것도 출제의 어려움을 가중시킨다는 의견이 있었다. 1지문 1문항 원칙을 고수하는 것이 고부담 평가에서 반드시 필요하지는 않으며 시험개발의 효율성을 방해할 수도 있는 요인이므로 일부 개정이 필요해 보인다고 결론지었다.

이어 Jun et al.(2014)는 『텝스 2.0 개발을 위한 기초 연구』에서 TEPS의 개정 방향을 수요자 분석 결과와 타 영어능력평가 분석 결과를 토대로 논의하였다. 우선 수요자 분석을 통해 영어 시험의 주요 수요자 집단의 개선 요구사항을 파악하였다. 영어 시험 점수를 신규 직원 채용이나 기존 직원의 인사 평가에 활용하는 기업이나 공공기관의 채용 및 인사 담당자, 그리고 신입생 입시와 배치에 영어 시험 점수를 사용하는 대학 또는 교육기관의 입시담당자와 교원을 대상으로 설문 조사를 실시하였다. 분석 결과 기존 TEPS의 난이도를 조정하여 TEPS에 대한 응시자들의 거부감을 해소하고 TEPS의 강점과 개선사항을 적극 반영할 필요가 있으며, 타 시험과의 정확한 환산 기준을 마련하여 TEPS 응시자들의 불이익을 해소해야 한다고 결론지었다. 둘째로, 기존 시험과 주변 아시아 3개국(대만, 일본, 중국) 및 미국, 영국의 영어능력평가를 분석하였다. 기존 TEPS의 특징과 타 국가의 여러 영어능력평가의 특징을 비교분석하여 어느 시험들의 어떠한 특징들이 우수하고 타당도 높은 영어평가의 기준에 부합하는지 살펴보고 어떤 특징들을 TEPS 개정 시 고려할 수 있을지 검토하였다. 마지막으로, TEPS 2.0 개정안을 제시하였는데, 난이도 조절과 성취수준 설정 문제는 수험자의 부담을 덜어주는 방향으로 가야 한다고 결론지었고, 문법 및 어휘 영역의 유지 여부 문제, 듣기 영역 질문 및 보기의 활자화, 그리고 듣기 지문 2번 청취 또는 1번 청취에 대한 판단을 내려야 한다고 제안하였다. 또한 듣기 및 읽기 영역에서 1지문 다문항 유형 도입을 고려해야 한다고 제안하였다.

다음으로 Lee et al.(2015)의 『신 TEPS 개편사업』이 수행되었다. 먼저, 수험자 요구 분석, 타 영어시험의 비판적 비교 검토, 타당도 구축을 위한 문헌조사, 그리고 영어평가 전문가들과의 회의를 통해 다음과 같은 개편의 방향을 설정하였다. (1) 우리나라의 일반적인 수험자의 영어 능력을 고려하여 난이도를 보다 현실적인 수준으로 개선한다. (2) 시험 응시자의 최대 집중 가능 시간을 고려한 적절한 길이의 시험이 될 수 있도록 한다. (3) 보다 많은 수험자들이 응시할 수 있도록 대중 친화적인 시험으로 개편한다. (4) 시험 응시자의 영어 능력을 보다 타당하고 정확하게 평가할 수 있도록, 보다 혁신적이고 실제적인 듣기나 읽기 평가 문항을 개발한다. 개편 방향(4)에 맞추어 새로운 문항을 개발하였다(item prototyping). 듣기 영역에서는 대화와 일치하는 시각자료 고르기(dialogue-based visual task), 문제 해결 과제(problem-solving task), 강연 내용 메모・요약하기(note-taking task) 유형을 개발하였다. 읽기 영역에서는 스캐닝(scanning) 능력을 평가하는 매칭(matching) 유형과 1지문 다문항 유형을 개발하였다. 그리고 나머지 개편 방향에 맞추어 기존 TEPS의 문항 유형을 유지하면서 새로 개발된 문항을

포함시키고 문항 수와 시험 시간을 조절하여 듣기 영역 40문항(45분)과 읽기 영역 55문항(60분)으로 이루어진 평가틀과 샘플 문항을 제작하였다. 이 샘플 문항을 가지고 듣기 영역은 3차례, 읽기 영역은 4차례에 걸쳐 10여명에서 100여명에 이르는 다양한 응시자 집단을 대상으로 예비 타당성 연구를 시행하였다. 또한 새로 개발된 문항 유형과 듣기 영역의 속도 및 읽기 영역의 지문 길이 등에 대한 응시자의 의견을 시험 후 설문을 통해 수집하였다(Lee, Lee, & Jun, 2016a; Lee, Lee, & Jun, 2016b).

뒤이어 Lee et al.(2016)의 『신 TEPS 이행 예비 타당성 검증 사업』이 수행되었는데, 우선 신 TEPS 체제 이행에 필요한 준비 검증 조사를 실시하였다. 또한 문항 점검 및 예비 타당성 실험이 시행되었다. 신 TEPS 개편사업에서 개발된 평가틀에 기반한 듣기 40문항, 읽기 55문항의 완전한 세트를 제작하고, 이 세트를 사용하여 예비 시행을 실시한 후 응시자의 반응과 시험 결과를 분석하였다. 마지막으로 외부 전문가 중심의 문항 타당성 검증을 위해 8명의 외부 언어 평가 전문가에게 시험 평가틀과 시험 세트를 전달한 후 이메일 설문을 통해 문항 타당성에 관한 의견을 수집하는 연구를 실시하였다.

4. 현 개정 TEPS 연구에서의 평가틀 확립 과정

4.1. 본 TEPS 개정 연구의 범위 설정

그동안의 이전 연구에서는 새로운 시험의 성격을 논의하면서 다양한 가능성을 열어 놓고 논의했던 것이 사실이다. 2012년 이전 연구에서는 기존의 4개 영역(청해, 문법, 어휘, 독해)의 문항을 축소하는 수준을 넘어서 말하기, 쓰기 영역을 추가하고 더구나 지필 시험이 아닌 컴퓨터 기반시험으로 개발하는 안을 고려하였다(이는 추후 i-TEPS 개발로 열매를 맺음). 2012년 이후 연구에서도 비슷한 전략을 설정하여 다양한 안들이 논의되었다. 기존의 TEPS 문항을 축소하는 것을 넘어 TEPS-Speaking & Writing(TSW) 시험을 일부 개정하여 추가하는 안부터 문법과 어휘 영역을 없애고 그 일부 문항 유형들을 나머지 청해・독해 영역으로 흡수하여 2개 영역 체제로 개편하는 안 등이 논의되었다.

본 연구는 이러한 사전 연구 논의 안들에 대한 심층적인 검토와 논의의 토대 위에서 출발하였다. 시대의 추세를 잘 반영하기는 했지만 응시자의 수나 점수 활용도 면에서 매우 제한적이었던 i-TEPS의 전례에 대한 비판적인 고민이 있었다. 또한 현재의 4개 영역 체제를 급격히 2개 영역 체제로 변경할 경우에는 기존 TEPS와의 연결성이 약해질 수 있어 오랜 시간 기존 TEPS 체제 하에서 축적해온 연구 성과 및 시행 노하우를 활용하기 어렵다는 점도 고려되었다. 특히 그렇게 할 경우 그동안 기존 TEPS를 위해 문항 데이터 베이스에 축적한 많은 양의 문항이 낭비될 수도 있고 새로이 개발된 시험을 안정화 시키고 공인을 받는 데 예상보다 훨씬 많은 시간이 걸릴 수 있어 수험자들에게 혼란을 줄 수 있는 점도 고려하였다. 또한 외국어로서 영어를 습득해야 하는 국내 환경의 영어학습에서는 언어의 기본이 되는 어휘와 문법이 중요하고 이를

평가하여 영어교육과 학습에 기여할 수 있다고 판단하였다. 따라서 완전히 새로운 시험을 개발하기 보다는 기존 TEPS에 대해 수험자가 그동안 요청한 개선 요구 사항, 최근 평가 분야의 이론적 추세 변화, 그리고 변화된 의사소통 상황을 반영하여 시험을 개정하는 쪽으로 큰 방향을 설정하였다.

이러한 논의에 기반하여 본 연구에서는 새로이 개발하는 TEPS시험은 완전히 새로운 시험이 아니고 기존 TEPS의 전통을 유지하는 범위 내에서 시험을 개정하는 것으로 큰 방향을 잡았고 기존 TEPS와 개정 TEPS 시험은 동일한 구인(construct)을 평가하는 시험으로 정의하였다. 이러한 방향에 부합하도록 개정 TEPS는 기존 TEPS의 총 4개 영역 체제(청해, 어휘, 문법, 독해)를 유지하고 거의 모든 문항 유형을 그대로 포함한다. 다만 영어 평가 이론의 변화, 시대의 변화, 그리고 수험자의 부담에 대한 고려사항을 반영하는 새로운 문항유형들을 일부 추가하고자 하였다. 또한 이전 개편 연구들의 결론을 분석하여 수용 가능한 점을 개정 TEPS에 반영하기로 결정하였다(<표 1> 참조).

표 1. TEPS 개정 초반 연구 결과의 평가틀 반영사항

	항목	평가
수용 가능한 점	문항 수와 시험 시간 축소	심리적 난이도 조절이 가능함
	1지문 다문항 유형 중 1지문 2문항 유형	출제자에게 큰 부담 없이 시험의 진정성을 높일 수 있음
	청해 영역 대화 1회 청취	대화 문항에 한정하여, 같은 대화를 두 번 들으며 생기는 피로도를 낮출 수 있음. 대신 상황 제시 문구를 대화 전에 제시하면 수험자가 어떤 내용을 기대해야 할지 파악 가능
수용하기 어려운 점	절대적 난이도 변경	기존 수험자 집단에게 혼란을 가져옴
	청해 영역의 대화와 일치하는 시각자료 고르기, 문제 해결 과제, 강연 내용 메모・요약하기와 읽기 영역의 스캐닝 문항 등 완전히 새로운 문항 유형	기존 유형과 너무 다름. 출제 및 난이도 조절의 어려움
	매우 긴 지문이 포함된 1지문 다문항 유형	출제 및 난이도 조절의 어려움
	문법 및 어휘 영역의 폐지	시험의 내용이 크게 바뀌어 기존 수험자 집단에게 혼란을 가져옴
	청해 영역 질문 및 보기의 활자화	읽기 능력이 듣기 능력 평가에 미치는 영향을 최소화하고 요령을 통한 정답 유추를 방지하여 결과적으로 청해 영역의 내용 타당도를 높인다는 TEPS의 원칙이 바뀜

4.2. 개정 TEPS 평가틀 개발 및 조정

개정 TEPS가 기존 TEPS와 동일한 구인을 평가하도록 하기 위해 기존 TEPS의 4영역 체제를 유지하면서 문항 수를 축소하는 방식으로 개정 TEPS 평가틀을 구성하였다. 이는 개정 TEPS가 기존 TEPS의 각 영역별 문항 유형 분포, 문항 소재 분포, 문항 난이도 분포를 최대한 동일하게 유지하는 범위 내에서 문항 유형별(파트 별) 문항 수, 세부능력・세부지식 별 문항 수, 문항 소재별 문항 수, 난이도별 문항 수를 축소했음을 뜻한다. 135문항으로 이루어진 i-TEPS 모델이 존재하지만, 일부 문항 유형이 빠졌던 i-TEPS와는 달리 기존 TEPS의 모든 유형을 유지시켰고, 여기에 1지문 2문항 유형을 청해 영역에 4문항, 독해 영역 10문항씩 각각 추가하였다.

기존 TEPS와 개정 TEPS의 평가틀에서의 일반적인 차이점은 우선 <표 2>와 같이 요약할 수 있다. 각 영역의 문항 수가 축소되었으며 각 영역의 시험 시간도 축소되었고 점수척도도 0-600점 체계로 변경되었다.

표 2. 개정 전・후 TEPS 하위영역의 문항 수, 검사 시간 및 점수 범위

하위영역	개정 전			개정 후		
	문항 수	검사 시간	점수 범위	문항 수	검사 시간	점수 범위
청해	60	55분	4-396	40	40분	0-240
어휘	50	15분	1-99	30	25분	0-60
문법	50	25분	1-99	30		0-60
독해	40	45분	4-396	35	40분	0-240
합계	200	140분	10-990	135	105분	0-600

또한 각 영역별 변경사항은 다음과 같이 요약할 수 있다.

청해 영역

- 파트 별 문항 수 축소
- 파트 3의 대화 및 질문 청취 횟수가 2회에서 1회로 변경
- 1지문 2문항으로 구성된 파트 5 추가

어휘 영역・문법 영역

- 파트 별 문항 수 축소
- 각 영역 내 대화 문항 및 단문 문항 비율 변경
- 두 영역의 시험 시간을 통합하여 시행
- 두 영역의 순서가 기존 문법→어휘에서 어휘→문법으로 변경
- 문법 영역의 파트 3과 4가 파트 3으로 통합

독해 영역

- 파트 별 문항 수 축소
- 파트 2와 파트 3의 순서 변경
- 1지문 2문항으로 구성된 파트 4 추가
- 일부 지문에 실제를 반영하는 다양한 디자인 도입

각 변경사항에 대한 사유, 판단 근거, 그리고 결정 과정은 다음과 같다.

1) 각 영역의 문항 수 축소 및 각 영역 내 파트 별 문항 수 축소

총 135문항으로 이루어진 평가틀 변경에는 다음과 같은 측정학적 지표들과 내용적인 측면들이 동시에 고려되었다.

- 연구에 기반하여 개발된 i-TEPS의 문항 수(청해 40문항, 문법 30문항, 어휘 30문항, 독해 35문항, 총 135문항) 모델이 이미 존재하기 때문에 시험 세트 제작 시 용이하다. 또한 i-TEPS를 수년간 시행한 자료를 분석한 결과, 135문항으로도 충분히 수험자의 영어능력을 정확하게 평가할 수 있다는 사실을 알 수 있었다. 다만 일부 문항 유형이 제외된 i-TEPS와는 달리 개정 TEPS에서는 시험의 연속성을 위해 기존 TEPS의 모든 문항 유형을 유지한다.
- 문항 수 축소를 통해 수험자가 느끼는 심리적 난이도 조절을 목표로 한다.
- 어휘와 문법은 총점 반영 비율과 배점이 낮으므로 문항 수가 청해와 독해만큼 많을 필요는 없으나 기존 TEPS의 평가틀을 충분히 반영할 수 있도록 최소 길이를 유지한다.
- Bell과 Lumsden(1980)에 따르면 검사 길이가 줄더라도 예측타당도는 크게 영향받지 않는다.
- Kim(2016a)은『신뢰도와 시험 길이(Reliability and test length)』에서 각 영역별 문항 축소 비율에 따른 예상 신뢰도를 산출하고, 개정 TEPS의 검사 길이에 대해 135문항, 140문항, 160문항의 세 가지 안을 제시하였다. 각 안의 장단점을 비교한 결과 135문항으로 축소했을 때에도 기존 TEPS 대비 총점 신뢰도는 거의 변화가 없다는 점, i-TEPS 모델을 반영하여 문항 출제 및 검사 구성의 안정성을 높일 수 있다는 점을 고려하여 135문항으로 축소하는 안을 채택하였다.
- 기존 TEPS의 영역별 신뢰도 평균을 바탕으로 문항 수 축소에 따른 예측 신뢰도 곡선을 산출하여 각 영역별 문항 수 증가에 따른 신뢰도 증가의 포화점(saturation point)을 고려하였다. 영역마다 차이가 있으나 전반적으로 신뢰도 0.70 이후로 문항 수 증가에 따른 신뢰도 증가 추이가 급격히 둔화됨을 확인하였다(<그림 1> 참조).

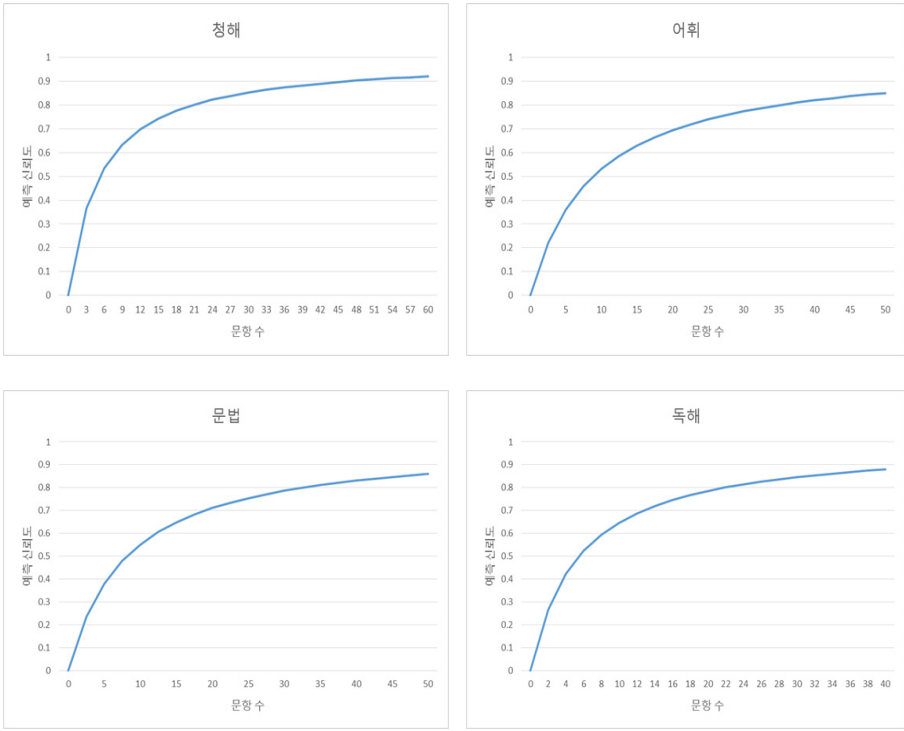

그림 1. 기존 TEPS 4개 영역의 문항 수 증감에 따른 신뢰도 변화 예측치

- 기존 TEPS의 총점 신뢰도 평균을 바탕으로 문항 수 축소에 따른 예측 신뢰도 곡선을 산출하였다. 분석 결과 총 120문항일 때 예측 신뢰도는 0.904로 대규모 시험에서 요구되는 0.90 신뢰도 유지를 위한 최소한의 문항 수가 120문항임을 파악하였고, 총 135문항일 때 예측 신뢰도는 0.914로 적정 신뢰도 수준을 유지하기에 충분하다는 것을 확인하였다. 또한 총점에 대한 예측 신뢰도는 영역별 문항 수를 같은 비율로 축소한다고 가정했을 때의 신뢰도를 예측한 것으로, 영역별 신뢰도가 상대적으로 낮은 어휘와 문법 축소비율이 높고 청해와 독해 축소비율이 낮으므로 실제 총점 신뢰도는 예측 신뢰도보다 높게 산출될 것으로 예상하였다(<그림 2> 참조).

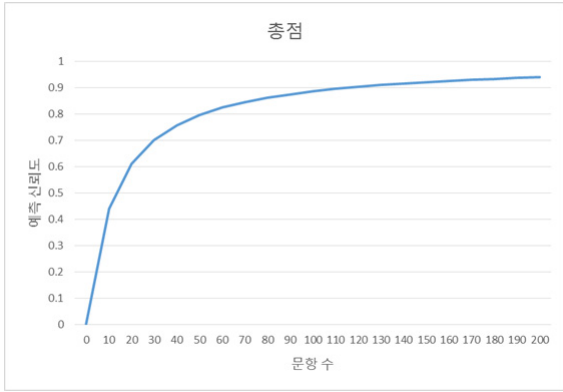

그림 2. 기존 TEPS 전체 시험 문항 수 증감에 따른 신뢰도 변화 예측치

2) 각 영역의 시험 시간 축소

각 영역별 축소된 문항 수에 비례하여 시험 시간을 축소하였다. 그 결과 문항당 시험 시간의 평균이 유사하다(<표 3> 참조). 그리고 미세하지만 문항당 주어진 시간이 조금씩 길어져 수험자의 부담을 경감할 수 있다.

표 3. 기존 TEPS와 개정 TEPS 시험 시간 비교

영역	기존 TEPS 시험 시간 (문항당 평균 시험 시간)	개정 TEPS 시험 시간 (문항당 평균 시험 시간)
청해	55분(0.917분)	40분(1.000분)
어휘	15분(0.300분)	25분(0.417분)
문법	25분(0.500분)	
독해	45분(1.125분)	40분(1.143분)
합계	140분(0.700분)	105분(0.778분)

3) 0-600점의 척도 점수 체계로 변경

Kim(2016b)은『시험 길이와 점수 척도(Test length & score scale)』에서 문항별 가중치, 척도의 안정성, 수험자의 점수 해석을 고려하여 0-400점, 0-600점, 0-1000점이라는 세 개의 점수 척도를 제안한 바 있다. 0-1000점 척도를 유지할 경우 기존 TEPS 점수체제와 연속성이 있으나 문항 수에 비해 가능한 점수 포인트가 너무 많고 문항당 가중치가 너무 커지므로 매 회차 점수 변동 폭이 커질 수 있다. 이러한 점을 고려하여 0-400점과 0-600점 척도를 적용했을 때 어떤 변화가 있는지 측정학적 분석을 통해 살펴보았다. 0-400점 척도의 경우 기존 TEPS의 영역별 가중치를 동일하게 유지했을 때 문항당 척도점수 가중치가 크게 바뀌므로 점수 산출 과정에서의 오차 발생 가능성 및 점수 해석의 어려움이 있다고 판단되었다. 0-600점 척도의

경우 문항당 척도점수 가중치가 기존 TEPS 점수체제에서의 가중치와 크게 달라지지 않고, 척도 범위 변경 정도가 검사 길이 축소 비율과 비슷한 수준이므로 점수 해석에 있어서 수험자들의 혼란을 최소화할 수 있다는 점을 고려하여 0-600점 척도를 최종 선정하게 되었다.

4) 청해 영역 파트 3의 대화 및 질문 청취 횟수가 2회에서 1회로 변경

Jun et al.(2014)의 『텝스 2.0 개발을 위한 기초 연구』에서 청해 영역 지문을 한 번만 들려주는 것을 제안하였다. 실제 커뮤니케이션 상황을 반영하기 위해 모든 대화는 한 번씩 들려주는 것으로 통일하였다. 대화 및 질문 청취 횟수를 2회에서 1회로 변경하면 같은 대화를 두 번씩 들으며 생기는 피로도를 낮출 수 있다. 대신 청취 횟수가 줄어든 점을 보완하기 위해 대화를 듣기 전에 간략하게 대화 상황이 제시되는데 이를 통해 수험자가 대화에서 어떤 내용을 기대해야 할지 파악이 가능하다.

5) 청해 영역과 독해 영역에 1지문 2문항 유형 추가

수험자가 실제로 듣거나 읽어야 하는 담화나 글의 분량을 반영하고 다면적이고 종합적인 지문 이해 능력을 평가하기 위해 1지문 2문항 유형을 추가하였다. Kim et al.(2012) 『정기 TEPS 전면개정을 위한 기초연구』, Jun et al.(2014) 『텝스 2.0 개발을 위한 기초 연구』, Lee et al.(2015) 『신 TEPS 개편사업』에서 1지문 다문항 유형의 도입을 제안하였다. 또한 Choi(2011)와 Yi(2013)는 실생활에서 사람들이 긴 글을 읽어야 하는 경우가 많기 때문에 긴 지문에 대해 여러 개의 질문을 하는 것이 더 진정성 있는(authentic) 시험이라고 하였고, TEPS에서 1지문 다문항 유형을 도입하는 것을 제안하였다. 1지문 다문항 유형은 대다수의 세계적인 영어 시험들(TOEFL, IELTS, GEPT, Cambridge 영어 시험, MELAB, CAEL, TOEIC 등)에서 사용하고 있다. 이러한 점들을 고려하여 신뢰도 유지에 중요한 1지문 1문항 유형을 다수 유지하면서 출제자에게 큰 부담 없이 시험의 진정성을 높일 수 있는 1지문 2문항을 일부 추가하였다. 단 센터 외부에서 초빙된 교육평가 전문가는 1지문 2문항 유형에서 문항반응이론의 국부독립성 가정이 유지되는지 파일럿 시험을 통해 알아볼 것을 제안하였다. 이에 대한 대응으로 개정 TEPS 1차 파일럿 시험에 사용된 1지문 2문항에 대한 측정학적 통계분석을 실시하였고 국부독립성 가정이 위배된 일부 문항을 발견하였다. 이러한 과정을 통하여 확인된 국부독립성 위반 문항 짝에 대한 연계 지문과 질문 및 선택지의 내용분석(Kim & Jun, 2017)을 통해 그 원인 제공 요소들을 파악하였고 이러한 분석 결과를 문항 출제자 지침에 반영하였으며 센터 내부 출제자를 대상으로 워크샵도 실시하였다.

6) 어휘 영역과 문법 영역 내 대화 문항 및 단문 문항 비율 변경

어휘 지식은 글쓰기 능력을 포함한 모든 언어 능력과 직결되고(Carter & McCarthy, 1988; Nation, 1990), 풍부한 어휘 구사력(a large vocabulary size)의 필요성에 대한 근거가 늘어나고 있다(Nation, 2011). 또한 문법 능력은 정확한 문장 구성력과 직결된다(Muncie, 2002). 따라서 어휘와 문법 평가는 영작문에 필수불가결한 올바른 어휘 구사력과 정확한 문장 구성력의 간접 평가이며 이를 위한 학습을 유도한다. 최근 글쓰기 능력의 필요성이 점점 더 커지고 있음을

감안하여 어휘 영역과 문법 영역에서 이를 간접적으로 평가하는 문항 비중을 확대하였다. 따라서 대화 문항과 단문 문항의 비율이 어휘 영역에서는 25문항, 25문항(1:1)에서 10문항, 20문항(1:2)으로, 문법 영역에서는 25문항, 25문항(1:1)에서 12문항, 18문항(2:3)으로 변경되었다.

7) 어휘 영역과 문법 영역의 시험 시간을 통합하여 시행

대부분의 대단위 표준화 인지능력 시험들은 속도시험(speed test) 보다는 역량시험(power test)을 지향한다. 역량시험이란 대부분의 수험자가 출제된 모든 문항들을 읽거나 듣고, 풀고, 그리고 답할 수 있는 충분한 시험 시간이 주어져야 한다는 의미이다. 하지만 영어능력시험의 경우에는 실생활에서 실시간으로 일어나는 의사소통 수행 능력을 평가하기 때문에 영역에 따라서는 속도화(speededness)의 요소를 일정 정도 도입하는 것이 불가피하며 오히려 그렇게 하는 것이 시험의 진정성(authenticity)과 예측타당도(predictive validity)를 높이는 것이라는 주장도 제기되고 있다(Choi, 1997, 1999; Oller, 1995).

기존 TEPS의 경우에는 전체적으로 볼 때 개정 TEPS보다는 훨씬 많은 수의 문항 제시와 더 짧은 문항당 응답시간 설정을 통해 속도화 시험의 요소를 가미해 수험자의 내재화되고 자동화된 영어숙달도를 평가하고자 하는 성격을 일정 정도 가지고 있었다고 할 수 있다. 다만 시험에 속도화 요소가 과도할 경우 일부 문항들이 수험자의 진정한 영어능력을 평가하는 기능을 제대로 수행하지 못할 수도 있다. 아울러 그동안 수차례 이루어진 수험자 요구 조사에서도 수험자가 어휘 영역에 할당된 시간의 부족으로 인해 부담을 호소해 왔다는 점이 반복적으로 확인되었다. 이러한 문제를 해결하는 방안으로 어휘와 문법 두 영역을 한 영역으로 통합하지는 않지만 시험 시간 관리 측면에서는 통합하여 운영하기로 결정하였다. 이러한 변화는 어휘 영역의 시간 부족 관련 수험자 부담을 덜어주는 효과뿐만 아니라 두 영역의 문항당 시간제한을 공평하고 균형적으로 설정할 수 있다는 장점이 있다. 원래 기존 TEPS는 문항당 평균 시험 시간이 각각 어휘 0.3분, 문법 0.5분이었으나, 이렇게 두 영역의 시험 시간을 통합운영 할 때에는 어휘・문법 영역의 문항당 시험시간이 평균 0.417분으로 통일된다.

두 영역 시험 시간 통합운영 결정에는 수험자가 집중할 수 있는 시험 환경 조성이라는 실질적인 효과도 고려되었다. 기존 TEPS의 시행 절차에 따르면 모든 수험자는 감독관의 감독 하에 한 영역을 마친 후에는 반드시 시험을 일단 중단해야 하고 다시 감독의 신호가 있을 때에만 그 다음 영역을 일제히 시작할 수 있다. 만약 개정 TEPS시행 시 어휘 영역과 문법 영역의 시험 시간을 이처럼 분리하여 운영하면 총 배정 시간이 어휘영역은 9분, 문법영역은 15분이 된다. 현 시험 감독 지침에 의하면 감독관은 각 영역 종료 5분 전, 1분 전, 그리고 마지막으로 종료 시점을 수험자에게 구두로 공지하게 되어 있고 종료선언 직후 수험자들이 시험을 완전히 중단하였는지 확인하기 위해 시험실을 돌도록 되어있다. 이미 두 영역에 배당된 수험시간이 짧아진 상황에서 이러한 시간고지 규칙을 적용할 경우 수험자의 문제를 푸는 흐름을 깨고 집중력을 현저히 저하시킬 수 있다. 감독관의 입장에서도 수험자들이 다른 영역의 문제를 풀고 있지 않은지 일일이 확인하는데 더 주의를 기울여야 한다. 이러한 시행상의 혼란을 방지하기 위해서도 두 영역의 시험 시간을 통합해야 한다는 필요성이 제기되었다.

다만 교육평가 외부전문가는 어휘 영역과 문법 영역의 영역 점수를 따로 부여하려면 두 영역은

별개의 영역으로 구분되어야 한다고 하였다. 따라서 시행상 어휘 영역과 문법 영역의 시험 시간은 통합되지만 두 영역이 별개의 영역임을 확실히 하기 위해 시험지 디자인에서도 별도의 영역으로 표시하여 구분하고, 문항 번호도 어휘영역이 끝나면 문법 영역이 새로 1번으로 시작되도록 시험지와 답안지를 제작하였다. 또한 이러한 변화가 실제로 시험 점수의 측정학적 질에 끼치는 잠재적 영향을 살펴보기 위해서 어휘 영역과 문법 영역의 시험 시간이 별개로 시행되었던 기존 TEPS 시험 결과와 어휘와 문법 영역의 시험 시간을 통합하여 시행한 1, 2, 3차 파일럿 및 필드 테스트 시험 결과를 비교 분석하였다. 분석결과 두 영역의 통합운영과 순서의 변화가 문항 특성 및 검사 특성에 영향을 주지 않는 것으로 확인되었다(Kwon et al., 2018).

8) 어휘 영역과 문법 영역의 순서 변경

언어를 습득할 때 어휘 덩어리(lexical chunks)로 먼저 습득하고 문법성은 그 후 자연스럽게 파악하게 한다는 언어습득 및 교육 이론인 어휘접근법(Lexical Approach; Lewis, 1993)을 반영하여 개정 TEPS에서는 어휘를 먼저 평가한 후 문법을 평가한다. 또한 긴 대화와 문단으로 구성된 문법 영역 파트 3 다음에 독해 영역으로 바로 이어질 수 있게 배치하였다. 문법-어휘 순으로 검사지를 구성했던 1, 2차 파일럿 시험 결과와 어휘-문법 순으로 시행된 3차 및 필드 테스트 시험 결과를 비교했을 때 문항 특성 및 검사 특성에 차이가 나타나지 않는 것을 확인하였다(Kwon et al., 2018).[1)]

9) 문법 영역의 파트 3과 4를 파트 3으로 통합

기존 TEPS 문법 영역의 파트 3(대화에서 문법상 틀리거나 어색한 부분 고르기)과 파트 4(문단에서 문법상 틀리거나 어색한 부분 고르기)를 각각 5문항, 5문항에서 2문항, 3문항으로 축소한 결과로 문항 수가 적은 반면 두 파트가 대화나 문단에서 문법상 틀리거나 어색한 부분을 고르는 같은 유형이므로 하나의 파트로 통합하여 시행해도 무리가 없다고 판단하였다.

10) 독해 영역 파트 2와 파트 3의 순서 변경

평가틀에 대한 자문의견을 제공했던 외부전문가는 지문을 읽고 문맥상 어색한 내용 고르기 유형으로 이루어진 기존 파트 3은 응집성을 평가하므로 파트 1(지문을 읽고 빈칸에 가장 적절한 답 고르기)의 마지막 두 문항이자 응집성을 평가하는 문항인 빈칸에 들어갈 연결어 고르기 유형 바로 다음에 이어지도록 해당 파트의 위치를 파트 2로 변경할 것을 제안하였다. 같은 능력을 평가하는 문항 유형을 한 데 모으는 것은 평가 제작 이론상 문제가 없고 오히려 권장되는 방법이다.

11) 독해 영역의 일부 지문에 실제를 반영하는 다양한 디자인 도입

Multicampus(2017a)는『NEW TEPS 분석 및 의견』에서 다양한 시각자료의 활용을 통해

1) 참고로 시험 내에서 영역의 순서를 바꾸는 것이 시험 결과에 영향을 미치지 않는다는 Graduate Management Admission Council(GMAC)의 연구 결과를 바탕으로 경영대학원 입학 시험인 Graduate Management Admission Test(GMAT)는 수험자가 영역의 순서를 선택하도록 하고 있다(GMAC, 2017).

시험지의 매력도를 높이면 수험자의 흥미도가 증가하고 시험 난이도 하락으로도 느껴질 수 있다는 의견을 제시하였다. 또한 TEPS가 첫 시행되었던 1999년 이후 이메일, 스마트폰, 태블릿 컴퓨터, 인터넷 신문, 블로그 등 IT 기술의 발달로 새로운 형식의 의사소통 상황이 만들어졌고, 시험에서 이러한 변화된 의사소통 상황과 언어사용을 반영할 필요가 생겼다. 독해 지문에 이러한 실제를 반영하는 다양한 디자인을 도입하여 진정성을 높이면서 실생활에서 사람들이 읽는 글이나 문서와 유사하게 보이도록 지문을 구성하였다. IELTS, TOEIC 등 세계적인 영어 시험의 독해 지문에도 실제를 반영하는 디자인이 적용되어 있다. 처음에는 독해 영역의 모든 파트에 디자인을 도입하는 것을 고려하였으나 외부전문가는 빈칸이 포함되어 있는 글에 디자인을 도입하는 것은 실생활에서는 볼 수 없는 것이라 진정성이 부족하다고 하였다. 따라서 빈칸 채우기 유형으로 이루어진 독해 파트 1과 문맥상 어색한 내용 고르기 유형으로 이루어진 파트 2에는 디자인을 도입하지 않기로 결정하였다.

12) 영역별 문항 항목 분포

개정 TEPS의 영역별 문항 항목 분포는 기본적으로 기존 TEPS의 항목 분포와 최대한 비슷하게 유지하는 틀에서 항목별 문항 수를 축소하는 것을 원칙으로 삼았다. 여기서 문항 항목이란 언어 기능(function), 상황(situation), 주제(topic), 그리고 요소(element)를 의미한다. 예를 들어 청해 영역에는 안부 묻기, 약속 잡기, 칭찬하기, 부탁하기 등의 언어 기능에 따른 문항 수 분포가 있고 문법 영역에는 시제, 수 일치, 어순, 분사구문 등의 문법 요소에 따른 문항 수 분포가 있는데, 이러한 분포들을 최대한 그대로 유지하면서 각 영역별 문항 수를 축소하였다.

4.3. 파일럿 시험과 필드 테스트를 통한 연구 결과

연구프로젝트 초기의 다양한 노력을 통해 준비된 개정 TEPS의 평가틀은 이후 수 차례의 파일럿 및 필드테스트 과정을 거쳐 추가로 다듬어지고 개선되었다. 매번의 사전 시험 당시 평가틀에 기반해서 실험적으로 시험 세트들이 제작되었고 일정한 표집계획에 맞춰 모집된 수험자들을 대상으로 실시하였다(본 특별호의 Lim, 2019와 Lim et al., 2019 참조). 이러한 수험자의 반응과 데이터 분석 결과를 가지고 문항틀을 계속적으로 수정하는 과정을 거쳤는데 그 내용을 간략히 요약하면 다음과 같다.

1) 1차 파일럿 시험

- 1차 파일럿 시험에서는 기존 TEPS의 평가틀과 측정하고자 하는 구인이 개정 TEPS에서도 유지되는지 점검하고자 하였다. 1차 파일럿 시험 결과를 바탕으로 요인분석을 실시한 결과 기존 TEPS와 같은 요인구조를 보여주었다.
- 1차 파일럿 시험 분석 결과 새로 개발된 1지문 2문항 유형 일부 문항에서 문항반응이론의 국부독립성(혹은 지역독립성) 가정이 위배되는 것을 발견하였고, 이를 해결하기 위해 국부의존성에 영향을 미치는 하위능력 문항 짝에 대한 분석(Kim & Jun, 2017)을 통해 출제자 지침을 마련하고 내부 출제자 워크샵을 실시하였다.

2) 2차 파일럿 시험

- 2차 파일럿 시험에서는 TEPS 개정 시기에 수험자들이 기존 TEPS와 개정 TEPS 점수를 상호 호환하여 사용할 수 있도록 점수 환산표를 개발하였다(KELTA, 2018). 이를 위하여 2차 파일럿 시험은 기존 TEPS 수험자 집단과 가급적 유사한 집단을 선발하여 시행되었다.
- 1차 파일럿 시험 결과를 바탕으로 1지문 2문항 유형의 국부의존성 해결을 위한 후속 조치가 있었고, 2차 파일럿 시험에서는 해당 유형의 국부독립성 가정이 위배되지 않는다는 분석 결과가 산출되었다.

3) 3차 파일럿 시험

- 3차 파일럿 시험에서는 문법 영역과 어휘 영역의 순서가 변경되었지만 요인분석을 실시한 결과 기존 TEPS 그리고 1, 2차 파일럿 시험과 동일한 요인구조를 보였다.
- 3차 파일럿 시험에 추가된 독해 지문 디자인에 대한 응시자들의 의견을 종합해 볼 때 중립적이거나 긍정적인 평가가 부정적인 평가보다 훨씬 많았다.

4) 필드 테스트

- 1-3차 파일럿 연구 결과를 바탕으로 확립된 최종 평가틀을 기존 TEPS 수험자 집단과 유사한 집단을 대상으로 최종 점검하기 위해 필드 테스트를 시행하였다.
- 문항 분석 결과 1-3차 파일럿 시험 결과와 마찬가지로 난이도, 변별도, 신뢰도 등의 수치가 기존 TEPS와 유사하였고 기존 TEPS와 동일한 요인구조를 보였다.

5. 맺는 말

새로운 시험을 개발하거나 기존의 시험을 개정하는 연구는 여러 단계에 걸친 일련의 토의, 조정, 결정 과정을 요구한다. 본 연구 초기에는 개정 TEPS 평가틀의 전체적 방향을 정하고 기본구조를 설계하는 작업이 이루어졌는데 이 과정에서 사전연구 검토, 언어평가・교육평가 분야의 최근의 이론적 추세 분석, 그리고 수험자의 요구 조사 분석 결과 등이 그 주요한 토대가 되었다. 이러한 과정을 거쳐 제작된 평가틀은 개정 TEPS의 시험 세트들을 제작하는 데 실험적으로 활용되었고 또 이러한 시험 세트를 사용하여 수 차례의 파일럿 및 필드테스트가 실시되었다. 이러한 평가틀 제작과정에서 언어교육원 TEPS센터 외부의 관련 전문가들에게 검토 및 자문을 의뢰하였고 그 조언 및 자문 내용도 평가틀 수정에 일부 반영되었다. 본 논문은 이 모든 과정에서 다양한 이해당사자들로부터 수합한 의견들이 개정 TEPS에 어떻게 반영되었는지를 논의하였다.

최근의 주요 타당도 검증 이론가들은 시험 개발 및 개정 과정에서 이루어진 주요한 결정과 그 이론적, 논리적, 경험적 근거들뿐만 아니라 시험의 다양한 이해당사자(stakeholder) 집단의 의견을 청취하고 반영하는 과정들을 문서의 형태로 기술하는 것이 바람직하다고 제언한다

(Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Kane, 2013). 왜냐하면 이러한 정보들이 시험 타당도 검증의 중요한 증거로 활용할 수도 있기 때문이다. 이러한 측면에서 볼 때 본 논문에 기술된 여러 내용도 추후 개정 TEPS의 타당도 검증의 중요한 근거 자료로 활용될 수 있을 것으로 기대된다.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163-181.

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bell, R., & Lumsden, J. (1980). Test length and validity. *Applied Psychological Measurement* 4.2, 165-170.

Carter, R., & McCarthy, M. (1988). *Vocabulary and language teaching*. Essex: Longman.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York: Routledge.

Choi, I.-C. (1997). Essential test method facets of a general English proficiency test and their validity as perceived by test-takers. *Language Research*, *33*(4), 773-799.

Choi, I.-C. (1999) Test fairness and validity of the TEPS. *Language Research*, *35*(4), 571-603.

Choi, I.-C. (2008). *Review of new TEPS development* (Research Report No. 52). Seoul: SNU Language Education Institute.

Choi, I.-C., Son, C. Y., & Ahn, J. (2008). *Development of table of specifications for the new TEPS* (Research Report No. 55). Seoul: SNU Language Education Institute.

Choi, M. S. (2011). *Analysis of the factors affecting passage dependency of multiple-choice English reading comprehension tests* (Unpublished master's thesis). Ewha Womans University, Seoul, Republic of Korea.

Davis, J., & Ferdous, A. (2005). Using item difficulty and item position to measure test fatigue. Retrieved from http://www.air.org/sites/default/files/downloads/report/AERA2005Test_Fatigue11_0.pdf

GMAC. (2017). Select Section Order & Removal of Test Center Profile Update. Frequently Asked Questions around the 2017 GMAT Release. Retried from https://www.gmac.com/frequently-asked-questions/gmat-select-section-order.aspx

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Jun, Y. C., Ryu, D.-S., Park, Y. J., Lee, Y., Shin, S.-H., Jun, H. . . . Byun, J. (2014). *Research for the development of TEPS 2.0* (Research Report No. 77). Seoul: SNU Language Education Institute.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of*

Educational Measurement, *50*(1), 1-73.

KELTA. (2018). *A study on the compatibility and score conversion between the revised TEPS and other certified language proficiency tests*. Seoul: KELTA.

Kim, J. Y. (2016a). *Reliability and test length* (internal document). Seoul: TEPS Center, Language Education Institute, Seoul National University.

Kim, J. Y. (2016b). *Test length and score scale* (internal document). Seoul: TEPS Center, Language Education Institute, Seoul National University.

Kim, J. Y., & Jun, H. (2017). An investigation of local item dependence in testlets and its causes in a large-scale English proficiency test. *Journal of Educational Evaluation*, *30*(4), 837-858.

Kim, M.-W., Jeong, S., Kim, J.-W., Lee, Y.-W., Lee, Y., Shin, S.-H. . . . Lee, G. (2012). *Basic research for a complete revision of the regular TEPS* (Research Report No. 73). Seoul: SNU Language Education Institute.

Kwon, H., Lee, Y.-W., Lee, Y., Park, Y.-J., Kim, J., Jun, H. . . . Park, H. (2018). *Development and validation of a pilot test form for the revised TEPS* (Research Report No. 80). Seoul: SNU Language Education Institute.

Lee, B., Kim, C., Park, Y. J., So, Y.-S., Lee, Y., & Jun, H. (2016). *Verification of preparation for New TEPS* (Research Report No. 79). Seoul: SNU Language Education Institute.

Lee, B., Lee, Y.-J., & Jun, H. (2016a). Evidence supporting a validity argument for an English listening comprehension test: Two prototyping studies. *The Mirae Journal of English Language and Literature*, *21*(4), 311-342.

Lee, B., Lee, Y.-J., & Jun, H. (2016b). The validity of the new item types of the listening section of an English proficiency test. *Secondary English Education*, *9*(4), 141-165.

Lee, B., Park, Y. J., So, Y.-S., Lee, Y.-J., Kim, C., Jun, H. . . . Yeom, S. (2015). *The development of New TEPS* (Research Report No. 78). Seoul: SNU Language Education Institute.

Lee, Y.-W., Anderson, P., Son, C. Y., Bong, J. S., Ahn, J., Park, M. K. . . . Ahn, H. (2009). *Research on the production, administration, and analysis of an i-TEPS pilot test and preparation for test operationalization* (Research Report No. 57). Seoul: SNU Language Education Institute.

Lee, Y.-W., Kim, S., & Moon, Y. (2008). *A preliminary psychometric investigation into optimal scenarios of section structure restructuring for New TEPS* (Research Report No. 49). Seoul: SNU Language Education Institute.

Lee, Y., Lee, J., Kim, J., Lee, C., & Lee, H. (2008). *Branding and marketing strategies for the TEPS* (internal document). Seoul: TEPS Council, Seoul National University Foundation.

Lewis, M. (1993). *The Lexical Approach: The state of ELT and a way forward*. Hove, UK: Language Teaching Publications.

Lord, R. G. (1985). An information processing approach to social perceptions, leadership and behavioral measurement in organizations. *Research in Organizational Behavior*, *7*, 87-128.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford: Information Age Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: American Council on Education and Macmillan.

Multicampus. (2017a). *Analysis of the New TEPS and feedback* (internal document). Seoul: Multicampus Co. Ltd.

Multicampus. (2017b). *Feedback from sales representatives on the previous TEPS* (internal document). Seoul: Multicampus Co. Ltd.

Muncie, J. (2002). Finding a place for grammar in EFL composition classes. *ELT Journal*, *56*(2), 180-186.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle Publishers.

Nation, I. S. P. (2011). Vocabulary research into practice. *Language Teaching*, *44*(4), 529-539.

Oller, J. W. Jr. (1995). Review of content and construct validation of a criterion-referenced English proficiency test. *English Teaching*, *50*(3), 161-168.

Ryu, D.-S., Park, Y.-Y., Kwon, H.-S., Song, M.-J., Min, E.-K., Ahn, J. . . . Jin, D. (2006a). *Reforming TEPS* (Research Repot No. 45). Seoul: SNU Language Education Institute.

Song, M.-J., Park, Y.-Y., Shin, S.-K., & Jin, D. (2007). *Research for the development of a new TEPS* (Research Report No. 48). Seoul: SNU Language Education Institute.

TEPS Council. (2012). *External feedback on the format and difficulty of the TEPS* (internal document). Seoul: TEPS Council, Seoul National University Foundation.

Yi, Y.-S. (2013b). On the optimal text length of reading comprehension tests. *The Jungang Journal of English Language and Literature*, *55*(4) 505-530.