TEPS 하위영역 점수 및 총점에 대한 신뢰도 분석

임의진[1†]
[1]서울대학교

Reliability of TEPS Section Scores and the Total Score

Euijin Lim[1†]
[1]Seoul National University

ABSTRACT

The purpose of the current study is to investigate the reliability and stability of the section and total (composite) scores of TEPS from the classical test theory perspectives. For reliability analyses, multiple sets of data were collected not only from the operational TEPS administrations before and after the revision but also from the four pilot tests administered during the TEPS revision process. Cronbach's (1951) alpha coefficients were computed for four different section scores of the tests while Feldt and Brennan's (1989) composite score reliability coefficients were computed for the total scores of these tests. These coefficients are examined and compared across different test forms of the original and revised TEPS. Coefficients of equivalence and stability and correlation coefficients between forms before and after the revision were also examined to see how stable TEPS scores were. The results showed that TEPS section scores and the total score were reliable and that the changes introduced by the revision did not reduce the stability. The total score had high reliability above 0.9 indicating that TEPS can be used as a dependable indicator of Korean English language learners' English proficiency to inform language-related, decision-making.

Keywords: TEPS, reliability, internal-consistency reliability, coefficient of equivalence and stability

1. 서 론

1999년에 TEPS가 처음 개발되고 시행된 이래로 영어 사용 상황의 변화와 영어교육 및 평가 분야에서의 연구 성과들이 축적되었고, 이에 따라 오랜 연구와 수험자의 요구를 바탕으로 TEPS에 대한 개정이 이루어졌다(Kwon et al., 2018). 개정에 따라 문항 수와 시험 시간 축소, 청해 및 독해 영역에 1지문 2문항 유형 도입, 청해 영역 대화 1회 청취, 어휘 영역과

† Corresponding author: ejlim.mail@gmail.com

https://doi.org/10.30961/lr.2019.55.S.51

문법 영역의 시험 시간 통합 등 TEPS에 다양한 변화가 있었다(좀 더 자세한 내용은 본 특별호의 Lee & Jun, 2019; Jun et al., 2019 참조).

이 중 TEPS의 문항 수와 시험 시간 축소는 모든 하위영역에 적용되었고 검사의 신뢰도와 직접적으로 연관되는 중대한 변화라고 할 수 있다. 개정 TEPS의 적절한 문항 수를 결정하기 위해 측정학적 지표와 내용적 측면이 동시에 고려되었다. 특히 Kim(2016)의 연구에서 검사 길이를 135문항으로 축소했을 때 예상 신뢰도를 산출하였고 기존 TEPS 대비 개정 TEPS의 총점 신뢰도가 높은 수준을 유지할 것으로 예측되어 개정 TEPS의 검사 길이를 결정하는 근거가 되었다.

TEPS의 경우 대단위로 시행되는 고부담 영어능력 시험으로, 검사의 신뢰도를 지속적으로 점검하고 그 수준을 유지하는 것이 매우 중요하다. 그러나 고부담 시험이라는 특징을 고려할 때 동일한 검사형이 수험자에게 한 번 이상 노출되어야 하는 검사-재검사 신뢰도나 한 회차에 하나 이상의 검사형을 사용해야 하는 동형검사 신뢰도를 산출하기에 현실적인 제약이 있다(TEPS Center, 2019). 내적 일관성 신뢰도는 일반적인 시행 상황에서 확보한 자료를 통해 매 회차 산출할 수 있고 검사의 질을 유지하는데 중요한 지표가 되지만 검사 내적인 오차요인만을 고려한다는 한계가 있다. 따라서 반복응시자를 대상으로 안정-동형도 계수를 산출하여 비교함으로써 내적 일관성 신뢰도가 포함하지 못하는 검사점수의 안정성 및 검사형 간 동형성 정도를 함께 고려할 수 있다.

본 연구에서는 네 차례에 걸쳐 시행된 파일럿 검사와 개정 전・후 시행된 TEPS의 하위영역 점수 및 총점에 대한 내적 일관성 신뢰도를 점검하였다. 또한 개정 전・후 시행된 TEPS 검사형에 대해 각각 안정-동형도 계수를 산출하여 시간에 따른 TEPS 점수의 안정성과 검사형 간 동형성 정도를 살펴보았다. 이와 더불어 개정 전과 후에 시행된 검사형 간 상관계수를 살펴봄으로써 개정으로 인해 점수의 안정성에 어떤 영향이 있었는지 점검하였다.

2. 이론적 배경과 관련 연구

2.1. 이론적 배경

전통적으로 신뢰도 지표는 응시자의 시험점수가 일관성을 지니는 정도를 나타낸다(Crocker & Algina, 1986; Ferguson & Takane, 1989). 고전검사이론에서는 관찰된 점수가 수험자의 능력을 나타내는 진점수와 오차점수의 합으로 이루어진다고 가정한다. 검사의 신뢰도는 이를 기반으로 얻어지는 측정학적 지표로, 진점수의 분산이 관찰점수 분산에서 차지하는 비율로 정의할 수 있다. 대단위 검사에 있어서 검사의 신뢰도는 타당도의 필요조건이자 검사의 질을 점검하는 데 있어서 핵심적인 지표라고 할 수 있다. 검사 점수가 오차 없이 일관성 있게 진점수를 나타내는 정도, 즉 신뢰도가 일정 수준 이상 도달해야 비로소 그 진점수가 과연 측정하고자 하는 바를 제대로 나타내고 있는지 타당도의 관점에서 논할 수 있기 때문이다.

신뢰도 지표를 산출하기 위하여 다양한 절차가 개발되어 왔다. 검사-재검사 신뢰도(test-retest reliability)는 동일한 검사형을 짧은 시간차를 두고 동일한 응시자에게 두 번 시행하였을 때 응시자의 점수가 일관되는지를 살펴 보는 방법이다. 보통 동일 응시자의 검사, 재검사 점수 간 적률상관계수를 구하여 신뢰도 지표로 사용한다. 동형검사 신뢰도(parallel-forms reliability)는 같은 평가틀(test specification)로부터 제작되어 평가의 내용 범위나 측정학적 속성이 동일한 서로 다른 두 검사형을 동일 응시자에게 시행하여 응시자의 점수가 일관된 정도를 검사하는 방법이다. 이 방법 역시 동일한 응시자들이 두 검사형에서 얻은 점수 간 적률상관계수를 구하여 신뢰도 지표로 사용한다. 내적 일관성 신뢰도(internal-consistency reliability)는 한 검사형의 모든 문항들에 대해 응시자가 얼마나 일관되게 반응하는지를 살펴보는 방식이다.

이러한 측정 방식에 따라 각 지표가 의미하는 바는 조금씩 달라진다. 검사-재검사 신뢰도의 경우 시간에 따른 검사 점수의 안정성을 의미하여 안정성 계수(coefficient of stability)라고도 불리며, 성격검사나 태도검사에 중요한 지표로 보고된다. 반면 동형검사 신뢰도의 경우 검사형 간 동형성 정도를 나타내는 동형성 계수(coefficient of equivalence)로도 파악될 수 있는데, 문항 노출에 민감하여 동형검사를 제작하는 인지능력 측정 검사의 경우에 특히 중요한 지표라고 할 수 있다. 경우에 따라 안정성과 동형성을 동시에 파악하고자 할 때에는 동일한 수험자가 시간 간격을 두고 서로 다른 동형검사에 응시하도록 할 수 있다. 이때 두 동형검사에서 얻은 점수 간 적률상관계수를 신뢰도 지표로 사용하며 이는 안정-동형도 계수(coefficient of equivalence and stability)라고 불린다. 안정-동형도 계수는 여러 오차요인을 복합적으로 포함하므로 다른 신뢰도 지표보다 낮게 산출될 수밖에 없으나 검사 점수의 안정성과 검사형 간 동형성을 동시에 고려하는 지표라고 할 수 있다.

한 수험자 집단에서 얻은 두 검사점수 간 적률상관계수를 통해 산출되는 검사-재검사 신뢰도나 동형검사 신뢰도와 달리 내적 일관성 신뢰도는 한 수험자 집단이 얻은 한 검사점수로부터 신뢰도 계수를 산출하기 위해 보다 복잡한 통계적 절차를 적용한다. 내적 일관성 신뢰도의 가장 단순한 형태는 검사를 둘로 나누고 원래 검사 길이의 절반인 각 이분검사 점수 간 상관계수를 구하는 것이다. 이 상관계수는 절반 길이 검사에 대한 신뢰도가 되므로, Spearman-Brown 공식(Brown, 1910; Spearman, 1910)을 적용하여 원래 길이 검사에 대한 신뢰도를 산출한다. 이러한 반분검사 신뢰도(split-half reliability)는 검사를 어떻게 반분하는지에 따라 신뢰도가 달라지므로 이를 일반화한 Cronbach's alpha(Cronbach, 1951), 즉 알파계수가 보다 널리 사용된다. 알파계수는 검사 점수의 내적 일관성을 측정하는 대표적인 통계치로, 다음과 같이 나타낼 수 있다.

$$\rho_{\alpha} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_{X_f}^2}{\sigma_X^2}\right)$$

이때 σ_X^2는 원점수 총점의 분산이고, $\sigma_{X_f}^2$는 각 문항의 분산이다. n은 검사에 속한 문항의

수를 가리킨다. 이는 검사 점수 전체의 분산 중 각 문항들이 일관성 있게 측정하는 공분산의 비율을 나타낸다고 할 수 있다.

일반적으로 알파계수는 문항의 수에 비례하여 문항의 수가 많을수록 그 값이 커진다. 질 좋은 문항이 많아질수록 측정하고자 하는 내재적 능력에 대한 정보가 많아지므로, 진점수를 추정하는 데 있어 오차가 작아지고 일관성이 높아지기 때문이다. 그러나 문항의 수가 일정 수준을 넘어서면 신뢰도 상승 폭이 둔화되므로 시험 시간, 수험자의 부담, 문항 제작 비용 등을 고려하여 문항의 수를 적절한 수준으로 결정할 필요가 있다.

만약 검사 점수가 여러 하위 점수의 가중합산인 선형 조합(linear composite)으로 이루어져 있다면 알파계수로 그 신뢰도를 산출하는 것은 적절하지 않다. Feldt와 Brennan(1989)은 여러 하위영역으로 구성된 검사의 총점, 검사 총집(test battery)에서 제공되는 점수 등에 대해 신뢰도를 구하는 방법을 제안하였다. 총 n개의 하위영역으로 이루어진 어떤 검사의 총점은 각 하위영역 h에 대한 관찰점수 X_h와 가중치 w_h의 선형 조합으로 다음과 같이 정의된다.

$$Z_p = w_1X_{p1} + w_2X_{p2} + \cdots + w_hX_{ph} + \cdots + w_nX_{pn}$$

이때 각 영역의 관찰점수는 각 영역의 진점수(T)와 오차점수(E)의 합으로 이루어지고, 진점수와 오차점수 간 상관이 없으므로 각 관찰점수 분산 및 공분산은 아래와 같다.

$$\sigma_{wX}^2 = w^2\sigma_X^2 = w^2\sigma_T^2 + w^2\sigma_E^2$$
$$\sigma_{(w_hX_h)(w_jX_j)} = w_hw_j\sigma_{X_h}\sigma_{X_j}$$

총점의 분산은 총점을 구성하는 각 영역별 점수의 분산 공분산 행렬을 따라 구할 수 있다. 또한 총점의 오차 분산은 오차점수 간 상관이 없으므로 영역별 점수 오차분산의 가중합산으로 구할 수 있다.

$$\sigma_Z^2 = \sum w_h^2\sigma_{X_h}^2 + \sum\sum_{h\neq j} w_hw_j\sigma_{X_h}\sigma_{X_j}$$
$$\sigma_{E_Z}^2 = \sum w_h^2\sigma_{E_h}^2$$

따라서 총점의 신뢰도, 즉 총점의 전체 분산에서 오차 분산을 제외한 진점수 분산의 비율은 다음과 같다.

$$\rho_{ZZ'} = 1 - \frac{\sigma_{E_Z}^2}{\sigma_Z^2}$$

신뢰도 계수를 산출하는 방법은 이 외에도 여러 가지가 있다. 문항반응이론, 일반화가능도 이론 등을 바탕으로 산출되는 여러 신뢰도 계수들은 그 방법에 따라 세부적인 의미가 달라질

수 있으나 공통적으로 0과 1 사이의 값을 가지며 값이 1에 가까울수록 해당 측정도구를 사용해 얻은 관찰점수가 진점수(혹은 일반화가능도이론의 전집점수)에 가깝다는 것을 의미한다.

2.2. 언어평가 분야의 관련 연구

1980년대에 신뢰도의 개념이 언어평가 분야에 소개된 이래로 언어평가 분야의 신뢰도 연구는 검사의 특성, 검사 상황, 점수 산출 방식 등을 고려하여 보다 정교한 신뢰도 지수를 산출하는 방향으로 발전하였다. Krzanowski와 Woods(1984)는 반분검사 신뢰도, 알파계수, KR-20 등 다양한 신뢰도 지수와 그 통계적 특성을 소개하며 언어평가 분야에서 실제로 신뢰도 개념을 적용하고 해석하는데 주의할 점을 정리하였고, Bolus, Hinofotis, 그리고 Bailey(1982)는 채점자가 개입되는 언어평가 상황을 고려하여 일반화가능도를 바탕으로 신뢰도를 산출할 것을 제안하였다. Weigle(1998), Knoch(2009) 등은 채점자의 개입에 주목하여 채점자 훈련을 통해 채점자간 신뢰도 및 채점자내 신뢰도가 향상된다는 것을 보였고, Attali, Lewis, 그리고 Steier(2013)는 에세이 문항에 컴퓨터 자동 채점을 적용하였을 때 일반화가능도 이론을 바탕으로 한 신뢰도를 보고하였다. Gessaroli와 Folske(2002)는 검사가 단위검사(testlet)로 이루어진 경우 알파계수로 신뢰도를 산출하는 것이 적절하지 않다는 것을 보이고 대안에 대해 논의하였고, Longabach와 Peyton(2018)은 총점에 더불어 하위점수를 제공하는 경우 하위점수 산출 방식에 따라 신뢰도가 어떻게 달라지는지 비교하였다.

대단위 검사의 경우 검사의 신뢰도를 점검하고 높은 수준으로 유지하는 것이 중요한 책무이므로 대부분의 주요 영어검사들은 총점과 하위영역별 점수에 대한 신뢰도 산출 방법과 결과를 공개하고 있다. TOEFL의 읽기와 듣기 신뢰도는 문항반응이론을 바탕으로 산출되었고, 약 0.87 수준으로 나타났다. 말하기와 쓰기 영역의 신뢰도는 알파계수를 통해 산출하였고, 각각 0.86, 0.80으로 나타났다. 총점의 신뢰도는 0.95로 나타났으나 어떤 방법으로 산출하였는지는 밝히지 않았다(Educational Testing Service, 2011). TOEIC은 알파계수 대신 내적 일관성 신뢰도의 지표 중 하나인 KR-20를 사용하였고, 모든 검사형에 대해 TOEIC 듣기와 읽기 영역 점수에 대한 신뢰도를 구했을 때 그 평균이 약 0.90이라고 보고하였다(Educational Testing Service, 2019). IELTS의 읽기와 듣기 영역에 대한 알파계수는 약 0.88수준이고(UCLES, 2007), Feldt와 Brennan(1989) 공식에 따른 총점 신뢰도는 약 0.95로 나타났다(Fazel & Ahmadi, 2011). Cambridge English Exams의 경우 2010년에 시행된 각 수준별 검사에 대해 영역에 따라 알파계수, 채점자간 상관계수, 일반화가능도이론에 기반을 둔 일반화가능도 계수 등을 사용하였고, 총점에 대해서는 복합신뢰도(composite reliability)를 보고하였다. 수준, 영역에 따른 차이는 있었으나 영역별 점수에서는 0.70-0.91 수준의 신뢰도를 보였고, 총점에서는 0.92-0.95수준으로 나타났다(UCLES, n.d.).

3. 연구 방법

3.1. 자료

본 연구에서는 총 12개의 검사형에 대한 응답자료를 사용하였다. 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 다양한 수험자 집단에서 시행되었고, 선행된 파일럿 검사의 분석 결과에 따라 이후의 파일럿 검사형에 개선점을 반영하였다. 본 연구에서는 파일럿 검사를 포함한 12개의 검사형 전체에 대해 총점과 하위영역별 점수에 대한 내적 일관성 신뢰도를 산출하였다. 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

<표 1>에 나타난 원점수 평균을 비교했을 때 문항 수가 축소됨에 따라 원점수 평균이 개정 전 113.13-128.93에서 개정 후 72.59-76.94로 바뀌었다. 네 차례에 걸친 파일럿 검사에서는 검사형 간 원점수 평균 변화가 약 20점 수준으로 크게 나타났으나 개정이 완료된 후에는 원점수 평균 변화가 크지 않았다. 문항 수 축소로 점수의 범위가 변화하면서 개정 후 표준편차 또한 개정 전에 비해 감소하였다. 왜도와 첨도는 모든 검사형에서 유사하게 나타나 점수 분포 형태가 크게 변화하지 않았음을 파악할 수 있다.

파일럿 검사와 개정 전・후 검사형에 대한 하위영역별 검사 길이와 전체 검사 길이는 다음의 <표 2>와 같다. 파일럿 시험으로 사용된 검사형과 개정 후 시행된 검사형의 검사 길이는 개정 전 검사형에 비해 청해, 어휘, 문법에서 각각 20문항, 독해에서 5문항 줄어들었다. 그에 따라 전체 검사 길이는 200문항에서 135문항으로 줄어들었다. 파일럿 검사형과 개정 후 검사형의 검사 길이는 서로 동일하다.

표 2. 검사형에 따른 문항 수

문항 수	개정 전(A-D)	파일럿	개정 후(E-H)
청해	60	40	40
어휘	50	30	30
문법	50	30	30
독해	40	35	35
전체	200	135	135

안정-동형도 계수는 파일럿 검사가 아닌 일반적인 검사 상황에서 시행된 A-H형을 치른 반복응시자를 대상으로 산출되었다. 안정-동형도 계수의 산출 대상이 되는 A-H 검사형 간 반복응시자 수는 다음의 <표 3>과 같다.

표 3. 검사형 A-H에 반복 응시한 수험자 수

검사형	A	B	C	D	E	F	G	H
A	약 6,900명							
B	2,388	약 5,600명						
C	2,223	2,324	약 6,000명					
D	2,156	2,317	2,765	약 6,700명				
E	1,550	1,556	1,762	2,177	약 6,200명			
F	1,143	1,153	1,283	1,621	2,055	약 5,300명		
G	882	924	1,011	1,245	1,578	1,748	약 4,500명	
H	985	895	989	1,160	1,382	1,419	1,709	약 5,800명

<표 3>에 나타난 바와 같이 시행 시점이 인접한 회차 간에는 2,000명 안팎의 많은 응시자가 두 회차를 모두 응시하였고, 시행 시점이 가장 떨어져 있는 A형과 H형 사이에는 985명의 반복응시자가 있었다.

3.2. 분석 방법

본 연구에서는 개정 이전에 시행된 네 개의 검사형, 개정 과정에 시행된 네 차례의 파일럿 검사, 개정 이후에 시행된 네 개의 검사형 모두에 대해 각각 내적 일관성 신뢰도를 산출하여 비교하였다. 또한 개정 전・후 시행된 동형검사 간 안정-동형도 계수를 산출하여 검사 점수의 안정성과 검사형 간의 동형성을 점검하였다. 마지막으로 개정 전 검사형과 개정 후 검사형 간의 상관계수를 검토하여 개정이 점수의 안정성에 미친 영향을 파악하였다.

먼저 TEPS의 각 하위영역 점수에 대한 내적 일관성 신뢰도는 알파계수를 통해 산출하였다. TEPS의 총점은 각 하위영역 점수의 가중합산으로 산출되므로 Feldt와 Brennan(1989)이 제시한 방법을 적용하였다. 각 영역의 가중치는 척도점수 대비 각 문항별 점수의 비율로 정의하였다. 따라서 개정 전 검사형의 문항당 척도점수는 청해, 어휘, 문법, 독해 영역에서 각각 396/60, 99/50, 99/50, 396/40점이고, 파일럿 검사 및 개정 후 검사형에서는 각각 240/40, 60/30, 60/30, 240/35점이 된다. 이때 영역별 가중치는 문항별 점수의 비율로 산출 된다.

안정-동형도 계수는 <표 2>에 나타난 반복응시자를 대상으로 각 검사형을 통해 측정한 원점수 간 적률상관계수로 측정되었다. 구체적으로 하위영역 점수와 총점 각각에 대해 개정 전(A-D형) 검사형 간 안정-동형도 계수와 개정 후(E-H형) 검사형 간 안정-동형도 계수를 산출하였다. 또한 개정 전・후 검사형 간 적률상관계수를 산출하여 개정으로 인한 영향을 파악하고자 하였다.

4. 결과

<표 4>는 검사형에 따른 하위영역별 점수와 총점의 내적 일관성 신뢰도를 나타낸다. [그림 1]은 <표 4>에 나타난 신뢰도 계수를 그래프로 표시하였다.

표 4. 검사형에 따른 하위영역별 점수 및 총점의 내적 일관성 신뢰도

검사형	청해*	어휘*	문법*	독해*	전체**
A	0.91	0.86	0.88	0.88	0.96
B	0.91	0.85	0.88	0.91	0.96
C	0.90	0.84	0.82	0.84	0.95
D	0.92	0.83	0.88	0.89	0.96
1차	0.87	0.81	0.78	0.85	0.94
2차	0.85	0.82	0.83	0.87	0.94
3차	0.88	0.76	0.79	0.87	0.95
4차	0.88	0.78	0.85	0.85	0.94
E	0.89	0.81	0.83	0.86	0.95
F	0.88	0.77	0.81	0.83	0.94
G	0.87	0.79	0.81	0.84	0.94
H	0.87	0.78	0.75	0.86	0.94

*: 알파계수
**: 선형 조합으로 이루어진 점수에 대한 신뢰도

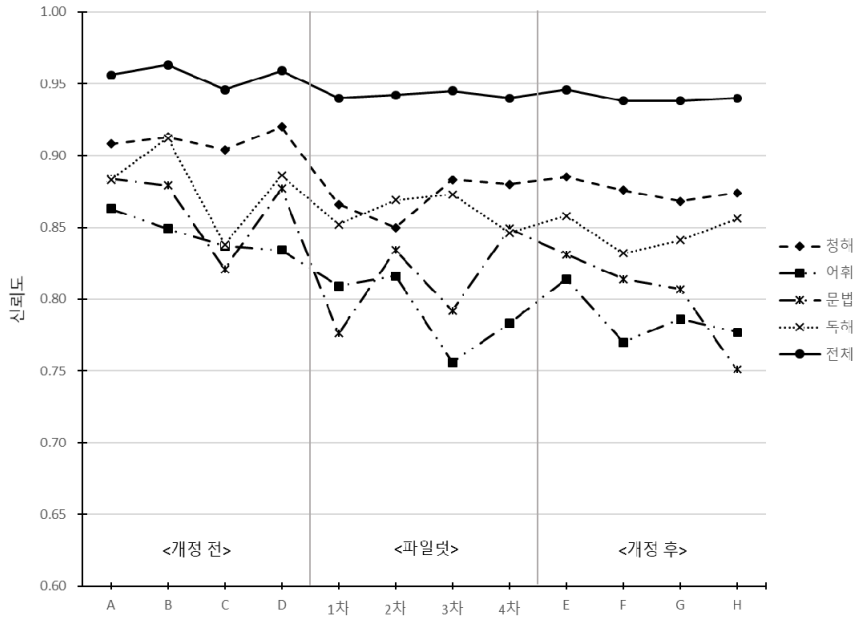

그림 1. 검사형에 따른 하위영역별 점수 및 총점의 내적 일관성 신뢰도

<표 4>에 나타난 알파계수를 하위영역별로 비교하면, 청해 영역의 신뢰도가 0.85-0.92 수준으로 가장 높았고 독해가 0.83-0.91 수준으로 뒤를 이었다. 어휘와 문법은 각각 0.76-0.86, 0.75-0.88 수준으로 비슷한 값을 보였다. 총점 신뢰도는 0.94-0.96 수준으로 매우 높게 나타났다. 개정 전후를 비교했을 때 전반적으로 개정 전 검사형인 A-D에 비해 파일럿 1-4차와 E-H 검사형에서 신뢰도가 소폭 하락한 것으로 나타났다. 특히 어휘와 문법 영역은 알파계수가 0.8 이하로 산출된 검사형이 일부 나타났다. 이는 해당 영역의 문항 수가 50에서 30으로 축소되며 예상된 변화라고 할 수 있다. 개정 후 청해 영역의 알파계수는 개정 전에 비해 소폭 하락하였으나 0.8 이상을 유지하였고, 독해 영역은 개정 전후 알파계수가 0.8 이상의 높은 수준으로 비슷하게 나타났다. 총점 신뢰도는 개정 전후 미세한 차이는 있었으나 0.9 이상의 매우 높은 수준을 그대로 유지하여 대단위 고부담 검사에서 요구되는 높은 신뢰도 수준을 충족하였다.

<표 5>는 A-H 검사형에서 나타난 청해 점수의 안정-동형도 계수 및 상관계수를 보여준다. 개정 전 A-D 검사형에서 청해의 안정-동형도 계수는 0.87-0.89 수준으로 0.8 이상의 높은 값을 보였다. 개정 후 시행된 E-H 검사형에서 관찰된 안정-동형도 계수는 0.83-0.86 수준으로 개정 전 검사형에서 나타난 값보다는 소폭 하락하였으나 여전히 높은 수준으로 나타났다. 이는 <표 4>에 나타난 청해 점수의 알파계수보다는 낮은 값이나 시간에 따른 수험자 능력의 변화와 검사형 간의 동형성 정도로 인한 오차를 반영한다는 점을 고려했을 때 충분히 높은 값으로, 청해 점수는 시간에 따라 급변하지 않는 안정성이 있으며 검사형 간 동형성이 높다고 할 수 있다. 이탤릭체로 표시된 개정 전과 후 검사형 간 상관계수도 0.83-0.87 수준으로 높게 나타나 TEPS 개정으로 인한 변화가 청해 점수의 안정성을 해치지 않음을 보였다.

표 5. 청해 점수의 안정-동형도 계수 및 상관계수

검사형	A	B	C	D	E	F	G	H
A	1.00							
B	0.88	1.00						
C	0.88	0.87	1.00					
D	0.88	0.89	0.89	1.00				
E	*0.85*	*0.85*	*0.86*	*0.87*	1.00			
F	*0.86*	*0.85*	*0.85*	*0.86*	0.84	1.00		
G	*0.86*	*0.84*	*0.83*	*0.85*	0.83	0.83	1.00	
H	*0.85*	*0.84*	*0.84*	*0.85*	0.86	0.83	0.83	1.00

<표 6>은 A-H 검사형에서 나타난 어휘 점수의 안정-동형도 계수 및 상관계수를 보여준다. 개정 전과 후 어휘의 안정-동형도 계수의 범위는 각각 0.80-0.83, 0.76-0.79로, 개정 과정에서의 문항 수 축소에 따라 신뢰도가 소폭 하락한 것으로 나타났다. 그러나 <표 4>에 나타난 어휘 점수의 알파계수와 비교했을 때 안정-동형도 계수는 시간에 따른 수험자 능력의 변화와 검사형 간 동형성 정도로 인한 오차를 반영했음에도 불구하고 그 값이 큰 폭으로 하락하지 않아, 어휘 점수가 시간에 따라 급변하지 않는 안정성이 있으며 검사형 간 동형성이 높다는 것을 보였다. 개정 전・후 검사형 간 상관계수도 0.74-0.80 수준으로 높게 나타나 TEPS 개정으로 인한 변화가 어휘 점수의 안정성을 해치지 않음을 보였다.

표 6. 어휘 점수의 안정-동형도 계수 및 상관계수

검사형	A	B	C	D	E	F	G	H
A	1.00							
B	0.83	1.00						
C	0.81	0.82	1.00					
D	0.81	0.80	0.81	1.00				
E	*0.80*	*0.78*	*0.79*	*0.79*	1.00			
F	*0.79*	*0.77*	*0.75*	*0.76*	0.79	1.00		
G	*0.78*	*0.76*	*0.78*	*0.78*	0.78	0.77	1.00	
H	*0.76*	*0.74*	*0.74*	*0.74*	0.77	0.76	0.78	1.00

<표 7>은 A-H 검사형에서 나타난 문법 점수의 안정-동형도 계수 및 상관계수를 보여준다. 개정 전 A-D 검사형에서 문법의 안정-동형도 계수는 0.81-0.86 수준으로 0.8 이상의 높은 값을 보였다. 개정 후 시행된 E-H 검사형에서 관찰된 안정-동형도 계수는 0.75-0.82 수준으로, 문항 수 축소에 따라 개정 전 검사형에서 나타난 값보다 소폭 하락하였다. 그러나 <표 4>에 나타난 문법 점수의 알파계수와 비교했을 때 안정-동형도 계수가 큰 차이를 보이지 않으므로 문법 점수가 시간 변화에도 안정적이고 검사형 간 동형성이 높다고 할 수 있다. 개정

전과 후 검사형 간 상관계수도 0.73-0.83 수준으로 높게 나타나 TEPS 개정으로 인한 변화에도 불구하고 문법 점수가 안정적으로 나타난다는 것을 보였다.

표 7. 문법 점수의 안정-동형도 계수 및 상관계수

검사형	A	B	C	D	E	F	G	H
A	1.00							
B	0.86	1.00						
C	0.83	0.81	1.00					
D	0.86	0.85	0.83	1.00				
E	*0.81*	*0.81*	*0.77*	*0.83*	1.00			
F	*0.78*	*0.77*	*0.75*	*0.80*	0.82	1.00		
G	*0.81*	*0.81*	*0.76*	*0.81*	0.78	0.77	1.00	
H	*0.78*	*0.77*	*0.73*	*0.77*	0.76	0.75	0.76	1.00

<표 8>은 A-H 검사형에서 나타난 독해 점수의 안정-동형도 계수 및 상관계수를 보여준다. 개정 전과 후 독해의 안정-동형도 계수는 각각 0.81-0.84, 0.80-0.83 수준으로 0.8 이상의 높은 값을 보였다. <표 4>에 나타난 독해 점수의 알파계수와 비교했을 때 독해 점수의 안정-동형도 계수는 다른 오차 요인들을 반영했음에도 불구하고 그 값의 차이가 작아, 독해 점수가 시간에 따라 급변하지 않는 안정성이 있으며 검사형 간 동형성이 높음을 나타냈다. 개정 전·후 검사형 간 상관계수도 0.78-0.83 수준으로 높게 나타나 TEPS 개정으로 인한 변화에도 불구하고 독해 점수가 안정적임을 보였다.

표 8. 독해 점수의 안정-동형도 계수 및 상관계수

검사형	A	B	C	D	E	F	G	H
A	1.00							
B	0.84	1.00						
C	0.81	0.82	1.00					
D	0.84	0.84	0.83	1.00				
E	*0.80*	*0.81*	*0.79*	*0.83*	1.00			
F	*0.80*	*0.79*	*0.78*	*0.81*	0.81	1.00		
G	*0.82*	*0.83*	*0.78*	*0.82*	0.80	0.81	1.00	
H	*0.83*	*0.82*	*0.78*	*0.81*	0.83	0.82	0.81	1.00

<표 9>는 A-H 검사형에서 나타난 총점의 안정-동형도 계수 및 상관계수를 보여준다. 개정 전 총점의 안정-동형도 계수의 범위는 각각 0.94-0.95로 매우 높은 수준을 보였다. 개정 후 문항 수가 200문항에서 135문항으로 줄어들었음에도 불구하고 안정-동형도 계수는 0.92-0.93 수준으로 높게 유지되었다. 이는 <표 4>에 나타난 총점의 알파계수에 비해 큰 차

이가 없어, TEPS 총점이 시간에 따라 급변하지 않으며 검사형 간 동형성이 높음을 나타낸다. 개정 전・후 검사형 간 상관계수도 0.91-0.94로 매우 높게 나타나 TEPS 개정으로 인한 변화가 총점의 안정성을 해치지 않음을 보였다.

표 9. 총점의 안정-동형도 계수 및 상관계수

검사형	A	B	C	D	E	F	G	H
A	1.00							
B	0.95	1.00						
C	0.94	0.94	1.00					
D	0.94	0.94	0.95	1.00				
E	*0.93*	*0.93*	*0.93*	*0.94*	1.00			
F	*0.92*	*0.92*	*0.92*	*0.93*	0.93	1.00		
G	*0.92*	*0.92*	*0.92*	*0.93*	0.92	0.93	1.00	
H	*0.92*	*0.91*	*0.91*	*0.92*	0.93	0.92	0.93	1.00

5. 결론 및 제언

본 연구에서는 대단위로 시행되는 고부담 영어능력 시험인 TEPS 의 영역 및 총점이 개정 후에도 충분한 수준의 안정성과 신뢰도를 유지하고 있는지 살펴보고자 하였다. 이를 위해 각 신뢰도 지표의 의미와 시행 환경을 고려하여 총점 및 각 하위영역 점수에 대해 알파계수로 측정한 내적 일관성 신뢰도와 반복응시자를 통해 산출한 안정-동형도 계수를 점검하였다. 또한 개정 전과 후의 검사형 간 상관계수를 산출하여 개정이 점수의 안정성에 미친 영향을 확인하였다.

TEPS와 같은 대단위 시험에 있어서 신뢰롭고 안정적인 점수를 산출하는 것은 시행 기관의 책무라고 할 수 있다. 본 연구에서는 개정 전과 후 TEPS가 지속적으로 신뢰도 높은 점수를 산출했으며, 개정으로 인한 변화가 TEPS 점수의 안정성을 해치지 않았다는 근거를 제시하였다. 특히 주로 활용되는 TEPS 총점의 신뢰도가 0.9 이상으로 매우 높게 유지되어 수험자의 영어 능력에 대한 중요한 판단 근거로 활용되기에 충분히 신뢰롭다는 것을 보였다. 또한 TOEFL, TOEIC, IELTS, Cambridge English Exams 등 국제적으로 널리 사용되는 영어능력검사와 비교했을 때 TEPS의 총점 신뢰도는 유사한 수준인 것으로 나타났다.

다만 본 연구에서는 활용 가능한 자료 구조의 한계로 내적 일관성 신뢰도와 안정-동형도 계수만을 산출하였다. 안정-동형도 계수는 시간에 따른 검사 점수의 안정성과 검사형 간 동형성을 동시에 고려하는 신뢰도 지표이나, 그 오차 요인 각각의 영향력을 분리하여 파악할 수는 없다. 추후 일반화가능도이론을 바탕으로 면밀한 설계에 따라 자료를 수집하여 각 오차 요인의 상대적 영향력을 파악할 수 있을 것이라 기대된다. 또한 장기적으로 문항반응이론에 기반한 신뢰도를 산출하고 해당 검사형을 구성하는 문항들을 개별적으로 분석하여 높은 신뢰도를 유지하기 위한 출제 지침을 개선하는 후속 연구가 가능할 것으로 생각된다.

References

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*, 125–141.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*, 245-258.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Educational Testing Service. (2011). *TOEFL® Research Insight Series, Volume 3: Reliability and comparability of TOEFL iBT scores*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Educational Testing Service. (2019). *User guide for the TOEIC® listening and reading test*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic-listening-reading-test-user-guide.pdf

Fazel, I., & Ahmadi, A. (2011). On the relationship between writing proficiency and instrumental/integrative motivation among Iranian IELTS candidates. *Theory and Practice in Language Studies, 1*, 747-757.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: MacMillan.

Ferguson, G. A. & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). New York, NY: McGraw-Hill.

Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2*, 277-295.

Kim, J. (2016). *Reliability and test length* (internal document). Seoul: TEPS Center.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275–304.

Krzanowski, W. J., & Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing, 1*, 1–20.

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.

Longabach, T., & Peyton, V. (2018). A comparison of reliability and precision of subscore reporting methods for a state English language proficiency assessment. *Language Testing, 35*, 297–317.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

TEPS Center (2019). *TEPS technical report: 2018 administration* (internal document). Seoul: TEPS Center.

UCLES. (2007). *IELTS handbook 2007*. Retrieved from http://www.ielts.org/pdf/IELTS_Handbook_2007.pdf

UCLES. (n.d.). *Quality and accountability*. Retrieved from https://www.cambridgeenglish.org/research-and-validation/quality-and-accountability/

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.