

ITSC G-TELP

A Validity Study of the IBT Speaking Certification Test

2022

Prepared by **Minjung Kim, PhD**

Conference **Korean Association for Educational Information and Media**

Date **Oct 22, 2022**

발표 순서

01

Introduction

말하기 시험

02

Literature Review

iBT 말하기 시험

03

Methodology

타당도 논의

04

Results

분석 결과

05

Conclusion

결론 및 제언

KEYWORDS

G-TELP 연구 키워드

타당성

iBT Speaking Test

Rasch Model

신뢰도

재택시험

말하기 시험이란?

Q. 통상적으로 언어 평가 영역에서 측정하고자 하는 것은 무엇일까?

언어 학습자가 가지고 있는 언어적인 능력에 대한 평가이다. 여기에서 '**언어적인 능력**'이라는 것은 해당 언어를 통해 타인과 의사소통을 수행하는 능력.

언어평가를 위한 영어 말하기 시험

증빙 자료로 활용

직장, 학업에서 서류 및 면접을 통과하기 위한 증빙 자료로 이용됨

실무 언어 능력

학문적 맥락이나 직장에서의 언어 능력 — 기업 및 실무에서는 실무 영어 능력 우수자를 선호함

연구 목적

1

타당성 검증

타당도는 뒷받침하는 증거의 적절성에 대한 평가적 판단으로, 시험 점수 해석 및 사용 논쟁의 첫걸음으로서, 영어 말하기 시험의 at home 테스트 점수 해석 및 활용의 타당성을 검증함.

2

시험 구조 파악

국내 영어 말하기 시험과 평가 시스템을 살펴보고 말하기 시험의 구조를 파악함.

3

객관성 검토

영어 말하기 시험의 의의: 평가 영역에서 객관성은 매우 중요한 항목이며, 실제 평가 기준을 활용하여 영어 말하기 능력을 객관적으로 평가하는지 살펴봄.

말하기 시험 사례 (1) — TOEIC Speaking & OPIc

TOEIC Speaking

- 온라인 플랫폼에서 학생들이 끊임없이 참여하고 도전하며, 지속적으로 언어 능력을 향상
- 채점 과정: 응시자 답변 → ETS 인증 Rater → Calibration Test 응시·통과 → 수험자 답변 채점 → 성적 산출·발표



OPIc

- 미국 ACTFL(American Council on the Teaching of Foreign Languages) 기관에서 출제
- 시험 직후 성적 확인, 25일 이후 재응시 가능, 응시 시간 40분
- 돌발 토픽·롤플레이처럼 예상하지 못한 질문도 정확히 이해하고 답변, 답변 내용과 문장에서 주제 전달이 명확한지를 평가

말하기 시험 사례 (2) — GST (GTELP Speaking Test)

- 미국 ITSC(International Testing Services Center)에서 개발된 국제 공인 영어 말하기 시험
- 시험 시간 약 30분, 11개 parts·30여 개 질문, 평가 기준(Content, Grammar, Vocabulary, Pronunciation)
- 교차 채점으로 수험자의 말하는 능력을 분석·진단함

채점 프로세스

수험자
녹음테이프

제1·제2 Rater
채점 (교차채점)

점수 차이 시
규정 절차에 따른 재평가

무작위 추출 검사 및
결과 보고

채점자 교육 매년 주관 기관의 채점 척도 적용 표준화 훈련 등, G-TELP Rater 양성 및 교육 프로그램을 통해 이루어짐

말하기 시험의 타당성

- 영어 말하기 평가의 인증 여부는 도구의 신뢰도와 평가의 타당성이 우선되어야 함
- 평가 도구마다 평가 기준, 평가 영역, 평가 등급 및 방법, 평가 시간이 모두 다르게 구성되어 있으며, 그중 지텔프 말하기 시험(GST)의 타당도를 살펴봄

GST 평가 기준 5가지 표현력 · 문법 · 어휘 · 발음 · 유창성 — 평가 도구마다 서로 상이하게 구성되어 있음

- 말하기 평가 기준이 다르므로, 동일 대상의 수험자를 어느 평가 도구로 측정하느냐에 따라 신뢰도 문제로 연결됨
- 말하기 평가는 의사소통 능력을 평가하는 것으로, 질문을 잘 이해해서 정확한 어법을 사용해 적절한 어휘를 정확히 사용하고 상대방이 알아들을 수 있는 발음으로 자연스럽게 말하는가를 평가하는 것임

공인 인증 말하기 시험의 타당성

평가 방식

- GST는 2인 1조의 평가자에 의해 평가가 이루어지며, 평가자 결과에 대한 평균 점수를 결과로 산출함

재택 시험 도입

- 말하기 평가는 시간, 날씨, 평가자의 컨디션에 따라 주관이 개입될 수 있어, 본인이 원하는 시간에 집에서 시험을 칠 수 있는 환경을 구성하여 신뢰성을 확보함

Rasch

Rasch 측정모형 도입

Rasch 측정모형을 이용하여 공인 말하기 시험(GST)에서 **신뢰도와 타당도**를 검증함

연구 절차 — Rasch 측정모형

- 타당도는 뒷받침하는 증거의 적절성에 대한 평가적 판단으로, 말하기 영어 인증 시험(GST at home)의 점수 해석 및 활용의 타당성을 검증하고자 함
- Rasch 측정모형은 원자료를 이용해서 문항 난이도 또는 피험자의 능력 추정값을 산출할 때, 피험자 집단의 수준 또는 검사 난이도에 따라 변하는 문제점을 해결해 줌

Rasch 모형 공식 (Rasch, 1960)

$$(P_{nik} / (1 - P_{nik})) = B_n - D_i$$

B_n 피험자의 능력 **D_i** 문항 난이도

3. METHODOLOGY

적합도 지수 (Infit / Outfit)

Rasch 모형에서는 개별 문항이 모형에 적합한지를 검증하기 위해 외적합도와 내적합도 지수 값을 제공함

외적합도 지수 (Outfit)

$$MS(WT)_i = \Sigma(X_{ni} - P_{ni})^2 / \Sigma P_{ni}(1 - P_{ni})$$

- 피험자의 능력 수준에 비추어 너무 쉬운 문항을 틀리거나 너무 어려운 문항을 맞히는 등, 이상 문항 반응 형태에 민감한 지수
- 능력 수준에서 많이 벗어난 문항에 대한 민감도를 나타냄

내적합도 지수 (Infit)

$$MS(UT)_i = (1/N) \Sigma(X_{ni} - P_{ni})^2 / P_{ni}(1 - P_{ni})$$

- 능력 수준에서 크게 벗어나지 않은 문항의 이상 반응 형태에 가중치를 둔 적합도 지수

기준: 외적합도·내적합도 지수 값은 1.0을 기준으로 1.5 이상이면 문제가 있는 문항으로 간주함 (Linacre, 2009)

연구 방법

연구 대상

GST 채점 기준표, 영어 말하기, GST 수험자의 파트별 시험 점수(2022년 시행)로 분석을 수행함

분석 도구

- 시험이 어떻게 개발·운영되어왔는지를 파일럿 테스트 및 전 과정 모니터 기간 데이터를 수집·분석
- 수험자의 파트별 시험 결과 및 demographic 분석, 시험 점수의 타당도 검증 연구를 수행
- 기술통계 및 시험의 신뢰도·타당도는 Rasch Model을 수행
- R 프로그램을 GUI 방식으로 구현한 오픈소스 통계 프로그램 jamovi(2021), snowRMM(Seol, 2021; Willse, 2014) 통계 모듈을 이용하여 분석을 수행함

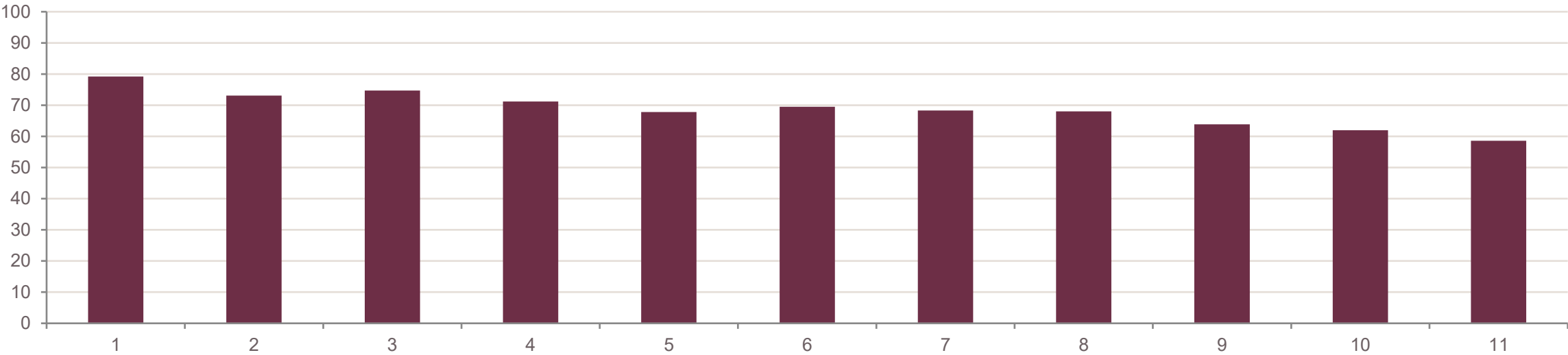
4. RESULTS

기술통계 — 전체 문항 및 Mastery Level

11문항(문항별), Mastery Level로 구성된 GST에 응시한 피험자 59명에 대한 기술통계값

N	평균 (Mean)	중앙값 (Median)	표준편차 (SD)	범위 (Range)
59	5.86	7	2.16	1 – 11

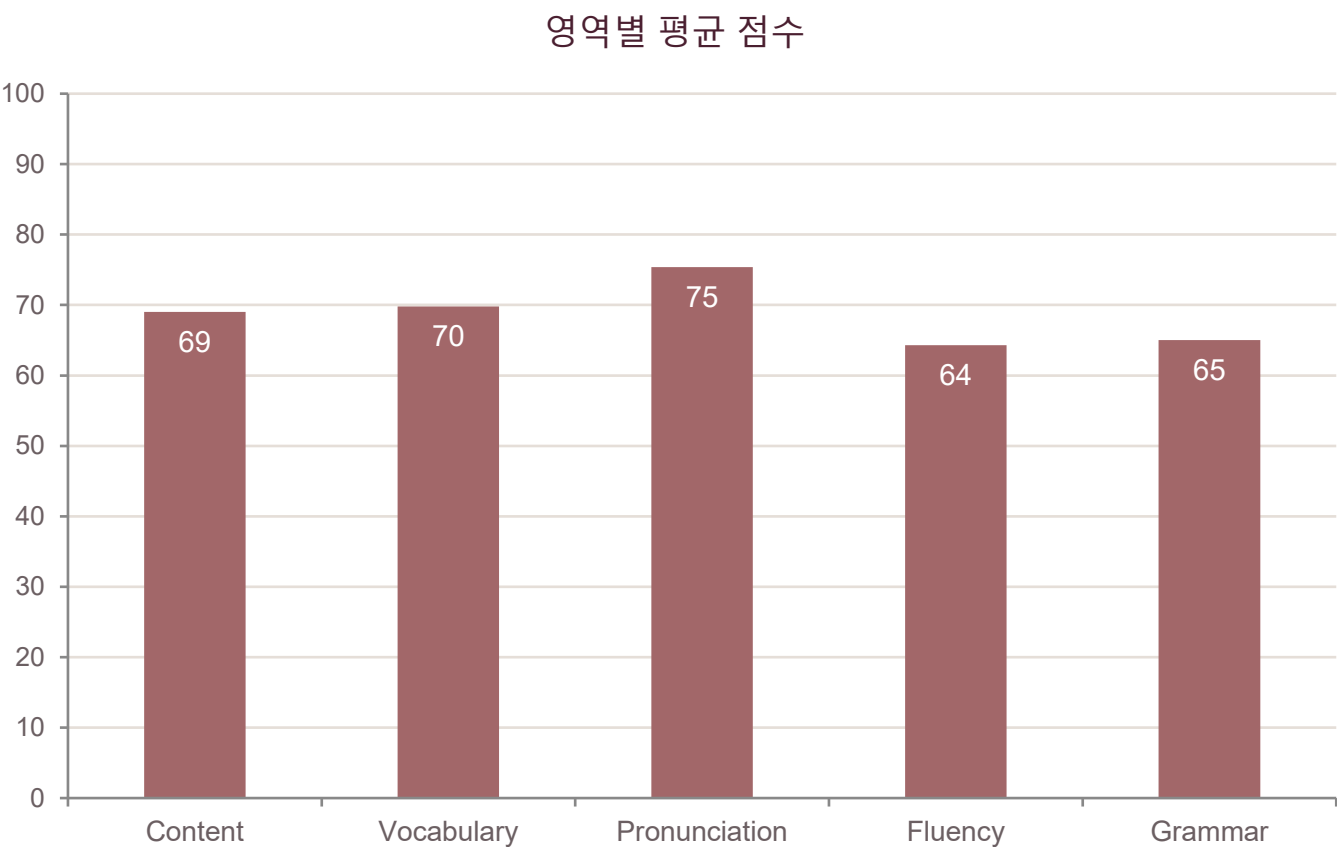
문항별 평균 점수 (1-11번)



4. RESULTS

기술통계 — 영역별(5개 영역)

11문항, 5개 영역(Content, Vocabulary, Pronunciation, Fluency, Grammar)으로 구성된 GST에 응시한 피험자 59명에 대한 기술통계값



영역별 상세 지표

영역	평균	SD
Content	69	9.68
Vocabulary	69.8	10.9
Pronunciation	75.4	4.19
Fluency	64.3	9.44
Grammar	65	9.88

4. RESULTS

신뢰도 분석 (Cronbach's Alpha)

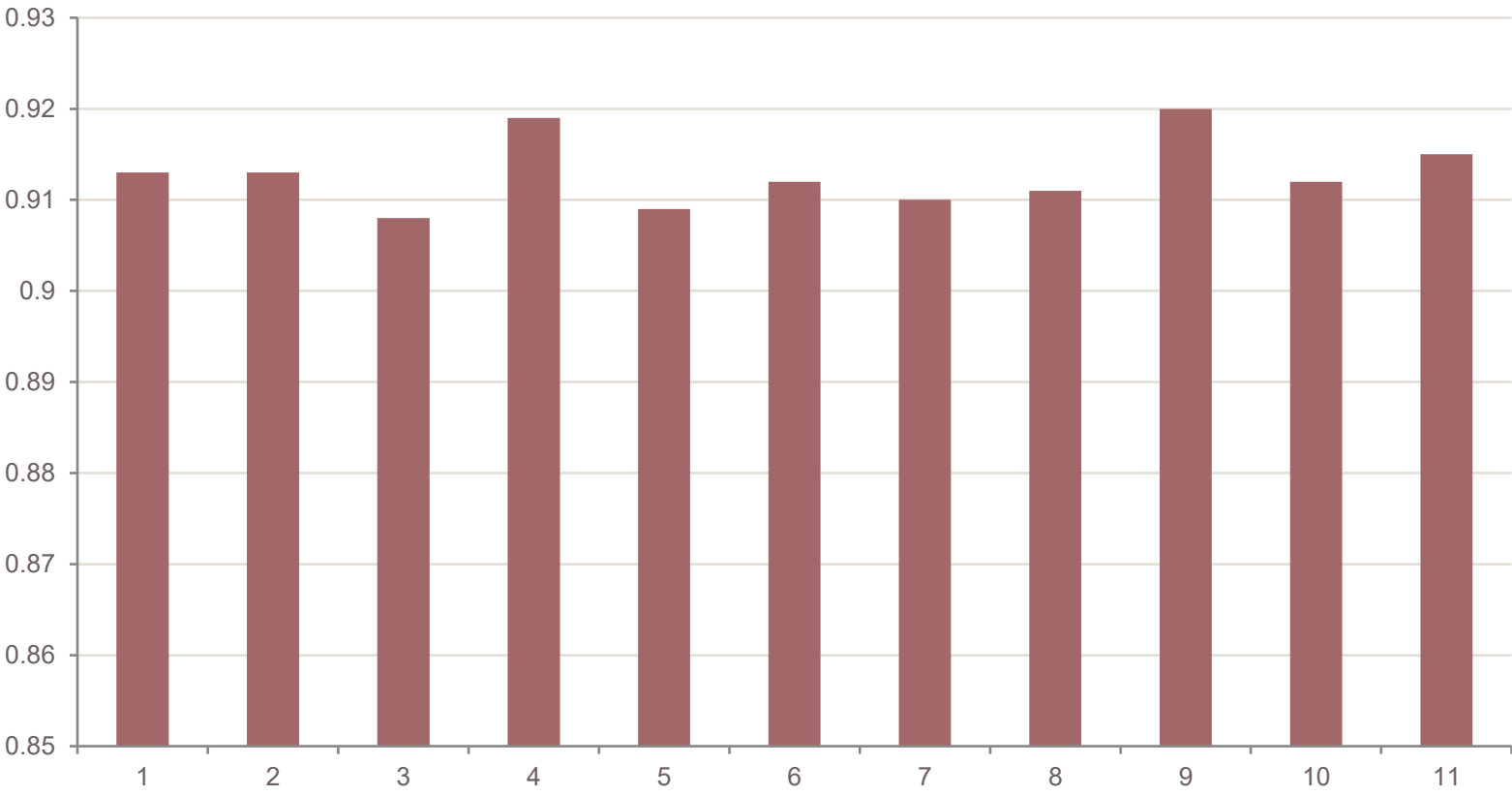
전체 신뢰도

.92

Cronbach's α

실증 기준인 0.7을 초과하여 높은 신뢰도를 보임

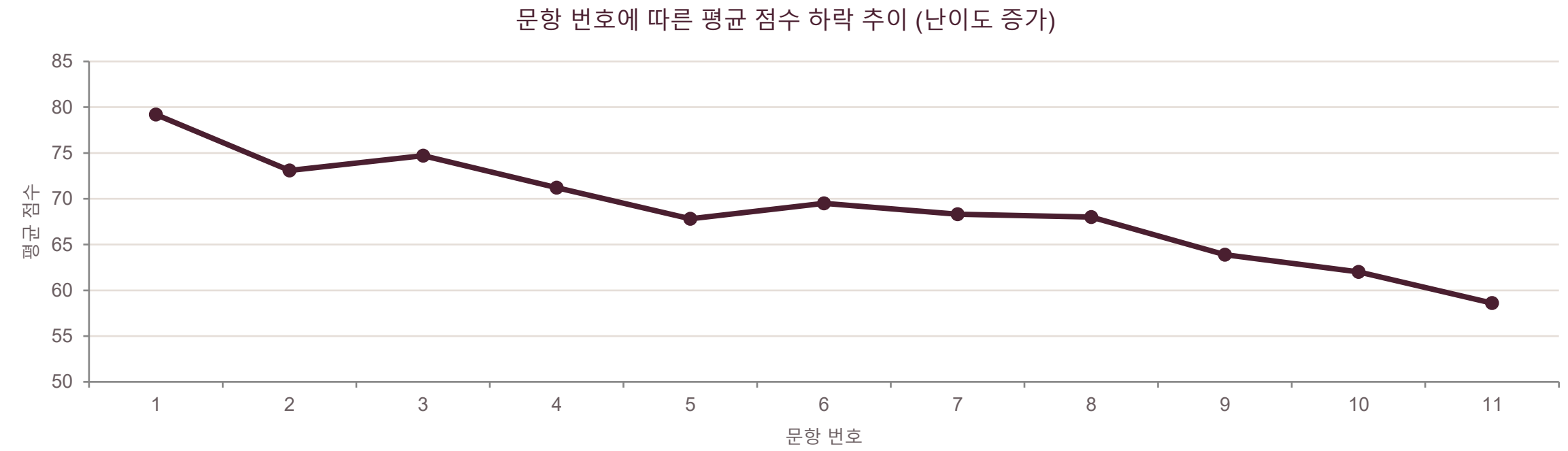
문항별 신뢰도 (해당 문항 제외 시)



4. RESULTS

타당도 분석 — 문항 난이도 분포

- 문항별 영역 피험자 문항 분포: 11번 문항이 제일 어려운 문항, 1번 문항이 제일 쉬운 문항으로 나타남 — 문항이 뒤로 갈수록 점점 어려워지는 것으로 나타남
- 문항 적합도 분석: 내적합도(infit)·외적합도(outfit) 통계값이 부적합 기준값 1.5를 동시에 초과하는 문항은 없는 것으로 나타남. 점이연 상관계수도 지나치게 낮은 문항은 없는 것으로 나타남



CONCLUSION

요약

- 본 연구의 목적은 Rasch 측정모형을 이용해서 지텔프 공인 영어 말하기 시험(GST) 문항의 신뢰도와 타당도를 알아보는 것으로 수행되었음
- 분석 결과, 문항 내적 일관성 신뢰도는 .92로 높게 나타났으며, 타당성에서는 개별 문항에 대한 적합도 분석 결과 모든 문항이 분석 모형에 적합으로 나타남
- at home test의 감독관이 서비스 없이 시행된 pilot test 피험자를 대상으로 분석하였고, 샘플 사이즈가 작아 전체 영어 말하기 성적을 측정하는 시험 분석 결과로 단정하기는 어려운 점이 있음
- 수험자들의 배경 변인을 포함하는 연구 설계가 필요함

논의 및 제언

논의

현재 파일럿 테스트로 샘플 수집 중 연구를 수행하였으며, 향후 보다 큰 샘플 사이즈(300~500명)로 타당도 높은 결과를 도출할 수 있음

제언

- 감독관 서비스를 도입하여 휴먼 팩터의 영향을 살펴보고, 애틀 테스트의 안정화에 대한 연구를 지속함
- 논리적·경험적 근거뿐 아니라 시험의 다양한 이해당사자 집단의 의견을 청취·반영하는 과정을 기술하여, 이러한 정보가 시험 타당도 검증의 중요한 증거로 활용 가능함을 검토할 수 있음
- 인공지능이 언어 평가 자동화 기술과 접목되면 영어를 포함한 외국어 평가 분야를 더 진보시켜 줄 것으로 기대됨

5. CONCLUSION

향후 연구 과제

- 1 샘플 수집 후, 최소 300~500명의 수험자로 파일럿 테스트 타당성 재연구
- 2 기존 말하기 시험(at home version, ex. OPIc, TOEIC Speaking)과의 비교 연구 및 Equating을 통한 점수 환산표 (연구 진행 중)
- 3 회차 간 동일한 수준의 난이도 유지, 난이도 차이의 통계적 보정 절차 실행, 수험자 모집단 특성의 안정적 유지 등과 관련된 연구
- 4 지역별, 연령별, 성별 등의 배경 변인으로 나누어 통계 기법으로 진행하는 후속 연구
- 5 파일럿 테스트 이후 성적 결과의 급상승·급저하 정도 및 객관성·일관성 유지, 문항별 관련성 및 레벨별 차이점을 살펴보는 후속 연구
- 6 자동 문항 채점 도입 이후의 타당성 논의 가능성

참고문헌

- Fulcher, G. (2009). When is test preparation questionable. Retrieved May 2, 2011, from <http://languagetesting.info/features/testprep/testprep.html>.
- Kang, Y. (2021). Validity and authenticity of 2015 Revised National Curriculum language forms. Master's Thesis, Korea University Graduate School.
- Kim, J. (2019). A validity investigation on the absolute grading system of the English section of the Korean Scholastic Aptitude Test, as well as some suggestions for its improvement. *English Language Assessment*, 14(1), 41–59.
- Kim, H. (2019). Foreign language teaching in Korea with reference to CEFR. *Institute for Humanities and Social Sciences (IHSS)*, 20(4), 79–96.
- Seok, J. (2017). A study on differences in using test-taking strategies by different achievement groups in TOEIC. *Journal of Humanities and Social Science (HSS21)*, 8(5), 597–612.
- Seol, H., Kim, S., & Kim, D. (2005). Using Rasch measurement model to evaluate the Marlowe-Crowne Social Desirability Scale. *Journal of Educational Evaluation*, 18(1), 101–123.
- Song, K. (2011). English as an international and global language: Language attitudes & pedagogical implications. *New Korean Journal of English Language & Literature*, 53(1), 201–221.
- Wall, D., & Horak, T. (2011). The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: Phase 3, The role of the coursebook, and Phase 4, Describing change. TOEFL iBT™ Report No. iBT-17. Princeton, NJ: ETS.

감사합니다!

Minjung Kim, PhD • ITSC G-TELP