ITSC International Testing Services Center G-TELP™ General Tests of English Language Proficiency

RESEARCH PAPER

A Validity Study of the IBT Speaking Certification Test

Prepared by

Minjung Kim

Ph.D., Educational Measurement, Department of Education, University of Kansas

Report Date

Oct 22, 2022

Korean Association for Educational Information and Media

목차

G-TELP 연구 키워드

- 타당성
- iBT Speaking test
- Rasch Model
- 신뢰도
- 재택시험

1. Introduction

- **말하기 시험**
  - 통상적으로 언어 평가 영역에서 측정하고자 하는 것은 무엇일까?

  - 그것은 말 그대로 언어 학습자가 가지고 있는 언어적인 능력에 대한 평가이다. 여기에서의
  - **'언어적인 능력'**이라는 것은 해당 언어를 통해 **타인과 의사소통을 수행하는 능력'**을 가리킴.

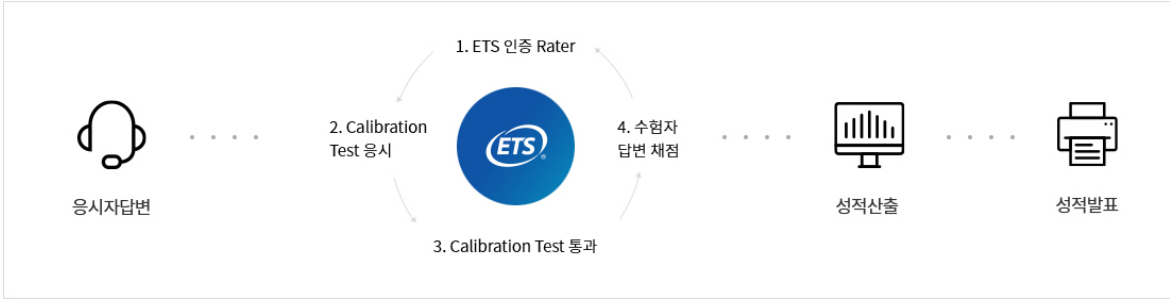
  - 언어평가를 위한 영어 말하기 시험
  - 직장, 학업에서 서류 및 면접을 통과하기 위한 증빙 자료로 이용
  - 학문적 맥락이나 직장에서의 언어 능력
    기업 및 실무 업무에서는 실무 영어 능력 우수자를 선호함.

1. Introduction

- **목적**
  - 타당도는 뒷받침하는 증거의 적절성에 대한 평가적 판단으로, 시험 점수 해석 및 사용 논쟁의 첫걸음으로, 영어 말하기 시험의 at home 테스트 점수 해석 및 활용의 타당성을 검증.
  - 국내 영어 말하기 시험과 평가시스템을 살펴보고 말하기 시험의 구조를 파악함.
  - **영어 말하기 시험의 의의**: 평가라는 영역에서 객관성은 매우 중요한 항목이며, 실제 평가 기준을 활용하여 영어 말하기 능력을 객관적으로 평가하는지 살펴봄.

2. Literature Review

- 말하기 시험
  - TOEIC Speaking
    - 온라인 플랫폼에서 학생들이 끊임없이 참여하고 도전, 지속적으로 언어 능력을 향상
    - 채점과정



  - 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 정확하게 이해하고 답변, 답변 내용과 문장에서 주제 전달이 명확한지

2. Literature Review

- 말하기 시험
  - GST (GTELP Speaking Test)
    - 미국 ITSC(International Testing Services Center)에서 개발된 국제공인 영어 말하기 시험으로,
    - 시험시간 약 30분, 11개 parts 30여개 질문, 평가기준(Content, Grammar, Vocabulary, Pronunciation)
    - 교차 채점으로 수험자의 말하는 능력을 분석, 진단함.

- 채점자 교육방법 : 매년 주관기관의 채점척도 적용의 표준화 훈련 등의 G-TELP Rater 양성 및 교육프로그램을 통해 이루어짐.

ETS TOEIC® Speaking YBM

OPIc Oral Proficiency Interview - comput

G-TELP™ General Tests of English Language Proficiency Speaking

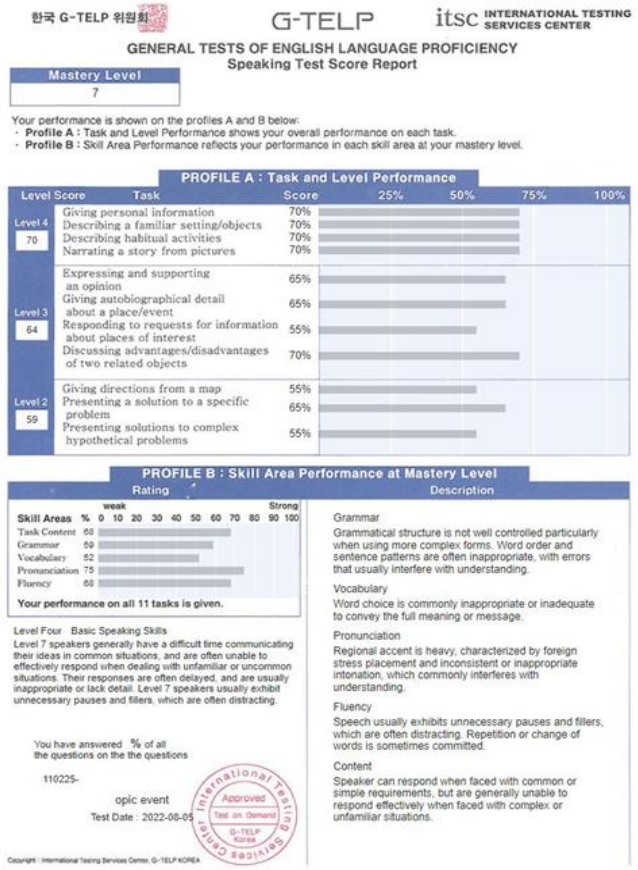
2. Literature Review

- **말하기 시험의 타당성**
  - 영어 말하기 평가의 인증여부는 도구의 신뢰도와 평가의 타당성이 우선되어야 함.
  - 평가도구마다 평가기준, 평가영역, 평가등급 및 방법, 평가시간이 모두 다르게 구성되어 있는데, 그 중 지텔프 말하기 시험(GST)의 타당도를 알아봄.

  - GST는 표현력, 문법, 어휘, 발음, 유창성 등의 5가지로 평가도구마다 서로 상이하게 평가기준이 구성되어 있음.
  - 말하기 평가가 기준이 다르므로 동일한 대상의 수험자를 어느 평가도구로 측정하느냐 에 따라 신뢰도 문제로 연결됨.
  - 말하기 평가는 의사소통능력을 평가하는 것으로 질문을 잘 이해해서, 정확한 어법을 사용해 적절한 어휘를 정확하게 사용하여 상대방이 알아들을 수 있는 발음으로 자연스럽게 말하는가를 평가하는 것임.

2. Literature Review

- **공인 인증 말하기 시험의 타당성**
  - 평가방식 : GST는 2인 1조의 평가자에 의해 평가가 이루어지며, 평가자의 결과에 대한 평균점수를 결과로 산출함.
  - 말하기 평가는 시간, 날씨, 평가자의 컨디션에 따라 주관이 개입될 수 있기 때문에 본인이 원하는 시간에 집에서 시험을 칠 수 있는 환경을 구성하여 신뢰성을 얻도록 함.
  - Rasch 측정모형을 이용하여 공인 말하기시험(GST)에서 신뢰도와 타당도를 검증함.



한국 G-TELP 위원회 G-TELP itsc INTERNATIONAL TESTING SERVICES CENTER

GENERAL TESTS OF ENGLISH LANGUAGE PROFICIENCY
Speaking Test Score Report

Mastery Level: 7

Your performance is shown on the profiles A and B below.
- **Profile A** : Task and Level Performance shows your overall performance on each task.
- **Profile B** : Skill Area Performance reflects your performance in each skill area at your mastery level.

PROFILE A : Task and Level Performance

Level	Score	Task	Score	25%	50%	75%	100%
Level 4	70	Giving personal information	70%				
		Describing a familiar setting/objects	70%				
		Describing habitual activities	70%				
		Narrating a story from pictures	70%				
Level 3	64	Expressing and supporting an opinion	65%				
		Giving autobiographical detail about a place/event	65%				
		Responding to requests for information about places of interest	55%				
		Discussing advantages/disadvantages of two related objects	70%				
Level 2	59	Giving directions from a map	55%				
		Presenting a solution to a specific problem	65%				
		Presenting solutions to complex hypothetical problems	55%				

PROFILE B : Skill Area Performance at Mastery Level

Rating

Skill Areas	%	weak 0 10 20 30 40 50 60 70 80 90 100 Strong
Task Content	68	
Grammar	59	
Vocabulary	62	
Pronunciation	75	
Fluency	68	

Your performance on all 11 tasks is given.

Level Four Basic Speaking Skills

Level 7 speakers generally have a difficult time communicating their ideas in common situations, and are often unable to effectively respond when dealing with unfamiliar or uncommon situations. Their responses are often delayed, and are usually inappropriate or lack detail. Level 7 speakers usually exhibit unnecessary pauses and fillers, which are often distracting.

You have answered % of all the questions on the the questions

110225-

opic event

Test Date : 2022-08-05

Approved Test on Demand G-TELP Korea

Description

Grammar
Grammatical structure is not well controlled particularly when using more complex forms. Word order and sentence patterns are often inappropriate, with errors that usually interfere with understanding.

Vocabulary
Word choice is commonly inappropriate or inadequate to convey the full meaning or message.

Pronunciation
Regional accent is heavy, characterized by foreign stress placement and inconsistent or inappropriate intonation, which commonly interferes with understanding.

Fluency
Speech usually exhibits unnecessary pauses and fillers, which are often distracting. Repetition or change of words is sometimes committed.

Content
Speaker can respond when faced with common or simple requirements, but are generally unable to respond effectively when faced with complex or unfamiliar situations.

3. Methodology

- **연구 절차**
  - 타당도는 뒷받침하는 증거의 적절성에 대한 평가적 판단으로, 말하기 영어인증시험(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(Rasch, 1960) 공식에서 Bn은 피험자의 능력, Di는 문항 난이도를 의미함.

$$\left(\frac{Pnik}{1-Pnik}\right) = B_n - D_i$$

3. Methodology

▪ 연구 절차

- Rasch 모형에서는 개별 문항이 모형에 적합한지를 검증하기 위해서 **외적합도와 내적합도 지수 값을 제공**하고 있음.
- **외적합도 지수**(Outfit : Outfit Mean Sauqre)는 피험자의 능력수준에 비추어서 너무 쉬운 문항을 틀리거나 너무 어려운 문항을 맞힐 경우와 같은 이상 문항반응 형태에 민감한 지수로 공식을 다음과 같이 제시함.

$$MS(WT)_i = \frac{\sum_{n=1}^{N}(X_{ni} - P_{ni})^2}{\sum_{n=1}^{N} P_{ni}(1 - P_{ni})}$$

- 외적합도 지수의 공식에서, 피험자의 **능력수준에서 많이 벗어난 문항에 대해 민감함을 알 수 있음.**
- 피험자의 능력수준에서 많이 벗어나지 않은 문항의 이상 문항반응 형태에 가중치를 둔 적합도 지수가 내적합도 지수(Infit: Infit Mean Saure)이며 공식을 제시하면 다음과 같음.
- **외적합도 지수와 내적합도 지수 값은 1.0을 기준으로 1.5 이상의 값을 보이면 문제가 있는 문항으로 간주함(Linacre, 2009)**

$$MS(UT)_i = \left(\frac{1}{N}\right)\sum_{n=1}^{N}\frac{(X_{ni} - P_{ni})^2}{P_{ni}(1 - P_{ni})}$$

3.Methodology

- 연구방법
  - 연구대상
    - GST 채점 기준표, 영어 말하기, GST 수험자의 파트별 시험점수(2022년 시행) 로 분석을 수행함.
  - 분석 도구
    - 시험이 어떻게 개발 운영되어왔는지를 파일럿 테스트 및 전 과정 모니터 기간에 데이터를 수집 및 분석.
    - 수험자의 파트별 시험결과 및 demographic 분석, 시험 점수의 타당도 검증 연구를 수행.
    - 기술통계 및 시험의 신뢰도 및 타당도는 Rasch Model을 수행,
    - R 프로그램을 GUI(Graphic User Interface)방식으로 구현한 오픈소스 통계 프로그램 jamovi(2021 : R Core Team, 2021), snowRMM(Seol, 2021; Willse, 2014) 통계 모듈을 이용하여 분석을 수행함.

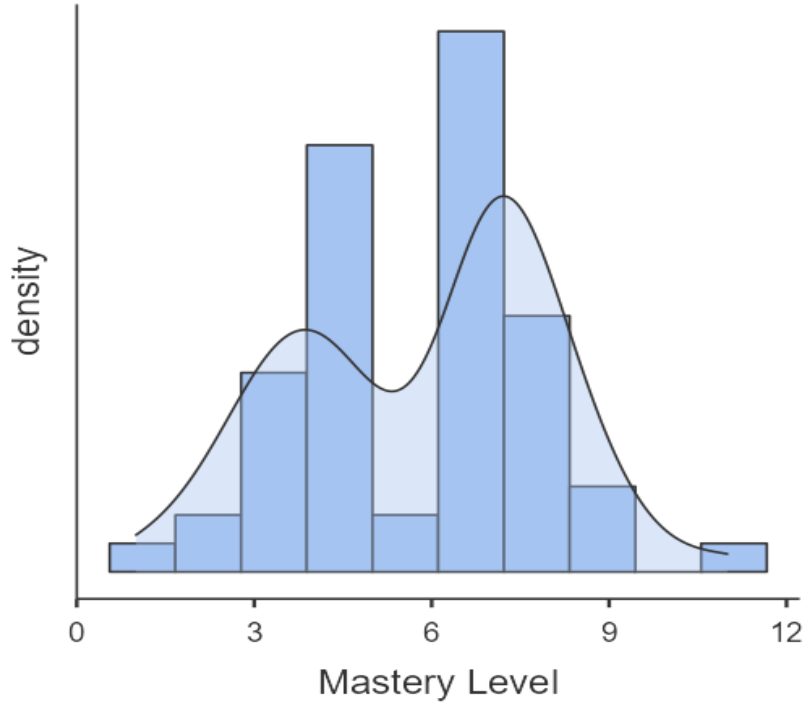
4. Results

- **기술통계**
  - 11문항(문항별), Mastery Level로 구성된 GST에 응시한 피험자 59명에 대한 기술통계값을 제시하면 다음과 같음. 총 등급 평균은 5.86이며, 표준편차는 2.16으로 나타남.

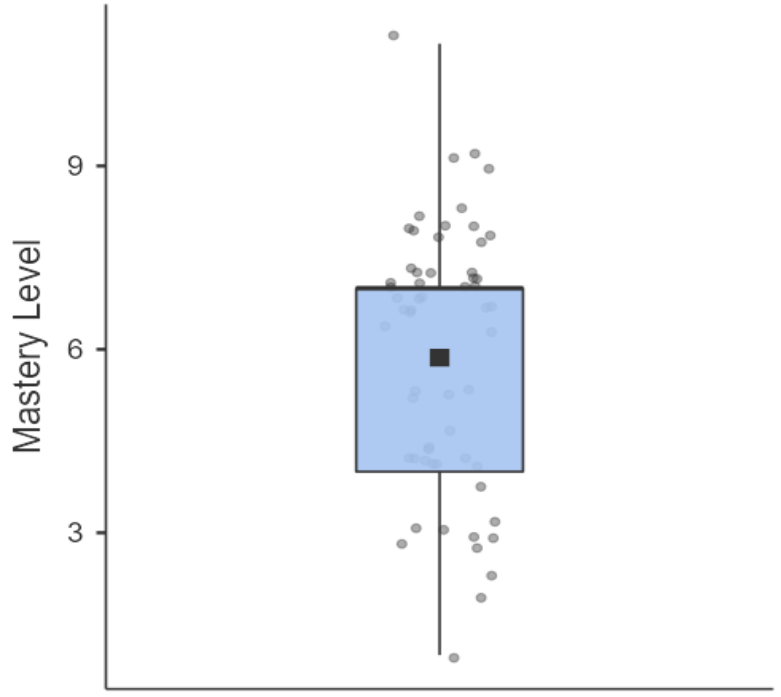
Descriptives

	1	2	3	4	5	6	7	8	9	10	11	Mastery Level
N	59	59	59	59	59	59	59	59	59	59	59	59
Missing	0	0	0	0	0	0	0	0	0	0	0	0
Mean	79.2	73.1	74.7	71.2	67.8	69.5	68.3	68.0	63.9	62.0	58.6	5.86
Median	75	75	75	70	65	70	65	65	65	60	60	7
Standard deviation	7.55	10.0	9.46	10.7	10.3	9.18	9.63	10.2	12.8	8.62	12.4	2.16
Variance	57.0	101	89.5	114	105	84.2	92.8	104	164	74.2	153	4.67
Minimum	60	35	35	35	45	50	35	35	0	50	0	1
Maximum	95	95	100	95	95	100	95	95	90	90	95	11

4. Results



- Mastery Level 히스토그램



- Mastery Level 박스 플랏

4. Results

- **기술통계**
  - 11문항, 5개 영역(Content, Vocabulary, Pronunciation, Fluency, Grammar)으로 구성된 GST에 응시한 피험자 59명에 대한 기술통계값을 제시하면 다음과 같음.

Descriptives

	Task Content	Vocabulary	Pronunciation	Fluency	Grammar
N	59	59	59	59	59
Missing	0	0	0	0	0
Mean	69.0	69.8	75.4	64.3	65.0
Median	68	68	75	61	64
Standard deviation	9.68	10.9	4.19	9.44	9.88
Minimum	45	43	61	32	43
Maximum	91	100	100	89	95

4. Results

- **신뢰도 분석**
  - 11문항으로 구성된 GST의 문항내적 일관성
  - **신뢰도(Cronbach Alpha)는 .92**로 실증기준인 0.7을 초과하여 높게 나타남.

  - 개별 문항의 개별 신뢰도를 오른쪽 표에 제시하였음.

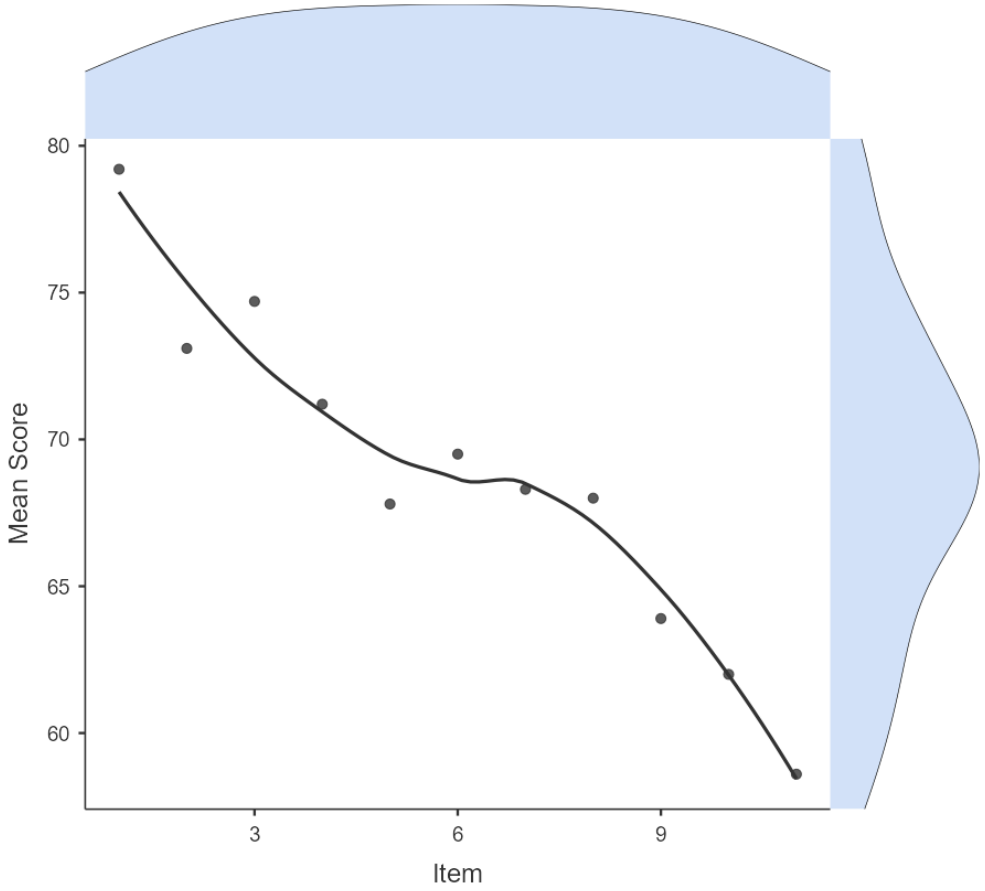
Item Reliability Statistics

	if item dropped
	Cronbach's α
1	0.913
2	0.913
3	0.908
4	0.919
5	0.909
6	0.912
7	0.910
8	0.911
9	0.920
10	0.912
11	0.915

4. Results

- **타당도 분석**
  - **문항별 영역 피험자 문항 분포** : 각 문항에서 11번 문항이 제일 어려운 문항으로 나타났으며, 1번 문항이 제일 쉬운 문항으로 나타남. 문항이 뒤로 갈수록 점점 어려워지는 것으로 나타남.

  - **문항 적합도 분석** : 말하기 시험 영역 11문항의 문항 타당도 분석 결과를 제시함. 분석 결과, 내적합도(infit), 외적합도(outfit) 통계값으로 부적합 기준값이 1.5를 동시에 초과하는 문항은 없는 것으로 나타남. 또한, 점이연 상관계수도 지나치게 낮은 값을 보이는 문항은 없는 것으로 나타남.



5. Conclusions

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

5. Conclusions

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

- **제언**
  - 감독관 서비스를 도입하여 휴먼 팩터의 영향을 살펴보고, 앳홈 테스트의 안정화에 대한 연구를 지속적으로 연구함.
  - 논리적, 경험적 근거뿐 만 아니라 시험의 다양한 **이해당사자 집단의 의견을 청취하고** 반영하는 과정들을 기술하여 이러한 정보들이 시험 타당도 검증의 중요한 증거로 활용 가능함을 검토할 수 있음.
  - 인공지능이 언어 평가 자동화 기술들과 접목되면 영어를 포함한 외국어 평가 분야를 더 진보시켜 줄 것으로 기대됨.

5. Conclusion

- **향후 연구 과제**
  - 샘플 수집 후, 최소 300~500명의 수험자의 파일럿 테스트 타당성 재연구
  - 기존 말하기 시험(at home version, ex. OPIc, TOEIC speaking)과의 비교 연구 및 Equating을 통한 점수 환산표(연구 진행중)
  - 회차 간 동일한 수준의 난이도 유지, 난이도 차이의 통계적 보정 절차 실행, 수험자 모집단 특성의 안정적 유지 등과 관련된 연구

  - 지역별, 연령별, 성별 등의 배경변인으로 나누어 통계기법으로 후속연구도 가능함.
  - 파일럿 테이스 이후, 성적 결과의 급성정, 급저하 정도 및 객관성, 일관성 유지를 보이고, 문항별 관련성 및 레벨별 차이점을 살펴보는 후속 연구 가치도 살펴봄.
  - 추후, 자동 문항 채점 도입 이후 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.

참고문헌

- 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. Retrieved August 1, 2013, from https://www.ets.org/Media/Research/pdf/RR-11-41.pdf.

감사합니다!