Investigating the Construct Validity of an ESL Test for Young Learners

Don Yao[1], Linyu Liao[2], & Xueting Ye[1]

[1] Department of English, Faculty of Arts and Humanities, University of Macau, Macau, China

[2] School of Foreign Languages, Guangdong Medical University, Dongguan, China

Correspondence: Linyu Liao, School of Foreign Languages, Guangdong Medical University, Dongguan, China

Received: January 11, 2022 Accepted: April 24, 2022 Online Published: May 2, 2022

doi: 10.5430/wjel.v12n5p59 URL: https://doi.org/10.5430/wjel.v12n5p59

Abstract

This study investigated the construct validity and reliability of the GEPT-Kids, which is composed of 25 listening items and 30 reading items. Data were collected from 742 test takers from Grades 5 and 6 from eight Chinese-speaking schools in Beijing, Shanghai, Guangzhou, Hong Kong, Macau, Taipei, Taichung, and Kaohsiung. Two CFA models (*Model 1*: a one-factor model with a general language proficiency construct; *Model 2*: a two-factor model with listening and reading constructs based on the test design) were proposed and analyzed using AMOS 24.0 to reveal whether the internal structure of the test reflected the test developers' design for the test. The results showed *Model 1* provided an adequate explanation of the data as most of the model fit indices were acceptable, and *Model 2* showed that there was a lack of discriminant validity between listening and reading because they were too highly correlated although the fit indices were high. While both models can be considered acceptable based on fit indices, and *Model 1* is more parsimonious than *Model 2*. To sum, the main finding was that GEPT-Kids is valid in terms of score interpretation of young learners' overall language ability but may lack discrimination in terms of assessing subskills of English language proficiency. Therefore, findings from this study suggest that GEPT-Kids is acceptable as a general language proficiency test but should be used with caveats while using test scores in terms of listening and reading in the classroom teaching-learning context.

Keywords: construct validity, ESL, young learners, CFA, GEPT-Kids

1. Introduction

Validity, defined as the degree to which evidence and theory support the interpretations of test scores by proposed uses of tests (AREA, APA, & NCME, 2014), is always an indispensable enterprise for a test in the arena of language assessment. Kunnan (2018) claims the validity framework is composed of four aspects comprising content-related evidence, criterion-related evidence, construct-related evidence, as well as social or consequential consequences. In other words, an authentic assessment is expected to accurately measure test takers' language ability by offering reasonable score interpretations, reaching high reliability, showing justice and fairness, and providing positive washback or consequences to the classroom teaching-learning context or even the society (Xi, 2010; Kunnan, 2012; Yao, 2021). Among different kinds of test validity, construct validity, the degree to which a test measures what it claims, or purports, to be measuring, is particularly emphasized by numerous applied linguists (Brown, J. 1996; Bachman & Palmer, 1996; Brown, H., 2004). However, previous construct validity research has mainly focused on the tests assessing adult language learners, research centered on young language learners, who are learning a foreign or second language and who are doing so during the first five to seven years of formal schooling, may remain a lacuna (McKay, 2006). Given that the last decade has witnessed a renewed upsurge of interest in assessing young language learners (e.g., McKay, 2006; Bailey, 2008; Hasselgreen & Caudwell, 2016; Papp & Rixon, 2018), the construct validity of tests for young language learners does merit scholarly attention.

Based on Kunnan and Liao's (2019) research on the assessment of young learners, the current study investigated the construct validity of the General English Proficiency Test-Kids (GEPT-Kids) in terms of its score interpretations. The GEPT-Kids for young learners, developed by the Language Training and Testing Centre (LTTC) in Taiwan, is a test specifically tailored to elementary school students. The results would indicate whether GEPT-Kids is valid or accurately assesses young test takers' overall language ability and subskills, and whether the test could be utilized in Taiwan or even all over China.

2. Literature Review

2.1 Construct Validity of GEPT and Other ESL Tests

There is a growing body of research on the construct validity of English as a Second Language (ESL) tests in the field of language assessment over the last two decades (e.g., Liao, 2009; Sawaki et al., 2016; Moore et al., 2006). Liao (2009) investigated the predictive validity of lexico-grammatical knowledge on L2 reading and listening abilities of the GEPT among 609 Taiwanese first-year students. Multivariate G-theory analysis was first conducted, revealing that test items were able to distinguish the English proficiency of test takers. Reading items measuring literal meaning were better than reading items appraising pragmatic meaning. In terms of listening, both kinds of items worked equally well. Confirmatory factor analysis (CFA) was secondly proposed to examine the internal structure of the GEPT. Results showed both reading and listening comprehension tasks intended to test the literal and pragmatic meaning. The structural equation modeling (SEM) approach was finally adopted to explore the relationship among lexico-grammatical knowledge, L2 reading ability, and L2 listening ability. Results showed that the lexico-grammatical knowledge was a more powerful predictor of reading abilities than listening abilities, and semantic meaning enumerated a stronger relationship with listening comprehension than reading comprehension, but they were both more potent than the grammatical knowledge.

Sawaki et al. (2016) examined the construct validity of the Foreign Language™ Internet-based test (TOEFL® iBT) in terms of its scored item responses among 2,720 English as a Foreign Language (EFL) test takers. The CFA model was proposed to examine the internal structure of the test: *Model 1* a single higher-order general factor (the entire test) and four first-order factors (the four sections), and *Model 2* a newly developed integrated speaking and wring section. Results showed that *Model 1* was overwhelmingly supported, and it was the best representation of the test's factor structure. *Model 2* also reasonably defined the target constructs, indicating that the construct of integrated speaking and writing section was minimally involved in the reading and listening constructs.

Moore et al. (2006) probed the construct validity of the International English Language Testing System (IELTS) academic reading test by comparing the reading requirements in IELTS test items and in university study. Weir and Urquhart's (1998) taxonomic framework was adopted focusing on two dimensions: level of engagement and type of engagement. Results showed that the bulk of the IELTS tasks had a "local-literal" configuration requiring a fundamental comprehension of small textual units. In terms of the academic corpus, the *local-literal* orientation was also found in a sizeable number of tasks, but a critical evaluation of material was more needed. Therefore, researchers suggested that reading tasks should be more *global-interpretative* in academic study to enhance the reading construct validity.

The abovementioned research has both theoretically and methodologically provided a comprehensive view of the construct validity of ESL tests. Nevertheless, the participants recruited in this research have only touched upon adult language learners. McKay (2006) maintains there has been little published research into the assessment of young learners. Since there has been a tendency that children begin to learn a foreign language at a younger age, especially in Asia and Europe (Alderson & Bachman, 2000), young ESL learners are supposed to be attached great importance.

2.2 Assessing Young Language Learners

The renewed lingering concern regarding the assessment of young language learners has started over the last decade (e.g., Kunnan & Liao, 2019; Butler & Lee, 2010; Elder & Zammit, 1992). Kunnan and Liao (2019) investigated the relationship between young learners' self-assessment, language attitude, and test performance. A total of 398 participants from Hong Kong, Taipei, Taichung, and Kaohsiung were recruited for the study. The instruments used in the study were a paper-and-pencil-based GEPT-Kids test paper and a bilingual survey designed by LTTC. Both CFA and SEM were adopted for statistical analyses. Results revealed that young language learners' learning attitude had little influence on test performance, whilst there is a high positive correlation between self-assessment and test scores.

Butler and Lee (2010) examined the effectiveness and self-assessment among 254 grade six ESL students in Seoul. The study proposed two types of assessment with a general assessment and a series of assessments designed for each lesson. Students' English performance was measured by two sets of tests: the movers' level of the Cambridge Young Learners English Test (CYLE) and the Test of English at Seoul City Elementary Schools (TESCES). Results showed that teachers and students perceived the effectiveness and self-assessment differently. In terms of young learners separately, they improved their ability of self-assessment and became more confident in learning English.

Elder and Zammit (1992) probed the validity and reliability of the Australian Language Certificates (ALC) among 8,300 young language learners. The ALC program, a language skill testing program in which students respond to

realistic texts and situations with questions in a multiple-choice format, offers students national recognition for levels of achievement in their studies of languages other than English (ALC website, n.d.). The results showed that most of the tasks were relatively easy, and few young learners failed to achieve level 1 in either reading or listening comprehension. Also, it was not difficult for the young learners to achieve level 3 in reading comprehension, particularly in Japanese, and in listening comprehension in Chinese and Italian. The test is, on the whole, superb to provide positive washback or consequences, but it may be a deficiency for students with higher language abilities.

These empirical studies have laid emphasis on language assessment for young learners. The impact of the test and the validity and reliability of the test are both covered. However, whether the construct underlying the test is valid is rarely mentioned. In other words, the scarcity lies in the research that investigates the construct validity of ESL tests.

2.3 Research Questions

It is noteworthy to point out that previous research has either overlooked the essentiality of young language learners or neglected the indispensability of the construct validity of an assessment designed for young language learners. Furthermore, since the GEPT-Kids attracts the attention of numerous young language learners, the construct validity issues become rather critical because an invalid test may bring about detrimental consequences to the classroom teaching-learning environment or even to society. In light of the gaps from the previous research, two research questions were articulated in this study:

RQ1: To what extent are the test items of GEPT-Kids reliable?

RQ2: To what extent does the GEPT-Kids measure test takers' receptive language ability?

3. Methods

3.1 Participants

The participants of this study were 791 Grade 5 and Grade 6 students from eight cities, namely, Beijing, Shanghai, Guangzhou, Macau, Hong Kong, Taipei, Kaohsiung, and Taichung. In each city, four classes of Grade 5 and Grade 6 students from one or two public schools participated in the current study. The participants were aged from 10 to 11 years old while data collection was conducted. They all share the same L1 Chinese and learn English as a foreign language, but their English learning experience varies to some extent. Most of the participants (43%) have started learning English since Grade 3 or Grade 4; 39% of the participants have started learning English since Grade 5 or Grade 6; the remaining participants (18%) have started learning English before Grade 3.

3.2 Instrument

A paper-and-pencil-based GEPT-Kids test paper was used for data collection. The GEPT-Kids, developed by LTTC in Taiwan, is a test specifically tailored to Taiwanese elementary school students (Lee, 2015a, 2015b). The content of the test contains both life experiences and the local curriculum of Taiwanese elementary school students. The test is designed to assess test takers' overall language ability of understanding and communicating in basic English equal to CEFR A1 level (Liao & Yao, 2021). All the test takers will receive diagnostic feedback on four subskills containing listening, speaking, reading, and writing (Kunnan & Wu, 2017).

The test paper used in this study only includes two sections: listening and reading. The listening section consists of four tasks: Judging Task 1 (Items 1 to 5), Judging Task 2 (Items 6 to 11), Matching (Items 12 to 18), and Multiple choice (Items 19 to 25). The two judging tasks require test takers to judge whether the sentence they hear is consistent with the picture they see; the matching task requires the test takers to match the sentence they hear and find out the correct picture; the multiple-choice task requires the test takers to choose the correct answer based on a brief conversation they hear. The reading section is composed of three tasks: Judging (Items 1 to 20), Cloze (Items 21 to 25), and Reading comprehension (Items 26 to 30). The judging task requires the test takers to judge whether sentences correctly describe the pictures; the cloze task requires test takers to choose a word from four choices that can complete the passage; the reading comprehension task requires the test takers to choose the correct answer to each question based on the reading passages.

3.3 Procedures

Four research assistants from a comprehensive university in Macau went to two primary schools each located in the abovementioned eight cities for data collection. The written consent was obtained from both principals and participants before the test with their signatures. The GEPT-Kids test paper was administered to the participants in their classrooms with similar settings. It took the participants no more than 60 minutes to finish the test. After test administration, all the test items were scored by Scantron dichotomously. Then, scores different from item level were input, and subtotal scores for each task were computed for further statistical analyses.

3.4 Analyses

Three types of statistical analyses were conducted for the current study: descriptive analysis, reliability analysis, and CFA. Firstly, descriptive analysis was conducted to gain an overall understanding of test takers' performance at item and task levels. Then, reliability analysis was conducted with item scores to examine the reliability of the test results. Lastly, based on the test design, two CFA models were proposed and examined to check the internal structure of GEPT-Kids. The first CFA model is a one-factor model, with general language proficiency being the latent factor or construct and subtotal scores for each test task being the observed variables. The second CFA model is a two-factor model, with listening and reading abilities being the two latent factors or constructs and subtotal scores for listening and reading tasks being the observed variables for each factor.

4. Results and Discussion

RQ1: To what extent are the test items of GEPT-Kids reliable?

4.1 Descriptive Analyses

Item level descriptive analysis shows that mean scores of each item range from .59 to .98 and most of the items have an average score of over .80. This result indicates that most of the items were easy for the participants. Table 1 shows mean scores at task levels, which further proves that the test difficulty is low for the test takers. According to the maximum mark column, it is found that the highest mark for each task is the full mark. These results are not surprising because the directions illustrated in the test paper are in Chinese, meaning that test takers have no difficulty understanding the intentions of test items. Besides, at the beginning of both listening and reading sections, a sample item has been displayed to ensure that every test taker could understand the meaning. Additionally, most of the items contain a colored picture beside, which may offer a better description to test takers. Furthermore, all the listening scripts are played twice, which may raise the accuracy of each listening item. Meantime, in terms of the reading section, high-level words that may be beyond the capacity of test takers' understanding are all illustrated with their Chinese meanings. Methodologically, the skewness and kurtosis values of each task show that normal distribution is violated. Thus, the bootstrapping technique was adopted for further CFA.

Table 1. Descriptive Statistics at Task Levels (n = 791)

	Min	Max	Mean	SD	Skewness	Kurtosis
LT1	1.00	5.00	4.59	.75	-2.06	4.31
LT2	.00	6.00	5.30	1.11	-1.73	2.63
LT3	.00	7.00	6.56	1.07	-3.22	11.65
LT4	.00	7.00	5.33	1.88	-1.01	-.08
RT1	7.00	20.00	17.76	2.69	-1.51	1.87
RT2	.00	5.00	4.18	1.22	-1.43	1.10
RT3	.00	5.00	3.72	1.53	-.95	-.32

Notes. LT: Listening Task; RT: Reading Task

4.2 Reliability Analyses

Reliability analyses were conducted three times among 1) all the test items, 2) all the listening items, and 3) all the reading items. As the following table shows, Cronbach's alpha is the highest when the reliability analysis is conducted among all the 55 items, revealing that all the 55 items measure the same construct. When the reliability analysis is conducted with listening and reading items, Cronbach's alpha is still high (.87 and .88 respectively), indicating high test reliability in each section. These results also illuminate that the GEPT-Kids test is reliable in assessing the overall language ability and individual language skills such as listening or reading. This result is consistent with the previous results that the test items are easy to test takers and the highest mark for each task is the full mark. Furthermore, it is also consistent with Kunnan and Liao's (2019) result that the items of GEPT-Kids are reliable. However, the reliability coefficients of the overall ability are higher than the individual skills. It may be because some test takers performed relatively better on the listening section, but others performed better on the reading section. Overall, they show consistency in terms of test scores.

Table 2. Results of Reliability Analyses

Category of items	No. of items	Cronbach's alpha
Total items	55	.93
Listening items	25	.87
Reading items	30	.88

RQ2: To what extent does the GEPT-Kids measure test takers' receptive language ability?

4.3 Confirmatory Factor Analysis (CFA)

CFA was conducted to examine the two models introduced earlier. Figure 1 and Figure 2 illustrate the one-factor and two-factor models respectively. CFA results of the two models show that all the regression weights and correlation coefficients are significant at the .001 level. According to Figure 1, all the factor loadings are above .59, indicating that the observed variables (i.e., tasks scores) could be well explained by the latent factor language ability. Similarly, in *Model 2*, the observed variables could also be well interpreted by the two latent factors (i.e., listening and reading abilities). However, in *Model 2*, the two latent factors are too highly correlated (r = .97), illuminating that the two factors lack discriminate validity. This is in line with Liao's (2009) outcome that GEPT shows a well internal structure in terms of listening and reading abilities.

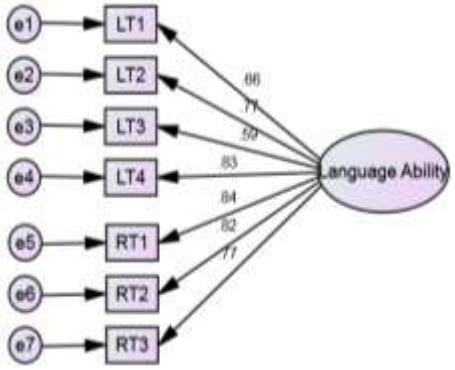

Figure 1. One-factor Model (Model 1)

Note. LT: Listening Task; RT: Reading Task

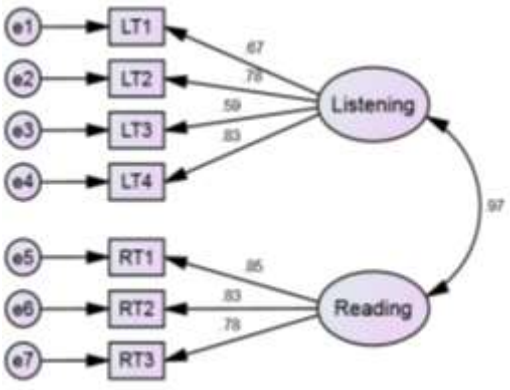

Figure 2. Two-factor Model (Model 2)

Notes. LT: Listening Task; RT: Reading Task

Table 3 further summarizes the model fit indices of the two models. As Table 3 shows, the two models have the same model fit indices. Three out of the four frequently used indices (TLI > .95, CFI > .95, SRMR < .03, see Heubeck & Neil, 2000, Hu & Bentler, 1999) demonstrate acceptable model fit. Only RMSEA is slightly higher than criterion .08 (Hu & Bentler, 1999). These results show that the two models are acceptable in terms of model fit.

Table 3. Model Comparison

Model	Factor(s)	Variables	Model fit indices				Evaluation
			TLI	CFI	RMSEA	SRMR	
Model 1 (Figure 1)	Overall Language ability	Task scores	.96	.97	.09	.03	Acceptable
Model 2 (Figure 2)	Listening ability & Reading ability	Task scores	.95	.97	.09	.03	Acceptable

To sum up, the two models present acceptable model fit and the factor(s) in each model could well explain the test scores. However, due to the discriminate validity issue in *Model 2*, *Model 1* (i.e., the one-factor model) seems to better explain the internal construct of GEPT-Kids. This indicates that GEPT-Kids only assesses one general

construct (i.e., overall input language ability) but not individual listening and reading constructs. That said, listening and reading constructs are intertwined and not able to be distinguished. It may be owing to the low difficulty of test items since most test takers got a high mark in the test. Additionally, as mentioned earlier, most test items include a picture aside. The hidden problem is that even though test takers may not understand the meaning of the item, they can still have a general sense of the context of the test item and, hence, guess the possible answer. The help of pictures may reduce the difference between listening and reading sections. Furthermore, the response format includes only the multiple-choice, which may dig out another reason why the construct of reading and listening sections is hard to distinguish. Thus, test developers may need to pay more attention to the hints or response format of the test while developing the test. This is also similar to Moore et al.'s (2006) investigation of IELTS with the conclusion that a critical evaluation of material is more needed.

5. Limitations and Further Research

Some concerns have been inevitably found in the present study. To begin with, language skills included in the study are sorely listening and reading skills. The other two skills (i.e., speaking and writing skills) are, therefore, recommended for further research. Additionally, the study is relatively large in sample size but fails to provide participants' in-depth understandings. Further research may consider interviewing several typical participants about their perceptions of the test. A mixed-methods approach (including both quantitative and qualitative methods) could hence be adopted to achieve better research outcomes. Lastly, the participants recruited from Mainland China are all from metropolitan cities (e.g., Shanghai and Beijing). Students from those cities may have more access to language learning, and the educational standard and level may also be relatively high. Therefore, the fairness issues might cause due to the selection of cities and participants. Further research is suggested to include a broader population.

6. Conclusion

The research on young language learners has re-drawn the attention in the field of language assessment. The GEPT-Kids as an increasingly prevalent test was then investigated in terms of its reliability and construct validity. The result showed quite high reliability coefficients of test items, revealing that the test is reliable overall and the test items measure the same construct. In terms of the construct validity, the GEPT-Kids assesses only the general construct but fails to distinguish the construct of subskills, i.e., listening and reading skills. The current study then suggests that the GEPT-Kids is suitable for assessing young learners' overall language ability. Nevertheless, when taking sub-skills into account, the test should be used with caveats in the classroom teaching-learning context because it may not be able to recognize test takers' listening or reading skills respectively. Also, since most of the participants are from mainland China, Hong Kong, and Macau, they got a high grade in GEPT-Kids. This implies that those participants are also adapted to the test that is specifically designed for pupils from Taiwan. The LTTC, therefore, may consider promoting the GEPT-Kids test to students from mainland China, Hong Kong, or Macau after slightly revising the test to achieve international recognition (Liao, 2021).

Acknowledgements

We sincerely thank Language Training and Testing Center (LTTC, Taiwan) for allowing us to use their data for this study.

References

Alderson, J. C., & Bachman, L. F. (2000). *Cambridge language assessment series*. Cambridge University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bailey, A. L. (2008). Assessing the language of young learners. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 379-398). SpringerLink. https://doi.org/10.1007/978-0-387-30424-3_188

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* Pearson/Longman.

Brown, J. D. (1996). *Testing in language programs*. Prentice-Hall.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-31. https://doi.org/10.1177/0265532209346370

Elder, C., & Zammit, S. (1992). Assessing performance in languages other than English: The contribution of the Australian Language Certificates. *The journal of the Australian Advisory Council on Languages and*

Multicultural Education, 6, 14-20.

Hasselgreen, A., & Caudwell, G. (2016). *Assessing the language of young learners*. Equinox Publishing.

Heubeck, B., & Neill, J. (2000). Confirmatory factor analysis and reliability of the mental health inventory for Australian adolescents. *Psychological Reports, 87*, 431-440. https://doi.org/10.2466/pr0.2000.87.2.431

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kunnan, A. J. (2012). Statistical analyses for test fairness. *French Journal of Applied Linguistics, 15*, 39-48. https://doi.org/10.3917/rfla.151.0039

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge. https://doi.org/10.4324/9780203803554

Kunnan, A. J., & Liao, L. (2019). Modeling relationships among young learners' self-assessment, learning attitude, and language test performance. *Journal of Asia TEFL, 16*(2), 701-710. https://doi.org/10.18823/asiatefl.2019.16.2.18.701

Kunnan, A. J., & Wu, J. (2017). *A survey of the English language proficiency of young learners in Chinese-speaking cities*. Unpublished paper submitted to Language Training and Testing Center (LTTC), Taipei, Taiwan.

Lee, K. Q. (2015a). *GEPT-Kids model test*. Learning Publishing.

Lee, K. Q. (2015b). *GEPT-Kids reading comprehension test*. Learning Publishing.

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in GEPT-Kids listening. *Frontiers in Psychology, 12*, 1-9. https://doi.org/10.3389/fpsyg.2021.767244

Liao, R. (2021). Advancing the international recognition of the locally-produced GEPT: And interview with Jessica Wu. *Language Assessment Quarterly*, 1-10. https://doi.org/10.1080/15434303.2021.1919116

Liao, Y. F. (2009). *A construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge*. ProQuest. Retrieved from https://search.proquest.com/docview/304869181?pq-origsite=gscholar

McKay, P. (2006). *Assessing young language learners*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733093

Moore, T., Morton, J., & Price, S. (2007). Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study. In L. Taylor, & C. Weir (Eds., Vol. 34). *IELTS Collected Papers: Research in reading and listening assessment* (pp.1-86). Cambridge University Press. https://www.ielts.org/-/media/research-reports/ielts_rr_volume11_report4.ashx

Papp, S., & Rixon, S. (2018). Examining young learners: Research and practice in assessing the English of school-age learners. In N. Saville, & C. J. Weir (Eds., Vol. 47), *Studies in language testing*. Cambridge University Press.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2016). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5-30. https://doi.org/10.1002/j.2333-8504.2008.tb02095.x

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170. https://doi.org/10.1177/0265532209349465

Yao, D. (2021). Automated writing evaluation for ESL learners: A case study of *Pigai* system. *The Journal of Asia TEFL, 18*(3), 949-958. https://doi.org/10.18823/asiatefl.2021.18.3.14.949

Copyrights