Human Nature Journal of Social Sciences
Vol.4, No.2 (June, 2023), Pp.223-231
ISSN(online): 2788-5240, ISSN(print):2788-5232

Human Nature Research Publisher

Original Article **http://hnpublisher.com**

Investigating the Reliability of Language Tests Using Classical Test Theory and Item Response Theory

Maham Masood[1], Iram Rubab[2], Rahma Shahid[3]

[1]BS Scholar, Department of English, GC Women University, Sialkot
[2]Assistant Professor, Department of English, GC Women University, Sialkot
[3]BS Scholar, Department of English, GC Women University, Sialkot
Correspondence: iram.sial@yahoo.com[2]

ABSTRACT

Aim of the Study: The study aims to investigate the concerns about the reliability of language tests. It uses both Classical Test Theory (CTT) and Item Response Theory (IRT) to examine the reliability of language tests with high-stake consequences.

Methodology: The study uses a mixed method approach using both qualitative and quantitative data. It conducts interviews with language experts and participants to examine the reliability of language tests. Furthermore, the study collects quantitative data through a survey with language professionals.

Findings & Conclusion: Using 15 questions, the value of Cronbach's Alpha was 0.939. It is above 0.70, which means it is good. This shows a great level of reliability. It means the language tests prepared were valuable and helpful for the participants. The tests of synonyms, reported speech, and analytical questions are a part of aptitude tests. CTT and IRT approaches were used, and which marks obtained from the participants helped to know the level of reliability.

Article History

Received:
March 11, 2023

Revised:
June 11, 2023

Accepted:
June 16, 2023

Published:
June 20, 2023

Keywords: Reliability; Language Tests; CTT; ITT; Mixed Method Approach.

Introduction

Language testing of individuals is a broad phenomenon that requires a complex approach to assess language proficiency using listening, reading, and speaking skills. Traditionally, the assessment method used for English proficiency was considered more time-consuming and insufficient to test the language proficiency of large data sets. For instance, the fixed-length test with pencils was used to measure the language proficiency of individuals who can assess language's limited characteristics. The items in the conventional paper-based test were suitable for the examinee's average ability. These tests also have restricted time on some sections that limit the response of individuals who respond slowly (Ozdemir & Gelbal, 2022). Therefore, advanced methods were developed to provide flexibility to both the examiner and the candidates. The computerised adaptive test is one of the advanced language proficiency assessment approaches that help examine language ability.

CAT (Computerised adaptive testing) method is conducted through a computer and provides flexibility to select the large test unit. The CAT test helps ensure the final score using as few items as possible that provide the accurate final score. Therefore, the CAT test is effective as it requires less assessment time, reduces test length, and provides each examinee with the same set of questions. Various CAT language proficiency tests, such as IELTS, TOEFL, Duo-lingo, Cambridge Assessment, and Oxford, are available for admission in institutes and employment purposes. However, there are concerns related to more reliable tests. Therefore, examining the effectiveness of English proficiency tests is crucial for tests with high-stake consequences. CTT (Classical test theory) and Item Response theory (IRT) is important in examining the quantify measurement errors.

CTT theory is based on a common estimation of measurement precision that explains that measurement is equal for all individuals regardless of their attribute levels. However, the IRT (Item response theory explains that measurement precision relies on the latent attribute level. This theory specification difference may result in the difference between IRT and CTT with the conclusion and statistical significance. Curi and Silvia (2019) claimed that CATs assessments based on the IRT (Item Response Theory) are important in providing comparable ability scores between different time tests and answering the different question items. Besides that, the development of computer technologies for language proficiency based on item response theory plays a major role in increasing the adaptability of the test, as per Lee et al. (2019). Ozdemir et al. (2022) also explained that the reliability of adaptive tests based on IRT (item response theory) must be assessed to explain the most reliable English proficiency test. Previous studies examined the reliability and validity of language proficiency CAT tests (LaFlair & Settles, 2019; Mizumoto et al., 2019). However, they should have focused on CTT and IRT-based items in the test. Besides that, they specifically focused on the Duolingo test (LaFlair & Settles, 2019). Therefore, this research aimed to investigate the reliability of language tests using both CTT and IRT theoretical approaches.

Literature Review

Language Testing

A test refers to a sample of the behaviour (Rahman, 2020). Thus, a language test will be a sample of the language behaviour. Language testing is a field of study in Applied Linguistics (Bachman and Palmer, 2022). Its major focus is assessing the first, second, or any other languages in the institute, using language at the workplace, and immigration, asylum contexts, or citizenship. Moreover, language testing refers to evaluating a person's proficiency level with the use of a specific language in an effective way. In addition, language tests work better when such tests are planned and developed for measuring explicit language skills. These include speaking, proficiency in writing, listening, comprehension, reading, and the skill of translating texts or interpreting spoken language.

Different Language Tests

Language is a complex thing to measure (Green, 2020). Subsequently, many kinds of language testing exist, and every type is measured using different skills for many reasons. For instance, one test may ask a person to read a whole passage loudly, whereas another test can be related to answering the questions of the passage. Following are some types of language tests,

Language Proficiency Tests

Proficiency testing is used to measure a person's skill level in a language independent of how a person has learned it (De Wilde et al., 2020). Whether someone has grown up speaking French or has taken lessons in adulthood, the proficiency test must score every individual similarly.

Aptitude Tests

An aptitude test is not used to measure the efficacy level of someone who speaks a particular language (Bokander and Bylund, 2020). However, this helps to know the ways that people used for acquiring language skills.

Diagnostic Tests

Proficiency tests generally give a general assessment of the complete language skillset of a person (Roth et al. 2019). Contrastingly, the diagnostic tests tend to identify particular strengths and weaknesses in the specific skillset.

Placement & Achievement Tests

The test is used entirely in the environments of language learning. Moreover, a placement test tends to measure the skill for grouping likewise skilled learners (Bachman and Palmer, 2022). Furthermore, an achievement test is used to measure a learner's progress over time.

Reliability of Language Tests

Reliability is based on how an individual measures it (Akhmedov, 2022). Moreover, the reliability of a test is about the constancy of scoring and the correctness of the management procedures of the test. Reliability is all about checking the consistency level in language assessment. For this, two theories play a relevant role.

Classical Test Theory

The classical test theory states that a score obtained in the measurement procedure is affected by two main things (Al Nima et al., 2020). These include the correct score of an object, person, or phenomenon being measured. The other refers to the error, which can be anything apart from the correct score of the idea of interest. For instance, when a candidate has completed an arithmetical reasoning test and attained a 15 out of 20, their “Observed score” will be 15. Though, there is no psychometric assessment that is 100 per cent reliable since error affects the result always. This means that the observed score of a candidate may differ from the “True score”.

Item Response Theory

IRT refers to item response theory, the latent response theory (Pliakos et al. 2019). It is meant to be a family of mathematical models which explains the link between the latent traits that are unobservable factors or trait, as well as their manifestations that involves the experiential outcomes, responses, or performance (Yeung, 2019). People with lower ability may have a poor level of chance, whereas those with a higher level of ability are more likely to answer correctly; for instance, the students with better math ability are quite probable to do a math item accurately.

Theoretical Framework

The theory of language, also called Generative Grammar and Dell Hymes’ Communicative Competence theory, is considered for this study. The concepts of competence and performance are needed to understand language testing or its overall assessment. Considering Generative Grammar theory, competence refers to the capacity to generate immeasurable sentences from a particular set of grammatical rules (Aprianto and Zaini, 2019). Moreover, the view postulates that competence rationally precedes performance and is based on generative for enhanced learning. On the other hand, Dell Hymes has given contradictory views about the perceived inadequacy of the Generative Grammar theory (Abdulrahman and Ayyash, 2019). He stated that the communicative aspect of the language tends to supersede the linguistic aspect of language.

Moreover, communicative competence, thus, is the language knowledge of the user of syntax, phonology, and social knowledge regarding the ways of using the utterances suitably. Thus, language tests are done to test the linguistic as well as communicative competence of students. This enables them to function correctly in situations where English usage is needed.

Methodology

The research is based on a deductive approach, in which the researchers deduced the main concept of checking the reliability of language testing through CTT and IRT. This is a primary mixed study in which both qualitative and quantitative data are used. A language test is selected that has high-stakes consequences as well as a sufficient number of participants (Munn et al. 2020). The researcher administered the selected language test to a sample of participants. The sample population includes professionals in language assessments from University of Sialkot and University of Management and Technology Sialkot (UMT) who are asked to share their views about the questionnaire provided. Quantitative data is obtained through a survey, whereas qualitative data is collected through participant interviews.

A questionnaire includes an academic reading section, keyword transformation, and report speech. These were asked to be critically reviewed by the participants of the study and share their views on it (Chen and Song, 2019). Their perceptions helped in finding the reliability of the tests as they are language professionals who have been in the field of language test assessment for years. Some professionals were asked to spare some time and give detailed views. Thus, the ones from whom a survey was conducted were 100 professionals. However, interviews were conducted with 5 participants.

All the test data is analysed using CTT and IRT to examine the reliability of the test. Later, the researcher compared the CTT and IRT analysis results of the chosen language test. Furthermore, based on the data obtained, recommendations are given to improve the reliability of language tests (Guest et al., 2020). The data gathered was analysed using a reliability test and thematic analysis. The reliability test helped in reaching a conclusion based on checking the reliability level of the language test that the participants conducted. On the other hand, thematic analysis was done while making different themes so that a clear idea could be depicted, considering the relevancy of the questionnaire conducted.

There were some ethical issues found while conducting the research. Professionals highly criticised the questionnaire, which was not very pleasant for the researcher. A few considered it a waste of time, considering that the researcher needed help to meet the level of actual Language testing assessment (Facca et al. 2020). Some were not ready to participate in this study, considering it might take much time. The researcher needed more help forming a questionnaire from his peer group. They were not interested in such studies. However, despite all these challenges, the researcher tried to conduct this study and prove the key findings.

Findings and Analysis

Table 1: *Scale: All Variables*

		N	**%**
Cases	Valid	100	100.0
	Excluded[a]	0	.0
	Total	100	100.0

Table 2: *Reliability Statistics*

Cronbach's Alpha	**N of items**
.939	15

A reliability test was run to check the reliability of the language tests. Cronbach's alpha measures internal consistency (Hayes and Coutts, 2020). It concludes how to find the close retable a set of items in a group. Moreover, it is known to be a reliability measure of scale. Attaining a high value for the alpha does not infer that the measure is not dimensional. Using 15 questions, the value of Cronbach's Alpha was 0.939. It is above 0.70, which means it is good. The Cronbach's alpha coefficient, close to 1.0, tends to portray a greater level of the internal consistency of the key items in a scale. Similar is the case with this study which shows great reliability. It means the language tests prepared were valuable and helpful for the participants (Baik et al. 2019). They had attempted them all and significantly found them to be relevant. Three tests of this study include synonyms, reported speech, and analytical questions. All these three were found to help assess the skills and competence of the participants.

Table 3: *Correlations*

		Synonym 1	**Synonym 2**	**Synonym 3**	**Synonym 4**	**Synonym 5**	**Synonym 6**	**Reported Speech 1**	**Reported Speech 2**
Synonym 1	Pearson Correlation	1	.490**	.315**	.717**	.847**	1.000**	.258**	.477**
	Sig. (2-tailed)		.000	.001	.000	.000	.000	.010	.000
	N	100	100	100	100	100	100	100	100
Synonym 2	Pearson Correlation	.490**	1	.634**	.381**	.588**	.490**	.454**	.488**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
	N	100	100	100	100	100	100	100	100
Synonym 3	Pearson Correlation	.315**	.634**	1	.231*	.405**	.315**	.791**	.413**
	Sig. (2-tailed)	.001	.000		.021	.000	.001	.000	.000
	N	100	100	100	100	100	100	100	100
Synonym 4	Pearson Correlation	.717**	.381**	.231*	1	.703**	.717**	.144	.361**
	Sig. (2-tailed)	.000	.000	.021		.000	.000	.152	.000
	N	100	100	100	100	100	100	100	100
Synonym 5	Pearson Correlation	.847**	.588**	.405**	.703**	1	.847**	.269**	.454**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.007	.000
	N	100	100	100	100	100	100	100	100
Synonym 6	Pearson Correlation	1.000**	.490**	.315**	.717**	.847**	1	.258**	.477**
	Sig. (2-tailed)	.000	.000	.001	.000	.000		.010	.000
	N	100	100	100	100	100	100	100	100
Reported Speech 1	Pearson Correlation	.258**	.454**	.791**	.144	.269**	.258**	1	.578**
	Sig. (2-tailed)	.010	.000	.000	.152	.007	.010		.000
	N	100	100	100	100	100	100	100	100
Reported Speech 2	Pearson Correlation	.477**	.488**	.413**	.361**	.454**	.477**	.578**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	
	N	100	100	100	100	100	100	100	100
Reported Speech 3	Pearson Correlation	.258**	.454**	.791**	.144	.269**	.258**	1.000**	.578**
	Sig. (2-tailed)	.010	.000	.000	.152	.007	.010	.000	.000
	N	100	100	100	100	100	100	100	100
Reported Speech 4	Pearson Correlation	.843**	.341**	.176	.564**	.675**	.843**	.296**	.496**
	Sig. (2-tailed)	.000	.001	.080	.000	.000	.000	.003	.000
	N	100	100	100	100	100	100	100	100

Reported Speech 5	Pearson Correlation	.379**	.731**	.372**	.239*	.392**	.379**	.568**	.587**
	Sig. (2-tailed)	.000	.000	.000	.016	.000	.000	.000	.000
	N	100	100	100	100	100	100	100	100
Analytical questions 1	Pearson Correlation	.258**	.454**	.791**	.144	.269**	.258**	1.000**	.578**
	Sig. (2-tailed)	.010	.000	.000	.152	.007	.010	.000	.000
	N	100	100	100	100	100	100	100	100
Analytical questions 2	Pearson Correlation	.477**	.488**	.413**	.361**	.454**	.477**	.578**	1.000**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
	N	100	100	100	100	100	100	100	100
Analytical questions 3	Pearson Correlation	.258**	.454**	.791**	.144	.269**	.258**	1.000**	.578**
	Sig. (2-tailed)	.010	.000	.000	.152	.007	.010	.000	.000
	N	100	100	100	100	100	100	100	100
Analytical questions 4	Pearson Correlation	.687**	.359**	.173	.518**	.770**	.687**	.261**	.445**
	Sig. (2-tailed)	.000	.000	.085	.000	.000	.000	.009	.000
	N	100	100	100	100	100	100	100	100

Correlation Analysis is a statistical method used to discover whether there is a link between variables (Senthilnathan, 2019). Moreover, it is important to find the strength of the relationship. It is a statistical technique to determine the possible linear connections among the variables. Considering the table above, there is a greater significance level for all questions of language testing. The standard test for synonymy refers to substitution (Taufiq et al. 2021). When an expression is replaced by another in a sentence while not changing the meaning of a sentence, those two expressions are called synonyms (Ngafif et al., 2022). Reported speech includes direct and indirect forms of verbs. Analytical questions are done after taking answers from a given passage. All these are a part of aptitude tests (Abd Gani et al. 2020). Thus, the researcher conducted aptitude tests on 100 people who considered this questionnaire helpful and reliable.

The researcher also conducted interviews to get insights on the topic. One of the participants stated,

> *You see, this questionnaire is useful to prepare oneself for any English language test. I have conducted many tests and it meets their level of standard. Moreover, I found it easy and less time consuming because answers are based on the options. Instead of writing long paragraphs, it was easy to fill them with all my best possible understanding. I wish you could share my results with me!*

These comments made the researcher satisfied that the questionnaire constructed was highly effective in measuring the skillset of the person (Sürücü and MASLAKÇI, 2020). The words of the participants assured the researcher that it was a highly reliable test.

Another participant had stated,

> *I did not get the idea of mixing all these tests. You could make it in an orderly way. I mean that analytical questions had to be in the beginning. However, despite the structural issues, these questions are strong enough to prepare oneself for a language test. I am quite satisfied with the content but not the structure. Your focus of the study is to find the reliability of Language tests that you have successfully attained. You used CTT and IRT approaches for it as per which, marks obtained from the participants had helped you in knowing the level of reliability.*

The researcher had used IRT and found that participants who scored high had better language skills.

Discussion

The highly common approaches used through psychometrics include the classical test theory (CTT) and item response theory (IRT). These are highly effective measuring instruments (Santhanadass et al. 2021). Moreover, CTT is used for relative simplicity and the lower skill level needed for analysis. Many researchers have used CTT to prove its significance. One of the examples is that it was used in the medical field to validate its need and reliability (Ngafif et al., 2022). It was found to be positive. CTT is used for analyzing the performance of a group of students on an assessment instrument. When another group of students tends to take that assessment instrument, comparisons, thus, cannot be done.

CTT undertakes that all of the items in the assessment instrument are useful for making an equal contribution to the overall performance of the students (Curi and Silvia, 2019). Contrastingly, IRT can be considered as some items are highly difficult rather than comparing them to others (Ozdemir et al., 2022). It means that the likelihood of success on the items is outstanding both to the students' ability and even to the item's difficulty. This study had not found any of the participants to face difficulty. Instead, all have appreciated the provided language tests.

Conclusion

This research paper has examined the effectiveness of English proficiency tests as crucial for tests with high-stake consequences. CTT (Classical test theory) and Item Response theory (IRT) is important in examining the quantify measurement errors. This is a primary mixed study in which both qualitative and quantitative data are used.

Using 15 questions, the value of Cronbach’s Alpha was 0.939. It is above 0.70, which means it is good. This shows a great level of reliability. It means the language tests prepared were valuable and helpful for the participants. Considering the table given of correlation analysis, it was found that there is a greater level of significance for all questions of language testing. The tests of synonyms, reported speech, and analytical questions are a part of aptitude tests. Thus, the researcher conducted aptitude tests on 100 people who considered this questionnaire helpful and reliable. Interviews revealed that the questionnaire was useful in preparing for any English language test. It was easy and less time-consuming as the answers were based on the options. The words of the participants assured the researcher that it was a highly reliable test. However, one of the participants found some structural issues in the questionnaire. However, the questions were strong enough to prepare oneself for a language test. CTT and IRT approaches were used, and which marks obtained from the participants helped to know the level of reliability.

Acknowledgments

None.

Conflict of Interest

Authors declared no conflict of interest.

Funding Source

The authors received no funding to conduct this study.

ORCID iDs

Maham Masood [1] https://orcid.org/ 0000-0002-4157-5105
Iram Rubab [2] https://orcid.org/0000-0002-1523-7065
Rahma Shahid [3] https://orcid.org/ 0000-0001-8788-0095

References

Abd Gani, N.I., Rathakrishnan, M. and Krishnasamy, H.N., (2020). A pilot test for establishing validity and reliability of qualitative interview in the blended learning English proficiency course. *J Crit Rev*, *7*(5), 140-143.

Abdulrahman, N.C. and Ayyash, E.A.S.A., (2019). Linguistic competence, communicative competence and interactional competence. *Journal of Advances in Linguistics*, *10*(1), 1600-1616.

Akhmedov, B.A. (2022). Analysis of the Reliability of the Test form of Knowledge Control in Cluster Education. *Psychology and Education*, *59*(2), 403-418.

Al Nima, A., Cloninger, K.M., Lucchese, F., Sikström, S. and Garcia, D. (2020). Validation of a general subjective well-being factor using Classical Test Theory. *PeerJ*, *8*, p.e9193.

Aprianto, D. and Zaini, N. (2019). The principles of language learning and teaching in communication skill developments. *VELES: Voices of English Language Education Society*, *3*(1).

Bachman, L. and Palmer, A. (2022). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baik, S.H., Fox, R.S., Mills, S.D., Roesch, S.C., Sadler, G.R., Klonoff, E.A. and Malcarne, V.L. (2019). Reliability and validity of the Perceived Stress Scale-10 in Hispanic Americans with English or Spanish language preference. *Journal of health psychology*, *24*(5), 628-639.

Bokander, L. and Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language learning*, *70*(1), 11-47.

Chen, C. and Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PloS one*, *14*(10), p.e0223994.

Cúri, M. and Silva, V. (2019). Academic English proficiency assessment using a computerized adaptive test. *TEMA (São Carlos)*, *20*, 381-401.

De Wilde, V., Brysbaert, M. and Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important?. *Bilingualism: Language and Cognition*, *23*(1), 171-185.

Facca, D., Smith, M.J., Shelley, J., Lizotte, D. and Donelle, L. (2020). Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. *Plos one*, *15*(8), p.e0237875.

Green, A. (2020). *Exploring language assessment and testing: Language in action*. Routledge.

Guest, G., Namey, E. and Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PloS one*, *15*(5), p.e0232076.

Hayes, A.F. and Coutts, J.J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1-24.

LaFlair, G.T. and Settles, B. (2019). Duolingo English test: Technical manual. *Retrieved April*, *28*, p.2020.

Lee, Y.L., Lin, K.C. and Chien, T.W. (2019). Application of a multidimensional computerized adaptive test for a Clinical Dementia Rating Scale through computer-aided techniques. *Annals of General Psychiatry*, *18*(1), 1-9.

Mizumoto, A., Sasao, Y. and Webb, S.A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101-123.

Mizumoto, A., Sasao, Y. and Webb, S.A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101-123.

Munn, Z., Barker, T.H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M. and Aromataris, E. (2020). Methodological quality of case series studies: an introduction to the JBI critical appraisal tool. *JBI evidence synthesis*, *18*(10), 2127-2133.

Ngafif, A., Sukarni, S., Nugraeni, I.I., Sahidah, N., Pithaloka, K.D. and Ningsih, S.C. (2022). Factors affecting the reliability of senior high schools' english teacher-made test in Kebumen. *Jurnal Pendidikan Surya Edukasi (JPSE)*, *8*(2), 162-177.

Ngafif, A., Sukarni, S., Nugraeni, I.I., Sahidah, N., Pithaloka, K.D. and Ningsih, S.C. (2022). Factors affecting the reliability of senior high schools' english teacher-made test in Kebumen. *Jurnal Pendidikan Surya Edukasi (JPSE)*, *8*(2), 162-177.

Ozdemir, B. and Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. Education and Information Technologies, 27(5), pp.6273-6294.

Pliakos, K., Joo, S.H., Park, J.Y., Cornillie, F., Vens, C. and Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, *137*, pp.91-103.

Rahman, M.S. (2020). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review**.**

Roth, D., Pace, N.L., Lee, A., Hovhannisyan, K., Warenits, A.M., Arrich, J. and Herkner, H. (2019). Bedside tests for predicting difficult airways: an abridged Cochrane diagnostic test accuracy systematic review. *Anaesthesia*, *74*(7), 915-928.

Santhanadass, A.R., Elumalai, G., Prasetyo, Y. and Zakaria, J. (2021). Devising and Testing Revised Validity and Reliability of Strategic Knowledge, Efficient Behaviour, and Affective Value in Outdoor Evaluation Questionnaire. In *ISMINA 2021: Proceedings of the 5th International Conference on Sports, Health, and Physical Education, ISMINA 2021, 28-29 April 2021, Semarang, Central Java, Indonesia* (p. 340). European Alliance for Innovation.

Sürücü, L. and MASLAKÇI, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), pp.2694-2726.

Taufiq, M., Putri, R., Fendi, H., Syafrilianto, S., Anggraini, D. and Indriyani, V. (2021). Validity and Reliability of Semester Tests Made by Teachers: An Evaluation Study of English Learning. In *Proceedings of the 2nd EAI Bukittinggi International Conference on Education, BICED 2020, 14 September, 2020, Bukititinggi, West Sumatera, Indonesia*.

Yeung, C.K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*.