JPAIR Multidisciplinary Research Volume 58 • October 2024
DOI: https://doi.org/10.7719/jpair.v58i1.913

Analysis of Model Fit and Test Information on Otis-Lennon School Ability Test Using Item Response Theory

REY-MARK G. BASAGRE[1], **RANIER I. CERNA**[2]
[1]Central Bicol State University of Agriculture, San Jose, Pili,
Camarines Sur, Philippines
[2]School of Saint Anthony, Lagro, Quezon City, Philippines

Corresponding author: reymark.basagre@cbsua.edu.ph

Originality 100% • Grammar Check: 98% • Plagiarism: 0%

ABSTRACT

Article History

Received: 06 Mar 2024
Revised: 15 Jul 2024
Accepted: 23 Aug 2024
Published: 31 Oct 2024

Keywords— Test Assessment, Model Fit, Test Information, Item Response Theory, Camarines Sur, Philippines

The Otis-Lennon School Ability Test (OLSAT) is widely used in different academic institutions for various purposes, such as admission tests, performance predictions, and intelligence tests. However, there has yet to be a study on its model fit, item evaluation, and overall test information. This study examined the OLSAT item and the overall test information used by a state university in the admission process through the lens of item response theory utilizing R Studio. Specifically, it seeks to determine the model fit of the data from the university's standardized entrance test and to examine the test information for the standardized entrance test. Comparing the nested models suggests that the more complex model fits the data better (p = 0.000), which is the 2 Parameter Logistic Model. Some items' p-values suggest local dependence, but it can be tolerated upon examination of some of the actual items. The index of difficulty ranges from -51.33198812 to 11.61952106, while the discrimination index ranges from -0.04759777 to 1.16897332 suggesting that some items have a very low discrimination index and some have a high difficulty index. In conclusion, the test provides more information about the middle portion of the ability scale, which is suitable for admission purposes of the test.

INTRODUCTION

Higher learning institutions worldwide are flooded by freshmen applicants who wish to enter the learning institution. With the limited capacity in terms of space, faculty, and funding, schools resorted to an admission policy that selects applicants based on founded criteria that measure the conceptual understanding of students (Basagre, 2023). With a large number of freshmen applicants during the admission process and considering its limited capacity, the university employed standardized tests to select the most qualified applicants to enter the university such as a graduate management admission test analyzed using item response theory (Kingston, 1985). The Otis-Lennon School Ability Test (OLSAT) is one of the standardized tests used for admission examinations and other purposes. The OLSAT is a multiple-choice K-2 assessment that tests reasoning skills using a variety of verbal, nonverbal, figurative, and quantitative reasoning questions (Otis, 1989). It is intended to evaluate a child's performance across multiple reasoning skill sets.

In the Philippines, most universities are employing stricter admission processes (Dominguez et al., 2023) to get the most qualified applicants out of the pool of enrollees due to free higher tertiary education (Guidang, 2016). State universities and colleges are forced to conduct stricter admission tests to balance adherence to the law and meet quality assurance requirements (Ordonez & Ordonez, 2009). The Otis-Lennon School Ability Test is the most sought-after test for selecting the most qualified students for university admissions. Furthermore, the OLSAT is also used by educators for other academic purposes. In a study of validity by Sapp et al. (2016), the Otis-Lennon School Ability Test was validated by comparing scores with those of the Wechsler Intelligence Scale of the Children-Revised using computation of correlations, t-tests, and regression equations on an overall group of 60 first-grade pupils was divided by ethnic group membership. The Otis-Lennon and the WISC-R IQ showed strong positive relationships, similar to the findings of Karrh (2009), between Reading and Math, and both the Stanford 10 achievement test and the Otis-Lennon School Ability Test. Comparisons by ethnic group membership revealed that the Otis-Lennon was equal for Black and White children, and the differences between the means were less than those reported for the national sample. Medallon and Cataquis (2011) conducted a study to verify if the OLSAT could predict the performance of freshmen students. Their correlation analysis revealed a significant direct correlation between the OLSAT and the final grades in English and Mathematics, between the total raw scores in the OLSAT and the GPA of the students, and between the verbal scores and the final grades in English, Mathematics, and the GPA.

Most of the studies on the Otis-Lennon School Ability Test focus on the determination of its validity by comparing it to other tests such as Peabody Individual Achievement Test (PIAT), and the Metropolitan Achievement Test (Davenport, 1976), Wechsler Intelligence Scale for Children-Revised (Sapp & Marshall, 2016), Wechsler Intelligence Scale for Children-Third Edition (Guilmette et al., 2011). Other research focuses on the predictive validity of OLSAT on different performances, such as the scholastic performance of Cebu Doctors' University physical therapy students (Borromeo et al., 2007), intelligence quotient (Avant & O'Neal, 1986), the performance of the first-year students, and achievement in Grades 2 and 4 (Antonak et al., 2014). The test is now widely used in different academic institutions for varied purposes. However, there has yet to be a study about the Otis-Lennon School Ability Test (OLSAT) model fits and items evaluation. Thus, this study focuses on analyzing the model fits and item information of the test.

FRAMEWORK

The Item Response Theory (IRT) by Orlando and Thissen (2001) was the only framework for this study. The IRT is a statistical framework used to model the relationship between respondents' performance on test items and their latent traits, such as ability, personality, and attitude. In test results, IRT acknowledges that different items may vary in difficulty and discriminatory power. It provides models to ensure test quality, reliability, and fairness in all stages, from test development to test utilization. In this study, the item response theory is valuable because it provides ways to assess model fit by comparing nested models, and it examines the test information leading to assessing the item's local dependence.

OBJECTIVES OF THE STUDY

This study aimed to analyze the Otis-Lennon School Ability Test (OLSAT) through the lens of item response theory. Specifically, it sought to find answers to the following research objectives: (1) determine the model fit evaluation from the test results of the Otis-Lennon School Ability Test (OLSAT), and (2) examine the test information provided by the test results of the Otis-Lennon School Ability Test.

METHODOLOGY

Research Design

This study is basic research since it tries to discover the model that best fits the data by experimenting with several Item Response Theory (IRT) models. It contributes to the generation of information required for test development theories. This study adopted a secondary data analysis, a type of research that employs data acquired by someone else for another purpose as research methodology (Johnston, 2014).

Participants

The participants of this study are freshmen from a state university in the Bicol region, Philippines. The data was gathered from over 1000 entrance examiners of the student's admission office from the admission year 2019, a year before the spread of COVID-19.

Data Gathering Procedure

A letter requesting access to the test responses from the admission test results of a state university in the Bicol region was addressed to the university president through the admission director and was then approved. The Otis-Lennon School Ability Test, administered as an admission test, is a multiple-choice test with 72 items and five options for each item.

From the responses of over 1000 entrance examiners of the student's admission office from the admission year 2019, the test papers were checked and coded for analysis as 1 for correct answers and 0 for incorrect answers. A nested model evaluation was carried out to assess the model fit. The test responses will be the basis for this study's secondary data analysis.

Statistical Analysis

For this study, the assessment of model fit was done by comparing the Rasch, 1PL, 2PL, 3PL, and 4PL using the MIRT package and the ANOVA function in the R studio, together with the analysis of the test information.

RESULTS AND DISCUSSION

Model Fit Evaluation of the Data from the Standardized Entrance Test

Models are nested when the simpler or reduced model can be expressed as a special case of the more complex or full model. Nested models are compared by comparing the log-likelihoods of the different models. Specifically, we examine whether the fit significantly worsens as a simpler model is used in place of a more complex model. For this study, we try to perform a pairwise comparison of the models such as the Rasch, 1 PL, 2PL, 3PL, and 4PL, but the 3PL does not converge after 10,000 cycles. We try to increase the number of cycles gradually from 20,000 to 30,000 until we reach 60,000 cycles, but still, it does not converge. Not converging up to this level means that the data that was derived might not truly represent the meaning of the data being analyzed. The 4PL (4 Parameter Logistic Model) converges after 50,000 cycles, but as we check the indexes, some values are so high, such as an item with a difficulty index above 500. This result is malicious because, in the actual evaluation of the items, more than 300 over 1000 are able to answer the item correctly, which suggests that the item is answerable and not very difficult.

For the reason mentioned above, we settle for the 2 PL (2 Parameter Logistic Model) as the best model to be used. The table below shows the statistics for the analysis of model fit.

Table 1
Model Fit Comparison using R Studio

	AIC	SABIC	HQ	BIC	logLik	X2	df	P
mod.rasch	82708.78	82835.19	82844.94	83067.04	-41281.39			
mod.1pl	2708.78	82835.19	82844.95	83067.04	-41281.39	-0.002	0	NaN
mod.2pl	1900.83	82150.19	82169.43	82607.54	-40806.41	949.951	71	0

In this table, the null hypothesis (Ho) is that the simpler model fits the data well, and the alternative hypothesis (Ha) is that the more complex model fits the data better. As a rule for the decision, we reject the null hypothesis (Ho) if the p-value is < 0.05. As shown, the three models (Rasch, 1PL, 2PL) are being compared, the Rasch model being the simpler model and the 2 PL as the more complex model. Since the derived p-value is 0.000, which is less than 0.05, we reject the null hypothesis (Ho) in favor of the alternative hypothesis (Ha). Hence, the more complex model, the 2 Parameter Logistic Model, fits the data better.

The 2-parameter Logistic Model was then used to find the local independence of the items in the test. Local independence can artificially inflate the item discrimination, hence the test information (Edwards et al., 2018; Rosenbaum, 1984; Dirlik, 2019). Violation of the local independence assumption can result in a distortion of the item, person, and test parameter estimates (DeMars, 2006; Sireci et al., 1991). Locally dependent items do not make unique contributions to the construct. As Zenisky et al. (2002) suggest, these items do not increase construct representation and exacerbate any construct-irrelevant factors that may be associated with an item, such as prior familiarity with the item context.

Chen and Thissen (1997) found G2 to be reasonable for detecting LD compared with several other LD measures in terms of power and Type I error rate. The table below shows the Local independence analysis using the G2 (Chen & Thissen, 1997).

Table 2

Sample Data Matrix for the Analysis of Local Independence

	Item.26	Item.27	Item.28	Item.29	Item.30	Item.31	Item.32	Item.33	Item.34	Item.35	Item.36	Item.37
Item.1	0.365	0.164	0.772	0.854	0.639	0.009	0.510	0.134	0.178	0.292	0.127	0.626
Item.2	0.223	0.258	0.637	0.883	0.727	0.098	0.074	0.164	0.194	0.336	0.27	0.233
Item.3	0.231	0.667	0.446	0.495	0.774	0.373	0.099	0.276	0.283	0.712	0.659	0.932
Item.4	0.510	0.762	0.406	0.358	0.204	0.768	0.000	0.629	0.138	0.188	0.033	0.333
Item.5	0.468	0.304	0.485	0.016	0.352	0.697	0.757	0.809	0.584	0.806	0.634	0.140
Item.6	0.018	0.248	0.384	0.158	0.697	0.374	0.699	0.604	0.858	0.287	0.716	0.772
Item.7	0.065	0.828	0.794	0.782	0.684	0.804	0.392	0.005	0.168	0.296	0.129	0.387
Item.8	0.276	0.738	0.040	0.097	0.032	0.168	0.876	0.455	0.355	0.350	0.180	0.770
Item.9	0.525	0.046	0.285	0.701	0.511	0.688	0.704	0.728	0.131	0.447	0.561	0.528
Item.10	0.355	0.847	0.467	0.440	0.009	0.133	0.715	0.025	0.149	0.814	0.100	0.424
Item.11	0.690	0.003	0.543	0.487	0.520	0.404	0.114	0.922	0.272	0.053	0.672	0.336
Item.12	0.503	0.091	0.569	0.310	0.117	0.012	0.717	0.188	0.477	0.400	0.205	0.207
Item.13	0.686	0.925	0.031	0.291	0.529	0.183	0.000	0.941	0.908	0.420	0.536	0.247

In this table, the null hypothesis is that items i and j (where i and j represent the items being compared) are locally independent, and the alternative hypothesis is that items i and j are locally dependent. If the p-value is less than 0.05, we reject the null hypothesis. Highlighted above are the items with a p-value below 0.05, which may be locally dependent and affect the test's total. These items do not necessarily mean that they are genuinely locally dependent items; hence, there is a need for further investigation. With the permission of the admission director, only two (2) items of the test were given due to the confidentially of the test, which is used for admission purposes.

For this academic purpose, items 8 and 28 have a p-value of 0.040, below 0.05, suggesting being locally dependent. However, in the actual visual examination of the items, the items are constructed in the same format but are not locally dependent. An examiner can correctly answer item 8 even without item 28 and vice versa. Hence, local dependent findings can be ignored, and the items can be retained.

We also look into the item fit of 72 items in the test using the item fit evaluation model according to Orlando & Thissen (2001). Here, the null hypothesis is that there is no item misfit, while the alternative hypothesis is that there is an item misfit. If the p-value is less than 0.05, we reject the null hypothesis. Below are the statistics from the evaluation.

Table 3

Sample Item Data from the Item Fit Evaluation

Item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2
Item.1	34.320	30	0.012	0.268
Item.2	40.041	35	0.012	0.256
Item.3	82.443	36	0.036	0.000
Item.4	53. 303	34	0.024	0.019
Item.5	46.669	31	0.022	0.035
Item.6	41.700	35	0.014	0.202
Item.7	38.721	34	0.012	0.265
Item.8	30.746	32	0.000	0.530
Item.9	27.286	28	0.000	0.503
Item.10	49.568	31	0.024	0.019
Item.11	41.166	34	0.015	0.186
Item.12	50.812	31	0.025	0.014
Item.13	65.580	36	0.029	0.002
Item.14	57.731	37	0.024	0.016
Item.15	39.729	34	0.013	0.230
Item.16	39.287	37	0.008	0.368
Item.17	46.327	35	0.018	0.095

Item.18	67.187	36	0.029	0.001
Item.19	35.815	31	0.012	0.253
Item.20	60.710	34	0.0282	0.003
Item.21	61.327	34	0.028	0.003
Item.22	35.829	29	0.015	0.179
Item.23	45.972	34	0.019	0.083
Item.24	41.905	34	0.015	0.165
Item.25	54.850	33	0.026	0.010
Item.26	29.685	32	0.000	0.584
Item.27	46.604	37	0.016	0.134
Item.28	68.336	37	0.029	0.001

Test Information of the test

In a test, knowing the information function can help us describe, select, and compare items or tests (Basagre, 2018; Mangubat, 2023). This information function tells us how much information we can get from the test. Depending on which area it provides greater information, we can decide on the test's purpose. This has been demonstrated but many empirical papers (Ning, 2017; Perera, et al., 2018; Chan et al., 2021). The table below shows the individual index of discrimination and difficulty parameter of the 72-item test. Here, we wanted an item to have a high value for the discrimination index (represented by a) and a moderately/middle/average value for the difficulty index (represented by b).

Table 4

Index of Discrimination and Difficulty of the Test Items

	a	B	g	U
Item.1	0.90431896	-0.95636020	0	1
Item.2	0.61547667	0.42082092	0	1
Item.3	0.18303605	4.76998512	0	1
Item.4	0.88884803	0.12654763	0	1
Item.5	0.97147662	-0.47214800	0	1
Item.6	0.70139778	0.87491963	0	1
Item.7	0.77581246	-0.09255358	0	1
Item.8	0.85157595	-0.72362335	0	1

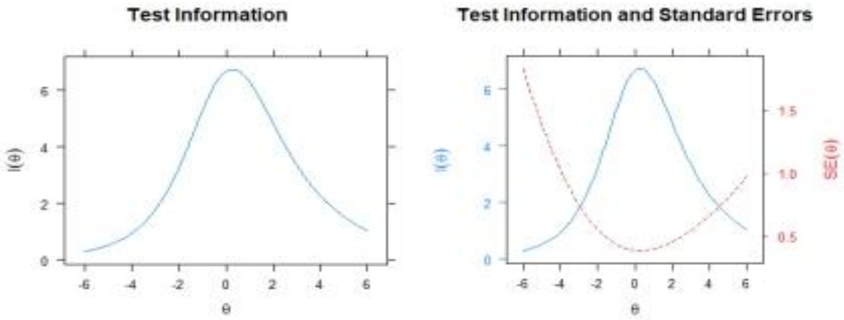
Item.9	1.16604643	-0.41502172	0	1
Item.10	0.80568109	-1.21583445	0	1
Item.11	0.69501206	1.52839155	0	1
Item.12	1.06162283	-0.03942185	0	1
Item.13	0.43183323	2.32270522	0	1
Item.14	0.38040074	0.88395125	0	1
Item.15	0.87449969	0.20357215	0	1
Item.16	0.16419159	5.75732153	0	1
Item.17	0.55858993	2.37813747	0	1
Item.18	0.50077576	1.51663970	0	1
Item.19	0.96622485	-0.17539247	0	1
Item.20	0.77592006	1.46062642	0	1
Item.21	0.71953807	1.47675536	0	1
Item.22	1.35562984	-0.15924888	0	1
Item.23	0.85794538	0.54406731	0	1
Item.24	0.73997031	0.15995106	0	1
Item.25	1.14332498	0.88788117	0	1
Item.26	1.02221039	0.14629582	0	1
Item.27	0.09855096	3.32481477	0	1
Item.28	0.07510279	6.24501484	0	1
Item.29	0.60814217	3.89682133	0	1
Item.30	1.01840713	-0.49733986	0	1
Item.31	0.76937386	0.06543455	0	1
Item.32	0.53694958	1.60690395	0	1
Item.33	0.40166530	1.45921734	0	1
Item.34	0.17227220	7.25965076	0	1
Item.35	0.33694744	2.72394736	0	1
Item.36	1.16897332	0.99447290	0	1
Item.37	0.50097469	1.03499160	0	1
Item.38	0.71652564	1.52079938	0	1
Item.39	0.79324480	1.83963000	0	1
Item.40	0.68973576	0.03068683	0	1

Item.41	0.60362888	1.26679222	0	1
Item.42	0.60772160	0.63029665	0	1
Item.43	0.31713379	0.15137610	0	1
Item.44	0.59129444	1.99117456	0	1
Item.45	0.41594005	1.93732331	0	1
Item.46	0.59857410	1.48448952	0	1
Item.47	0.61870942	2.45678745	0	1
Item.48	1.00464240	0.70886637	0	1
Item.49	-0.11734362	-6.60256243	0	1
Item.50	0.66470316	-0.15878095	0	1
Item.51	0.22539483	8.75971402	0	1
Item.52	0.18029997	3.45928801	0	1
Item.53	0.40548281	2.37426315	0	1
Item.54	0.91058125	0.35260888	0	1
Item.55	0.35808031	1.38018049	0	1
Item.56	0.29672671	8.74578089	0	1
Item.57	0.55032136	4.09313869	0	1
Item.58	0.26037825	4.09801061	0	1
Item.59	0.11487589	11.61952106	0	1
Item.60	0.53350676	2.14972832	0	1
Item.61	0.37336474	3.13508274	0	1
Item.62	0.39154387	5.07032193	0	1
Item.63	0.33850419	5.24268295	0	1

The index of difficulty ranges from -51.33198812 to 11.61952106, while the discrimination index ranges from -0.04759777 to 1.16897332. For this purpose, let us take a look at item 59. It has the highest difficulty value of 11.61952106. This item suggests that it is very difficult. This means that even the high-ability students have only approximately 40% of getting the item correctly. A good difficulty index is a difficulty level value ranging between -2 and +2 (Istiyono, 2017). Also, item 36 have the lowest discriminating index of -0.04759777. This suggests that it does not discriminate the students well, and the graph suggests that good and challenged students have approximately 15% of getting the answer correctly.

For this study, the test information was identified using graphs derived from R Studio. This data was then compared to other available data, such as the individual parameters (discrimination and difficulty parameters) derived from the 2-parameter Logistic Model. Below is the graph for the overall test information of the admission test used by the university.

Figure 1
Graph of the Test Information



As seen from the graph of the test information on the entrance test used by the university, it is a bell-shaped graph in the middle part where the peak is located at zero (0). This means that the test provides more information to middle-ability students, which is generally suitable for admission purposes. The standard error suggests the same.

CONCLUSIONS

Upon comparison of the nested model, the 2 Parameter Logistic Model is the best model that fits the data. Some items' p-values suggest local dependence, but upon examination of some of the actual items, it can be tolerated. Since only two (2) items were analyzed, other items marked as locally dependent should also be reviewed. There are item misfits noted upon objective examination of items using Orlando and Thissen's (2001) model, but it needs further examination. The proposed regression model can be used to predict the performance of students in taking the licensure examinations for electronics engineering graduates. An intensive review program on the academic courses must be adopted by the college since these are significant predictors of licensure performance. The curriculum must be continuously upgraded and strengthened.

The difficulty index of the items on the test ranges from -51.33198812 to 11.61952106, while the discrimination index ranges from -0.04759777 to 1.16897332. Some items need to be examined by looking at the actual items due to their very low discrimination index and high difficulty index. However, the test provides more information about the middle portion of the ability scale, which is suitable for test admission purposes.

TRANSLATIONAL RESEARCH

The findings of this study may be translated into a policy on using the Otis-Lennon School Ability Test (OLSAT) for university admission purposes where the test is determined to be best suited, especially for those without a standardized test for university admissions. Alongside this, a school or university policy to restrict the admission officers from using the Otis-Lennon School Ability Test (OLSAT) for purposes aside from admission could be another policy translation of this study. Furthermore, the results of this study could help other institutions to invest in developing their own well-crafted admission tests using item response theory by investigating its model fit and item information weather or not will serve the purpose of selecting qualified students in an admission test.

LITERATURE CITED

Antonak, R. F., King, S., & Lowy, J. J. (1982). Otis-Lennon Mental Ability Test, Stanford Achievement Test, and three demographic variables as predictors of achievement in grades 2 and 4. The Journal of Educational Research, 75(6), 366-373.

Avant, A. H., & O'neal, M. R. (1986). Investigation of the Otis-Lennon School Ability Test to Predict WISC-R Full Scale IQ for Referred Children.

Basagre, R. M. (2023, April). Effects of hands-on structured inquiry activities into students' conceptual understanding. In AIP Conference Proceedings (Vol. 2619, No. 1). AIP Publishing.

Basagre, R. M. G. (2018, July). Inquiry-Based Formative Assessment in Grade 10 Electricity and Magnetism. In *Ascendens Asia Journal of Multidisciplinary Research Conference Proceedings* (Vol. 2, No. 4).

Borromeo, M., Briones, M., Mahipos, J., Mamites, M., Namoc, G., Requillo, A., Tan, E., Tejero, Z., Yu, K., & Marquez, S. (2007). The predictive validity of Otis-Lenon school ability test on the scholastic performance of Cebu Doctors' University physical therapy students batch 2006 and 2007. Bachelor's thesis, Cebu Doctors' University, College of Rehabilitative Sciences, Department of Physical Therapy.

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213-236.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22(3), 265-289.

Davenport, B. M. (1976). A comparison of the peabody individual achievement test, the metropolitan achievement test, and the otis-lennon mental ability test. Psychology in the Schools, 13(3), 291-297.

DeMars, C. (2010). Item response theory. Oxford University Press.

Dirlik, E. M. (2019). The Comparison of Item Parameters Estimated From Parametric and Nonparametric Item Response Theory Models in Case of The Violance of Local Independence Assumption. International Journal of Progressive Education, 15(4), 229-240.

Dominguez, O. R. J., San, A. L. J. O., & Ferrer, S. F. M. (2023). Admission and retention requirements of the Bachelor of Science in Accountancy Program in Camarines Norte State College: basis for policy enhancement. DIU Journal of Business and Entrepreneurship, 16(01), 153-178.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. Psychological methods, 23(1), 138.

Guidang, E. P. (2016). Multiple Choice Test Randomizer of ICT Department of Abra State Institute of Science and Technology. *JPAIR Multidisciplinary Research*, *23*(1), 46-62.

Guilmette, T. J., Kennedy, M. L., & Queally, P. T. (2001). A comparison of the WISC-III and the Otis-Lennon School Ability Test with students referred for learning disabilities. Journal of Psychoeducational Assessment, 19(3), 239-244.

Istiyono, E. (2017). The analysis of senior high school students' physics HOTS in Bantul District measured using PhysReMChoTHOTS. Nucleation and Atmospheric Aerosols. https://doi.org/10.1063/1.4995184

Karrh, K. D. (2009). Predictors of student achievement in grade 7: The correlations between the Stanford Achievement Test, Otis-Lennon School Ability Test, and performance on the Texas Assessment of Knowledge and Skills (TAKS) math and reading tests. Liberty University.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the graduate management admission test 1. ETS Research Report Series, 1985(2), i-56.

Mangubat, A. (2023). Development of Diagnostic Test in Reading for Grade 7. *JPAIR Multidisciplinary Research*, *53*(1), 236-252.

Medallon, M. C., & Cataquis, R. E. (2011). Predictive Validity of the Otis-Lennon School Ability Test (OLSAT) to the First Semester Performance of Incoming Students at Lyceum of the Philippines–Laguna. Lyceum of the Philippines–Laguna Research Journal, 1(1), 1-1.

Ning, H. K. (2017). A psychometric evaluation of the achievement goal questionnaire–revised in Singapore secondary students. *Journal of Psychoeducational Assessment*, *35*(4), 424-436.

Ordonez, V., & Ordonez, R. M. (2009). Accreditation in the Philippines: A case study. In Higher Education in Asia/Pacific: Quality and the public good (pp. 201-215). New York: Palgrave Macmillan US.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement, 27(4), 289-298.

Otis, A. S. (1989). Otis-Lennon school ability test. Psychological Corporation, Harcourt Brace Jovanovich.

Perera, C. J., Sumintono, B., & Jiang, N. (2018). The psychometric validation of the principal practices questionnaire based on item response theory. *International Online Journal of Educational Leadership*, *2*(1), 21-38.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Sapp, G. L., & Marshall Jr, J. (1984). The Otis-Lennon school ability test: A study of validity. Psychological reports, 55(2), 539-544

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Zenisky, A. L., Hambleton, R. K., & Sired, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. Journal of Educational Measurement, 39(4), 291-309.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In Test scoring (pp. 85-152). Routledge.