Development and Calibration of Geometry Diagnostic Test for Upper Basic Education Students in Plateau State, Nigeria

Stephen Banwar Wanlor.[1], Prof. Sayita G. Wakjissa.[2], Prof. Yusuf A. Mustapha[3]

[1]Primary Education Studies Department, Federal College of Education, Pankshin, Plateau State, Nigeria

[2,3]Research Measurement and Evaluation Unit, Department of Educational Foundations, Faculty of Education, University of Jos, Nigeria

DOI: https://dx.doi.org/10.47772/IJRISS.2025.90300202

Received: 10 March 2025; Accepted: 18 March 2025; Published: 07 April 2025

ABSTRACT

This paper developed and calibrated a geometry diagnostic test for Upper Basic Education students in Plateau State, Nigeria. The purpose of the paper was to improve the performance of students in mathematics through the mastery of geometry contents which many Nigerian students found difficult. The study utilized instrumentation and survey research designs. The population was 20,550 UBE 3 students from Plateau Central Senatorial Zone drawn using probability proportionate to size sampling method. The validity of the UBE 3-GDT was estimated through experts' judgment, test blue prints and data model fit while reliability was estimated using test information function. The study was guided by four research questions. Data analyses were by chi- square goodness of fit statistics and item analysis indices. The finding of the study revealed that the diagnostic items fit the 3PLM. The test information function 3.278 obtained through the use of the maximum likelihood estimate statistics indicated a high reliability of between 0.75 and 0.91. In terms of difficulty, 93% of items had indices between 0 and 0.5, indicating that they were acceptable for diagnostic tests. The discrimination indices range from fair to moderate, that is, 0.20 to 0.50 and 0.50 to 1.00, and pseudo-guessing of 96% of items was within the acceptable range, that is, $0.20 < c \leq 0.30$. Thus, the UBE 3-GDT appears multidimensional in nature. It was concluded that diagnostic tests improve students' performance in mathematics and recommended that the UBE mathematics teachers should make adequate use of diagnostic tests in teaching geometry.

Keywords: development, calibration, geometry, diagnostic tests

INTRODUCTION

The primary reasons Nigerian students score poorly in mathematics include low-quality teacher-made tests, instructors' inability to design quality items, evaluation and students' lack of foundational skills. These make students perform poorly in the West African Examination Council (WAEC) (Dadughun, 2015; WAEC, 2019 - 2023). The chief examiner's annual report shows that Nigerian students perform more poorly in geometry than any other branch of mathematics in the Senior Secondary Certificate Examination (SSCE), while Wanlor, Dalong and Olakunle (2023) found that Upper Basic Education students in Plateau State perceive geometry concepts to be more difficult to learn than other branches of mathematics too.

It is common for students to fail to accomplish tests without figuring out why. Therefore, researchers and assessment specialists need to develop geometry diagnostic tests to assist students in remedying their difficulties and misconceptions in mathematics. These evaluations help teachers tailor their instruction, develop individualised lesson plans, and provide targeted support through remedial education by identifying students' strengths and shortcomings (Nuraini, Cholifah, & Laksono, 2018).

The development of diagnostic tests entails producing top-notch items to evaluate students' proficiency in geometry (Mawak, 2019). Using test results, calibration entails establishing item difficulty, discrimination, and guessing parameters. Item calibration makes use of test theories such as Generalisability Theory, Item Response Theory, and Classical Test Theory (Mawak, 2019). In order to detect misconceptions and close learning gaps in geometry instruction at the Upper Basic Education Schools, the paper set out to develop and calibrate a diagnostic test in geometry using the Item Response Theory (IRT). In order to ensure that the items had quality, psychometricians were given the GDT to evaluate the test information function, dimensionality, item characteristics, and data fit.

A data model fit in IRT is equivalent to validity in CTT. Thus, one of the significant benefits of IRT over CTT is in its ability to fit data to a model. Following the UBE 3-GDT trial test, if an item or items was or were determined not to match a specific model, the model or item(s) would be dropped. In certain cases, the data might not fit a model. When this happens, the tester should do pairwise item fit to reveal the items that fit the model. The goodness of fit chi-square was used to determine fitness while Xcaliber was used to run statistics created for the 3-PL models on the UBE 3-GDT data. The item fit information guideline is presented in

Figure 1.

FIT	Interpretation
Infit MNSQ (0.7 - 1.3) Outfit MNSQ (0.7-1.3), z (-2 to 2) Infit/outfit MNSQ (<0.7 or > 1.3), z (<-2 or > 2) Infit/outfit MNSQ (> 1.3), z (> 2)	Good fit misfit Underfit (unpredictable item)
Infit/outfit MNSQ (< 0.7), z (< -2)	Overfit (item too predictable)

Figure 1: Interpretation of item fit Indices of a Diagnostic Test.

Source: American Educational Research Association. (2020).

The diagnostic tests that employ the 3PLM have a difficulty parameter (b) of less than -2.50 to -1.50 for very easy items, -1.50 to -0.50 for easy items, -0.50 to 0.50 for moderately tough items, 0.50 to 1.50 for hard items, and 1.50 to 2.50 for extremely hard items (Embretson, 2020). The item difficulty parameter must be greater than or equal to 0 but less than or equal to 3 for the 3 PLM in order to meet the selection criterion for item difficulty (Wu & Adams, 2020). Three PLM items with b values ranging from -0.50 to 0.50 are used in the diagnostic test item selection process (de Ayala, 2022), and the optimal item is one where b = 0.00 (Wu & Adams, 2020).

For dichotomous items, the location on (theta) where the probability of a correct answer equals c/2 + 0.50 is where the "a" parameter represents the slope of the Item Response Function (IRF) or the Item Characteristics Curve (ICC) at its maximum. For item discrimination in a diagnostic test using the 3PLM, items with values ≤ 0.00 are extremely low and should be removed from the test (Kim & Lee, 2022); 0.20 to 0.50 have low a-values and are performing poorly (Li & Wang, 2022); 0.50 to 1.00 have moderately discriminating values among test takers (Wu & Adams, 2020); 1.00 to 2.00 have high a-values and are excellently discriminating among test takers; and >2.00 have very high discrimination (de Ayala, 2022) among them.

The c parameter in 3PLM represents a test taker who correctly guesses an item. Also, for a five-option test, a ‘c’ value of less than or equal to 0.20 is good, while for a four-option test, a ‘c’ value of 0.25 is ideal (Milfont & Fischer, 2020). Ideal c values for diagnostic tests fall between 0.00 and 0.20 or 0.25, while moderate acceptance of c values as high as 0.30 is possible (de Ayala, 2022). There is no guessing when c = 0, low

guessing when $c < 0.20$ and optimum when $c \geq 0.20$, moderate guessing when $0.20 < c \leq 0.30$, high guessing when $c > 0.30$, and complete guessing when $c = 1$ (de Ayala, 2022).

A diagnostic test item is considered unidimensional if it assesses only one trait and multidimensional if it evaluates multiple traits. Unidimensional items provide dichotomously scored data that are calibrated using the 1, 2, 3, and 4 PLM, while multidimensional items produce polytomously (multiple response) scored data that are calibrated using the Partial Credit Model (PCM) or Graded Response Model (GRM) (Columbia Public Health, 2020). In order to help choose the best model to employ, the paper ascertained the dimensionality of the geometry diagnostic multiple-choice test items. The IRT provides a test information function as an alternative to CTT's reliability and provides detailed information on the accuracy level for various ability levels. By using properly selected items, psychometricians can precisely design the degree of reliability information for various ranges of skill (Ayanwale, 2021). Credible information about each item is provided when the test data is plotted. The test information function is low when it is less than one (Tif < 1) and moderate when one (1) is less than or equal to the TIF value and the TIF value is less than or equal to three, that is, $1.00 \leq Tif \leq 3.00$. while high TIF values are those greater than three, that is >3.00 (Milfont & Fischer, 2020).

Primi's and Primi's research in 2014 on Rasch-Master's Partial Credit Model in children's drawing assessment indicated good fit indices for most characteristics. The Positive Affect Scale features moderate discrimination and measures responses below the average score more accurately. Eleje, Nkedi, Esomonu, Koye, Obasia, and Onah (2016) found good item difficulties with discriminating indices ranging from 0.22 to 0.65 and suitable difficulty ranging from 0.24 to 0.79. Dadughun (2015) found that the Primary School Mathematics Diagnostic Achievement Test (PRISMADAT) was reliable and unidimensional, with item difficulty and discrimination parameters ranging from -0.97 to 3.21 and -0.29 to 4.95, respectively. Bichi, Hafiz, and Bello (2016) found that qualifying examination items were not stable based on item discrimination and difficulty indices, while Ayanwale (2021) found that the DEVessay-MAT and NECO-MAT tests were unidimensional, with significant differences in overall item difficulty. Thus, these studies are different but similar to the current study.

The following research questions were raised to guide the study:

1. What are the items fit parameter indices of the UBE 3 Geometry Diagnostic Test in Plateau State?
2. What are the items analysis indices of the UBE 3 Geometry Diagnostic Test in Plateau State?
3. What is the dimension of the UBE 3 Geometry Diagnostic Test in Plateau State?
4. What is the test information function of the UBE 3 Geometry Diagnostic Test in Plateau State?

METHODOLOGY

The paper utilised instrumentation and survey research designs to develop and calibrate the UBE 3-GDT. Instrumentation research is a scientific method used to test human abilities and improve curriculum development. The instrumentation and survey design aimed to establish the GDT test qualities using the 3-PLM. The population was made of 20,550, consisting of 9,938 male and 10,612 female 'students. The sample size of 1445 students, comprising of 692 males and 753 females, was drawn from 24 schools in Plateau Central Senatorial Zone using the Probability Proportionate to Size sampling technique. The instrument of the study was the Upper Basic Education Three Geometry Diagnostic Test (UBE 3-GDT). The validity of the UBE 3-GDT was estimated using the Kendall coefficient of concordance and unidimensional trait structure using the 3PLM and Chi-square goodness of fit statistics, while reliability was estimated using the maximum likelihood function of the 3PLM.

RESULTS

Research Question One

What are the items fit parameter indices of the UBE 3 Geometry Diagnostic Test in Plateau State?

Table 1 Items Fit Parameter Indices of the UBE 3 Geometry Diagnostic Test

Item ID	Key	X^2	Df	P	Item ID	Key	X^2	Df	P
1	B	10.2755	347	0.5918	36	D	11.8655	347	0.4565
2	B	8.3932	347	0.7537*	37	C	8.8216	347	0.7181*
3	A	9.3019	347	0.6769	38	C	11.2853	347	0.5046
4	C	8.7963	347	0.7202*	39	D	12.208	347	0.4291
5	D	16.7332	347	0.1599	40	B	7.9556	347	0.7886*
6	D	19.003	347	0.8885*	41	A	9.9771	347	0.618
7	B	9.7351	347	0.6392	42	A	9.2947	347	0.6776
8	A	10.1104	347	0.6063	43	C	10.3182	347	0.5881
9	C	11.1655	347	0.5148	44	B	7.7031	347	0.8079*
10	C	10.1434	347	0.6034	45	A	12.2406	347	0.4266
11	D	25.403	347	0.013	46	B	10.8994	347	0.5376
12	B	36.2927	347	0.0003	47	C	14.6102	347	0.2634
13	A	11.9192	347	0.4522	48	A	8.2465	347	0.7656*
14	D	17.9595	347	0.1169	49	B	12.8597	347	0.3793
15	D	14.5768	347	0.2654	50	B	14.3336	347	0.2799
16	C	9.2201	347	0.684	51	B	12.6317	347	0.3964
17	A	17.2262	347	0.1413	52	A	17.3127	347	0.1382
18	B	10.8617	347	0.5408	53	C	15.0092	347	0.2409
19	A	14.1578	347	0.2907	54	A	14.89	347	0.2475
20	B	11.073	347	0.5227	55	A	17.1113	347	0.1455
21	C	17.5392	347	0.1304	56	D	12.4098	347	0.4134
22	B	12.1387	347	0.4346	57	B	28.3307	347	0.0049
23	B	8.1869	347	0.7704*	58	C	12.995	347	0.3694
24	B	15.5083	347	0.2148	59	A	13.5421	347	0.3309
25	C	12.0756	347	0.4396	60	A	16.8302	347	0.1561
26	A	12.0043	347	0.4453	61	C	9.5312	347	0.657
27	A	13.5133	347	0.3329	62	B	25.107	347	0.0143
28	C	7.6801	347	0.8096*	63	D	18.774	347	0.0941
29	B	10.5165	347	0.5707	64	D	14.3048	347	0.2817
30	A	5.5741	347	0.936*	65	D	23.3258	347	0.0251
31	D	10.8183	347	0.5446	66	D	12.826	347	0.3818
32	A	19.1988	347	0.0838	67	A	20.392	347	0.05
33	C	11.8814	347	0.4552	68	A	17.6583	347	0.1265
34	A	10.7877	347	0.5472	69	C	21.3113	347	0.046
35	B	13.5383	0.6179	0.3312	70	D	14.1335	0.6105	0.2923

Note: n = 1445. Items with asterisk indicate fit for the GDT

Table 1 shows the item fit parameter indices of the UBE 3 Geometric Diagnostic Test (GDT). The results indicate that only 10 items (2, 4, 6, 23, 28, 30, 37, 40, 44, and 48) fit the model with the p-values between 0.7 and 1.3. The other 60 (85%) items misfit the model with p-values < 0.7, and hence they needed to be checked. Statistically, it means that there was no difference between expected and observed frequencies of correct answers to the item at various ability levels of UB3 students in Plateau State.

Research Question Two

What are the item analysis indices of the UBE 3 Geometry Diagnostic Test in Plateau State?

Table 2 Item Analysis Indices of the UBE 3 Geometry Diagnostic Test

Item ID	Difficulty	Discrimination	Guessing	Item ID	Difficulty	Discrimination	Guessing
1	0.260	0.492	0.176	36	0.200	0.594*	0.231
2	0.180	0.647*	0.160	37	0.320	0.457	0.243
3	0.680	0.178	0.318	38	0.480	0.402	0.248
4	0.420	0.337	0.255	39	0.180	0.723*	0.228
5	0.260	0.342	0.239	40	0.200	0.184	0.237
6	0.220	0.313	0.238	41	0.300	0.449	0.242
7	0.300	0.580*	0.239	42	0.200	0.599*	0.231
8	0.200	0.444	0.234	43	0.340	0.400	0.246
9	0.400	0.262	0.248	44	0.420	0.395	0.249
10	0.480	0.416	0.248	45	0.400	0.432	0.246
11	0.400	0.200	0.250	46	0.320	0.447	0.242
12	0.440	-0.086	0.255	47	0.260	0.523*	0.239
13	0.620	0.234	0.250	48	0.360	0.484	0.243
14	0.180	0.566*	0.230	49	0.280	0.532*	0.239
15	0.260	0.432	0.238	50	0.380	0.371	0.247
16	0.500	0.339	0.250	51	0.460	0.052	0.369
17	0.420	0.209	0.249	52	0.380	-0.015	0.346
18	0.360	0.415	0.246	53	0.220	0.078	0.373
19	0.280	0.345	0.243	54	0.240	0.233	0.243
20	0.380	0.344	0.250	55	0.240	0.124	0.239
21	0.360	0.337	0.250	56	0.180	0.264	0.233
22	0.560	0.240	0.249	57	0.520	0.060	0.249
23	0.360	0.309	0.247	58	0.180	0.050	0.236
24	0.220	-0.225	0.245	59	0.220	0.144	0.242
25	0.460	0.352	0.250	60	0.380	-0.015	0.249
26	0.240	0.495	0.236	61	0.180	0.073	0.236
27	0.500	0.325	0.253	62	0.340	0.154	0.243
28	0.340	0.405	0.246	63	0.240	0.124	0.244
29	0.340	0.451	0.245	64	0.300	0.306	0.247
30	0.440	0.302	0.251	65	0.280	0.053	0.248
31	0.300	0.382	0.244	66	0.060	0.404	0.221
32	0.380	0.240	0.254	67	0.300	0.079	0.244
33	0.300	0.488	0.241	68	0.260	-0.014	0.242
34	0.340	0.266	0.248	69	0.180	0.027	0.237
35	0.340	0.316	0.249	70	0.300	0.479	0.240

Note: n = 1445. Items with asterisk indicate acceptable discrimination for the GDT.

Table 2 presents the indices (a, b, and c parameters) of the UBE 3-GDT items. The outcome indicates that 65 items (93%) were within the acceptable difficulty level of between 0 and 0.5, which were moderately difficult and acceptable for diagnostic tests. Items 3 and 22 were easy with $b > 0.5$, while items 66, 58 and 39 were hard with $b < 0.2$. In terms of discrimination indices, none of the items discriminated high with index value 1.00 to 2.00 or very high with index value greater than 2.00, but 8 items (11%), that is, items 2, 7, 14, 36, 39, 42, 47, and 49, moderately discriminated with index value between 0.5 and 1.00, while 51 (73%) were low or fair items, which were 1, 4, 5, 6, 8, 9, 10, 11, 13, 15-23, 25-35, 37-38, 41, 43-46, 48, 50-51, 53-54, 56-58, 61, 64-67 and 69-70 with index values between 0.20 and 0.50. Also, 11 (16%) were very low or poor items, which were items 12, 24, 40, 52, 55, 59, 60, 62, 63, and 68 because they had index values $a \leq 0.00$, showing that they were too bad items that needed to be revisited because the majority of students from the low-ability group got them correctly.

In terms of pseudo-guessing (parameter c), no item had c = 0, that is, no guessing parameter. Two items representing 3%, that is, items 1 and 2, had $c \leq 0.20$. These were considered low guessing parameters that were ideal for a 5-option test; 64 (91%) of the items, that is, items 3-50, and 54-70, were moderate items ideal for a four-option diagnostic test because p values ranged from $0.20 < c \leq 0.30$. Four items representing 6% had high guessing with $c > 0.30$, and no item was completely guessed, that is, with c = 1. The finding of the study implies that most of the UBE 3-GDT items had good indices in terms of difficulty, discrimination, and pseudo-guessing among the students in Plateau State.

Research Question Three

What is the dimension of the UBE 3 Geometry Diagnostic Test in Plateau State?

Table 3 Dimension of the UBE 3 Geometry Diagnostic Test (Total Variance Explained)

Component	**Initial Eigenvalues**		
	Total	**% of Variance**	**Cumulative %**
1	8.785	12.550	12.550
2	3.522	5.032	17.582
3	2.153	3.076	20.657
4	1.946	2.780	23.438
5	1.729	2.470	25.907
6	1.530	2.186	28.093
7	1.498	2.140	30.233
8	1.405	2.008	32.241
9	1.369	1.955	34.196
10	1.311	1.873	36.069
11	1.275	1.821	37.890
12	1.249	1.785	39.675
13	1.191	1.702	41.377
14	1.163	1.662	43.039
15	1.146	1.637	44.676
16	1.137	1.624	46.301
17	1.087	1.553	47.853
18	1.081	1.544	49.397
19	1.071	1.530	50.927
20	1.055	1.507	52.434
21	1.035	1.479	53.913
22	1.009	1.442	55.355

Table 3 shows the total variance explained for the GDT items. The data revealed 22 underlying factors, that is, 22 dimensions with eigenvalues higher than one (factors/components were reliable and explained substantive dispersion) with 8.785 total variance of the first factor, which is almost three times greater than the second factor, which had 3.522. The first factor explained 12.5% variance, and the second explained 5.032% of the residual variance, while the remaining variance was accounted for by the other factors. Also, the 22 factors with eigenvalues greater than one accounted for 55.36% of the total variance. This indicated that the variance has a dispersed distribution across multiple components, suggesting that the test was multidimensional, corresponding to various aspects of geometry proficiency rather than being dominated by a single factor.

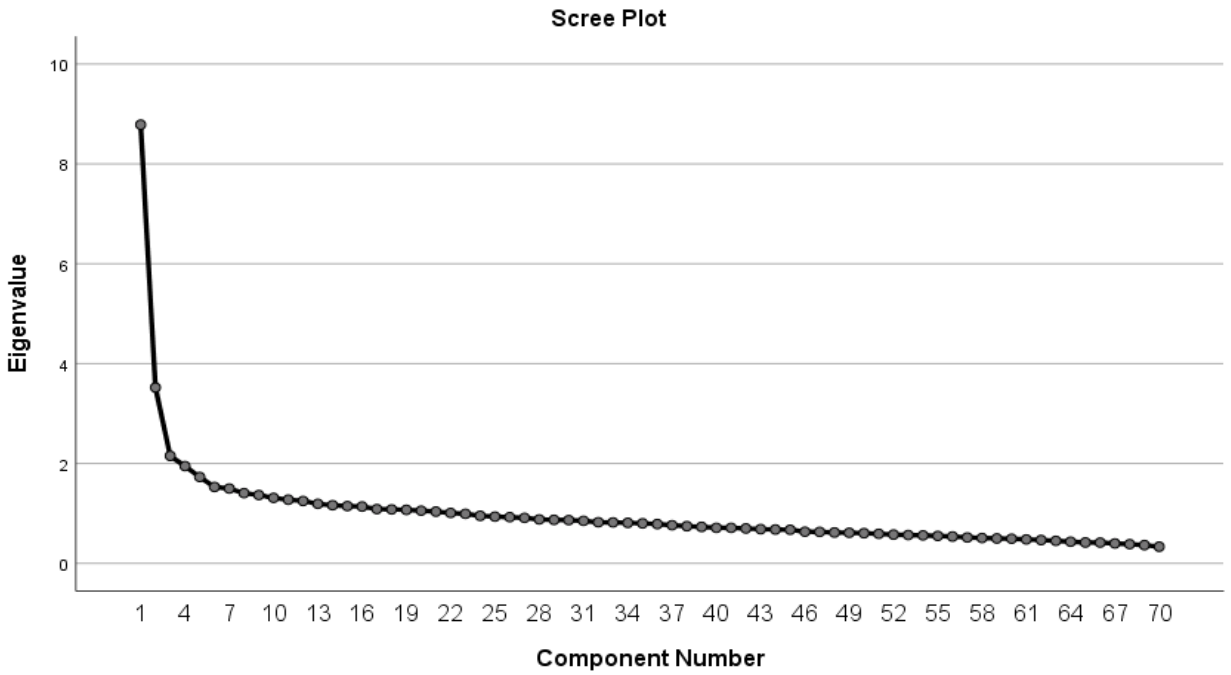

Figure 2. Scree Plot for UBE 3 Geometry Diagnostic Test (Multiple-Choice).

Figure 2 shows that the eigenvalues decreased gradually from Component 1 (8.785) without a clear "elbow point. This gradual decline supports a multidimensional structure, as there was no single, dominant factor or a small set of dominant factors.

Research Question Four

What is the test information function of the UBE 3 Geometry Diagnostic Test in Plateau State?

Table 4 Test Information Function of the UBE 3 Geometry Diagnostic Test

Item ID	Max Info	a SE	b SE	c SE	Item ID	Max Info	a SE	b SE	c SE
1	3.60	1.35*	0.65	0.23	36	0.59	0.58	0.61	0.27
2	4.78	1.38*	0.59	0.22	37	0.48	0.48	0.53	0.29
3	0.94	0.51	0.19	0.48	38	0.40	0.46	0.33	0.38
4	0.59	0.48	0.46	0.31	39	0.59	0.57	0.60	0.27
5	0.45	0.58	0.64	0.28	40	0.41	1.01*	0.74	0.26
6	0.44	0.81*	0.69	0.27	41	0.51	0.51	0.56	0.29
7	0.53	0.47	0.49	0.30	42	0.59	0.56	0.61	0.27
8	0.48	0.69	0.66	0.27	43	0.48	0.49	0.54	0.29
9	0.32	0.43	0.48	0.33	44	0.42	0.44	0.42	0.33
10	0.44	0.46	0.31	0.38	45	0.42	0.44	0.40	0.34
11	0.26	0.43	0.54	0.33	46	0.48	0.47	0.52	0.30
12	0.16	0.43	0.61	0.34	47	0.54	0.56	0.61	0.28
13	0.35	0.54	0.28	0.51	48	0.48	0.44	0.42	0.32
14	0.55	0.67	0.64	0.27	49	0.52	0.51	0.56	0.29
15	0.42	0.52	0.61	0.29	50	0.43	0.44	0.45	0.32
16	0.37	0.46	0.33	0.39	51	0.90	2.84*	3.86	0.25
17	0.29	0.42	0.46	0.34	52	0.94	0.87*	0.57	0.28
18	0.39	0.45	0.50	0.31	53	0.54	1.72*	1.72	0.28
19	0.47	0.61	0.65	0.28	54	0.63	0.77*	0.61	0.26
20	0.38	0.46	0.54	0.31	55	0.43	1.18*	0.68	0.26
21	0.41	0.50	0.59	0.30	56	0.36	0.41	0.41	0.35

22	0.32	0.51	0.31	0.45	57	0.58	0.77*	0.62	0.26
23	0.34	0.45	0.55	0.31	58	0.53	0.65	0.64	0.27
24	0.42	1.47*	0.94	0.26	59	0.54	0.72*	0.63	0.27
25	0.39	0.44	0.37	0.36	60	0.42	0.47	0.57	0.29
26	0.56	0.57	0.61	0.28	61	0.45	0.56	0.64	0.28
27	0.35	0.45	0.36	0.38	62	0.54	0.66	0.63	0.27
28	0.46	0.48	0.53	0.29	63	0.52	0.57	0.63	0.27
29	0.47	0.48	0.53	0.29	64	0.47	0.57	0.64	0.28
30	0.37	0.43	0.41	0.34	65	0.48	0.60	0.64	0.28
31	0.45	0.55	0.62	0.29	66	0.36	0.41	0.37	0.38
32	0.39	0.51	0.60	0.30	67	0.35	1.31*	1.14	0.26
33	0.52	0.49	0.54	0.27	68	0.38	0.51	0.63	0.29
34	0.41	0.50	0.59	0.29	69	0.40	0.53	0.64	0.28
35	0.42	0.52	0.61	0.29	70	0.40	0.53	0.64	0.28

Note: n = 1445.

Number in asterisks* = show high SE > or = 0.70 indicating IIF of low Reliability.

Table 4 shows the item information function (IIF), and figure 3 displays the test information function (TIF) for all the GDT items. The maximum information was 4.78 at theta = 4.000. At the cut-point of theta = 2.60 (Expected Precision Curve [EPC] = 0.50), the TIF equalled 3.278, which implied that the test had high information or a good reliability and precision. For item information functions, the standard error (SE) for item discrimination (a SE) was used, and the rule of thumb is that an SE less than 0.70 is ideal (Thissen, 2000). Therefore, 57 items (81%) had a standard error less than 0.70, while the other 13 items (19%) had an SE above 0.70. This implied that 81% of items provided much information at each ability level of the UBE 3 students in responding to the Geometric Diagnostic Test (GDT).

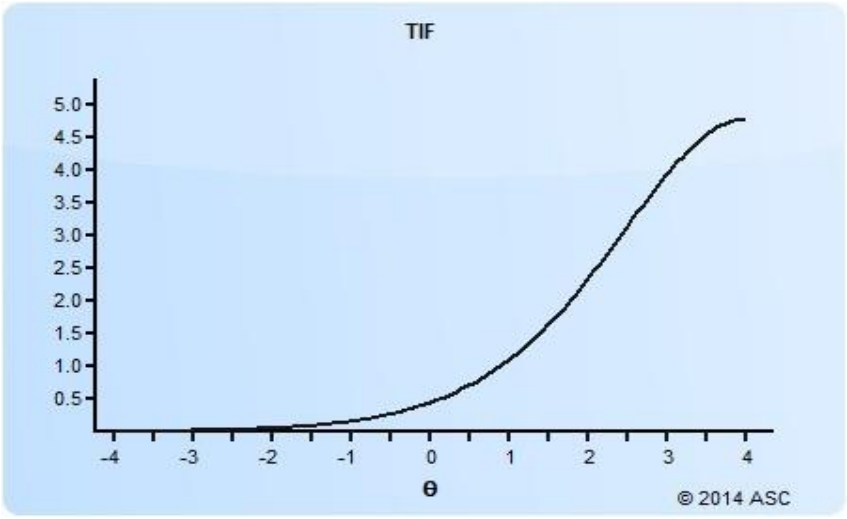

Figure 5. Test Information Function.

DISCUSSION OF FINDINGS

The experts' judgement, table of specification and data item fits of the instrument showed that only 10 (14%) out of the 70 items fit the 3PLM with p-values between 0.7 and 1.3, which were ideal criteria for diagnostic test item selection, while 60 (86%) items with p-values less than 0.7 showed misfit. This is in contrast with the findings of Primi and Primi (2014), who posited that when data fits a model, it is valid and will lead to accurate judgement of testees abilities.

Since most of the UBE 3-GDT did not fit the 3 PLM, the researchers conducted pairwise item fit, in which case some of the items that fit the model were discovered. The Pairwise Item Fit helped the researchers to

refine and/or remove misfitting items to consider alternative models or group locally dependent items (Reckase, 2020; Embretson & Reise, 2020). The finding is also in contrast to Dadughun findings that a significant difference existed between the proportions of PRISMADAT Forms A and B items under Practical and Descriptive Geometry that the 3-PLM fitted.

The item information function of the GDT revealed that 57 items representing 81% of the total items were reliable, with a reliability coefficient ranging from 0.60 to 0.83 and SE < 0.70. Only 13 items (19%) of the UBE 3 GDT were not reliable, with ranges from 0.00 to 0.59 and SE ≥ 0.70. Furthermore, the TIF equalled 3.278, which implied that the test had high information or a good reliability and precision. This finding was similar to Dadughun's finding, with a reliability coefficient of 0.88 and 0.93 established using the Cronbach alpha reliability method. The present studies differ in the method of determining reliability; that is, the current study used IRT methods, while Dadughun (2015) used CTT methods.

In terms of item parameters, it indicated that 65 of the items (93%) were within the acceptable difficulty level of 0 to 0.5, which was moderately difficult and acceptable for a diagnostic test. Items 3 and 22 were easy (b > 0.5), while three items, 66, 58 and 39, were difficult (b < 0.2). In terms of discrimination indices, none of the items were found to discriminate highly with index values between 1.00 and 2.00 or to discriminate very highly with index values greater than 2.00, but 8 (11%) of the items – 1, 7, 14, 36, 39, 42, 47, and 49 – discriminated moderately with index values between 0.50 and 1.00, while 51 (73%) had low or fair item discrimination indices of between 0.20 and 0.50. Also, 11 (16%) were very low or poor items with discrimination indices of a ≤ 0.00.

In terms of pseudo-guessing (parameter c), no item had c = 0; that is, no guessing parameter. Two (2) (3%), that is, items 1 and 2 had c ≤ 0.20, that is, low guessing parameter that is ideal for five-option test items, 64 (91%) items; 3-50 and 54-70 had 0.20 < c ≤ 0.30, which were moderate items ideal for a 4-option diagnostic test, and 4 (6%) items had high guessing parameters with c > 0.30, and no item was completely guessed, that is, with c = 1. The findings of the study generally implied that most of the GDT items had good indices in terms of difficulty, discrimination, and pseudo-guessing. These findings corroborate those of Eleje, Nkedi, Esomonu, Koye, Obasia and Onah (2016) and Zanon, Hutz, Yoo and Hambleton (2016), who found good item discrimination and suitable difficulty indices ranging from 0.22 to 0.65 and suitable difficulty ranging from 0.24 to 0.79, and in the 2008 SSCE Biology multiple-choice test had a discriminating power of 0.39, but the findings are in contrast to Bichi, Hafiz, and Bello (2016), who found that qualifying examination items were not stable based on their discrimination and difficulty indices.

The study found that the UBE 3 Geometry Diagnostic had multidimensional structures, as the variance was distributed across many components with no single dominant factor. The presence of numerous components with eigenvalues > 1 suggested that test assessed multiple underlying constructs, likely corresponding to various aspects of geometry proficiency. This study is at variance with that of Dadughun (2015) who found that both the PRISMADAT Forms A and B were unidimensional as well as Ayanwale (2021) who also found that the instrument used in Mathematics Achievement Test Using Generalized Partial Credit were unidimensional

CONCLUSION

Mathematics teachers should not only depend on achievement tests but should also develop diagnostic tests whose psychometric properties have been determined to ensure test quality to specifically identify students' areas of difficulties and misconceptions in tasks. This will help both students and teachers to find remedies and close gaps through remedial instructions in the teaching and learning of geometry concepts.

RECOMMENDATION

The following recommendations were made based on the findings of the study:

1. The UBE mathematics teachers should develop and make adequate use of diagnostic tests in teaching geometry concepts so as to detect students' areas of difficulties and misconceptions for remediation.

2. The UBE mathematics teachers should ensure that they determine the psychometric properties of their diagnostic test items before administering them to students.
3. The Plateau State SUBEB mathematics teachers in the study area need training and re-training in the development and calibration of instruments using the IRT approach.
4. The mathematics teachers should obtain and use calibrated diagnostic test items in their classrooms.

REFERENCES

1. American Educational Research Association. (2020). Standards for educational and psychological testing. American Educational Research Association.
2. Ayanwale, M. A. (2021). Calibration of Polytomous Response Mathematics Achievement Test Using Generalized Partial Credit Model. British Journal of Education, 5(2), 21-41.
3. Bichi, A. A., Haiz, H., & Bello, S. A. (2016). Evaluation of Northwest University, Kano Post-UTME Test Items Using Item Response Theory. International Journal of Evaluation and Research in Education, 5(4), 261-270.
4. Columbia Public Health (2020). Item response theory. Retrieved from htttps://www.publichealth.columbia.edu/research/population-health-methods
5. Dadughun, S. I. (2015). Development and calibration of a primary school Mathematics diagnostic test based on item response theory [Doctoral thesis, University of Nigeria, Nsukka, Nigeria]
6. de Ayala, R. J. (2022). Item response theory. Oxford University Press.
7. Eleje, L. J., Nkedi, P., Esomonu, M., Koye, R. O., Obasia, E., & Onah, F. E. (2016). Development and validation of diagnostic Economic Test for Secondary schools. Retrieved from https://www.researchgate.net/publication/304357127. DOI: 10.5430/wje. v6n3p9.
8. Embretson, S. E. (2020). Cognitive psychology and educational assessment. Journal of Educational Psychology, 112(4), 541-563.
9. Embretson, S. E. & Reise, S. P. (2022). Item response theory for psychologists (2nd ed.). New York: Psychology Press.
10. Kim, J., & Lee, Y. (2022). Applying item response theory. Willey.
11. Li, Z., & Wang, W. (2022). Item response theory: Principles and applications. San Diago, CA: Academic Press.
12. Mawak, J. J. (2019). Development and calibration of economics achievement test for secondary school students in Plateau State, Nigeria. [Graduate Theses and Dissertations, University of Jos, Nigeria]
13. Milfont, T. L., & Fischer, R. (2020). Measurement invariance and bias in cross-cultural reseach. Journal of Cross-Cultural Psychology, 51(4), 349-364.
14. Nuraini, P. S., Cholifah, N. L. S., & Laksono, W. C. (2019). Mathematics errors in elementary school: A meta-synthesis study advances in social science, education and humanities research (ASSEHR), 1st International Conference on Early Childhood and Primary Education, 24 (4), 244 259. Retrieved from https://www.researchgate.net/publication/328313601
15. Primi, T. D. N., & Primi, R. (2014). Rasch-Master's partial credit model in the assessment of children's creativity in drawings. The Spanish Journal of Psychology, 17(35), 1–16.
16. Reckase, M. D. (2020). Item response theory. In Sage Handbook of Research Methods in Psychology, Thousand Oaks, CA: Sage
17. Wanlor, S. B., Dalong, O. M., & Olakunle, F. J. (2023). Perception of geometric concepts learnt by Upper Basic school students to enhance quality education in Plateau State. Nigerian Journal of Educational Research and Evaluation, 22 (22), 138-148.
18. West African Examination Council (WAEC, 2019 - 2023). Chief examiners annual reports. https: www.waecgh.org>examiner
19. Wu, M., & Adams, R. J. (2020). Item response theory. SAGE Publications.
20. Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R, K. (2016). An applying of item response theory to psychological test development. Psicologia: Reflexão e Crítica, 2(16), 2-18.