Effect of the Age and Gender on the Reliability of Draw-a-Person Test

Ochilbek Rakhmanov[1], Senol Dane[2*]

[1]*Department of Computer Science, Faculty of Natural and Applied Sciences, Nile University of Nigeria, Abuja, Nigeria*

[2]*Department of Physiology, Faculty of Basic Medical Sciences, College of Health Sciences, Nile University of Nigeria, Abuja, Nigeria*

ABSTRACT

Introduction: Draw-a-Person test (DAPT) continues to be one of top 10 tools used by practitioners in the education and psychology field. There are several studies on the reliability of this test, all of which presented a strong correlation and significance between retests. In this study, we aimed to observe how features like age and gender affect the reliability of the test among schooling children.

Materials and Method: A total of 447 students aged between 48 and 130 months, from different regions of Nigeria took part in this experiment.

Results: show that correlation coefficient between tests was as high as 0.807, thus presenting enough evidence of the reliability of DAPT.

Conclusion: Female students tend to improve their DAPT score on retest, regardless of their age, while male students at younger age tend to improve their DAPT score on retest. Older male students on the other hand, fail to show the same level of improvement on their DAPT on retest.

Key words: Draw-a-person test, Age, Gender, Reliability

INTRODUCTION

Human figure drawing HFD tests have been employed as information gathering tools to determine the cognitive abilities of the person. These tests are popular among practitioners because they are unbiased, easy to conduct, easily interpreted from several aspects, etc. [1,2]. Some of the most popular HFD tests include those developed by Goodenough, et al. [3-6], to name a few.

The draw-a-person test

The Draw-a-Person Test (DAPT), also known as Draw-a-Man Test (DAMT), was introduced by Goodenough, e al. to assess children's mental development. In her study, she proposed a standardized scoring system based on the drawings of 2,300 primary school children, aged four to ten years. The DAMT had 51 scoring items [3]. Although the DAMT was primarily employed as a measure of cognitive ability, Goodenough did not rule out the possibility of using the test in a more interpretive manner that might provide insight into a child's personality. This

motivated many researchers to study the DAMT from different perspectives. After several reviews by researchers, Harris, et al. [4] updated the research of Goodenough, et al. [3] and sought to improve the DAMT. It was expanded to include the drawings of a woman and the self. Thus, in the Goodenough-Harris Drawing Test (GHDAMT), three drawings were collected instead of one [4]. Over the years, the Goodenough DAMT has been revised many times with added measures for assessing intelligence, but the origin of the test has remained unchanged.

Over the years, DAMT has become a controversial tool due to doubts over the validity and reliability of the results. Cox, et al. found significant differences between the drawings by normal children and those with developmental delays [7]. Fabry, et al. [8] reported significant correlations between the amount of details in DAMT results and performance IQ, verbal IQ, and general IQ, respectively in children with behavioural and emotional problems. One similar study conducted by Naglieri, et al. reported significant differences in emotional adaptation between normal students and those with conduct and oppositional disorders who attended a psychiatric day-care treatment centre [9].

Despite these miscellaneous results, DAMT is still dominated by the global approach. Researchers have assumed that children's drawings reflect their basic personality and adaptation. Nevertheless, the instrument is among the top 10 tools used by practitioners [10].

Reliability of DAMT

There are two important factors that researchers focus on when conducting the Draw-a-Man test: validity and reliability of the test. According to Anastasi [11], the reliability of a test refers to the "consistency of the scores obtained by the same person, when re-examined by the same test on different occasions". Dunn [12] was one of first to question the reliability of the Draw-a-Man test in his study, where he conducted the test twice and checked the correlation between both results. He achieved a correlation coefficient of 0.93. Considering this, Dunn [12] and several other researchers deem the DAMT a reliable tool for assessing children's mental development. Another study conducted to test reliability of DAMT was done by William, et al. [13], who tested the reliability of DAMT: IQ, which was recently developed by Reynold, et al. [14]. Reynold, et al. [14] used Cronbach's alpha coefficient to test the reliability of the test and found a median alpha coefficient of 0.82 for the entire norm group. On retest, they calculated that the correlation between scores was 0.86; however, this test focused on one group and not the whole test population. The outcome of the test conducted by William et al. [13] which was a mean alpha coefficient.82 for the entire norm group—was like that carried out by Reynold, et al. [14].

This leads to the question of whether the test's reliability is affected by different age groups or gender. How well are children performing on retest? According to the Goodenough-Harris Drawing Test scoring criterion (51 points in total), if children score +2 points or more during retest, they have a mental age of +6 months, in comparison to the first test, which is a serious difference and may affect judgment. Another question that was raised was whether we should expect an increment or decrement in retest, and how statistically significant this outcome is. So, all these questions caused us to conduct further studies on DAMT, as well as a comparative study between age groups and tests.

Picard also studied the effect of age and gender on the reliability of DAMT [15]. In his research, he compared the DAMT scores of boys and girls from a sample of 336 students, aged 5 to 12 years (from kindergarten to Grade 6). He found that girls outperformed boys in Grades 3 and 6. He also established relationship between age and gender, claiming that boys' scores increased with their age, while girls showed no such improvement with age. Reynold and Hickman (2004) stated that the mean alpha coefficient for both male and females was 0.8 [14].

The results presented from Brown's research are also highly related to our study. In his study, Brown tested the reliability of DAMT by comparing two tests to identify the number of different items that appeared on both [16]. He compared those differences with respect to the age and found that children between 7 and 8 years and older tend to have a higher number of differences between the two sketches of DAMT.

The purpose of this study is (1) to compare the average scores of normal school children in Nigeria using the DAMT to determine if it is applicable in this region, (2) to find, compare, and present results on the reliability of DAMT, (3) to identify the factors that can affect the reliability of DAMT (age and gender).

METHOD

Participants

The Non-probability sampling method was used to select children aged from 48 to 130 months, who are students in private educational institutions. These students were in Nursery 1-2 and Primary 1-2-3-4. For the sake of diversity, we have selected students from different parts of Nigeria, including Kano, Abuja, Kaduna and Lagos [17]. Kano is in Northern Nigeria and is dominated by the Hausa-Fulani ethnic group. Lagos is in South-Western Nigeria and is dominated by the Yoruba and Igbo ethnic groups. Kaduna, on the other hand, is predominantly a Hausa-Fulani state by virtue of its location in Northern Nigeria. However, there are also many other ethnic groups present in the region. Finally, Abuja, which is Nigeria's capital city, has a diverse population comprising Hausas, Igbos, and Yorubas. However, the Hausas are the dominant group in the region since Abuja is in the North.

Table 1 shows a demographic summary of participants. The total number of participants is 447. It is easy to observe that grades are fairly distributed with respect to the age, with a difference of 1 year (approximately) between each grade. Thus, we can assume that every grade is related to one age group. For example, Nursery 1: 4.00 years- 4.99 years, Nursery 2: 5.00 years- 5.99 years, and so on. So, our graphical representations will be based on grades, rather than ages.

Table 1: Demographic summary of participants.

Grade	Total	Mean Age	Std. of Age	Gender
				27 Male (51%)
Nursery 1	53	4.65 years	1.76 years	26 Female (49%)
Nursery 2	68	5.55 years	1.66 years	41 Male (60%)

				27 Female (40%)
				39 Male (52%)
Primary 1	74	6.52 years	1.48 years	35 Female (48%)
				42 Male (49%)
Primary 2	86	7.47 years	0.98 years	44 Female (51%)
				47 Male (55%)
Primary 3	85	8.7 years	1.08 years	38 Female (45%)
				45 Male (55%)
Primary 4	81	9.48 years	0.95 years	36 Female (45%)

Ethics and privacy

About ethical concerns, the purpose of the study was briefly explained to all the parents of the participants, and the researchers promised that the test results would be kept confidential and only accessible to the school management. The Guidance counsellors of the respective educational institutions were also involved in the study as volunteers, to monitor the progress of their students and ensure the transparency of the process.

Procedure and measurement

Teachers asked students to draw a human figure (a man) during school hours. Extra-curricular activities are frequently conducted in those schools, so this experiment was an ordinary educational activity for students. Students usually put in their maximum effort in such activities due to their eagerness to outperform their classmates. So, the sketches obtained can be assumed as each student's finest work. The students were given about 10 to15 minutes to complete their drawings, which proved sufficient for them. The same experiment was conducted one week later for the same students. Figure 1 represents some sample sketches of the students.

All sketches are collected and passed to a group of trained experts. We used university students and school guidance counsellors to grade the pictures. Clear instructions were given to them about the Goodenough-Harris Drawing Test scoring (51 points in total). Presence and the ratio of the items were dominant features during scoring. Some pictures contained uncertain shapes; these ones were evaluated by other experts and a vote of majority was used to grade them. Figure 2 shows some sketches that were judged through vote of majority. Once the sketches are marked, all scores, alongside the students' personal info, were passed for further step, data analysis, and result comparison.
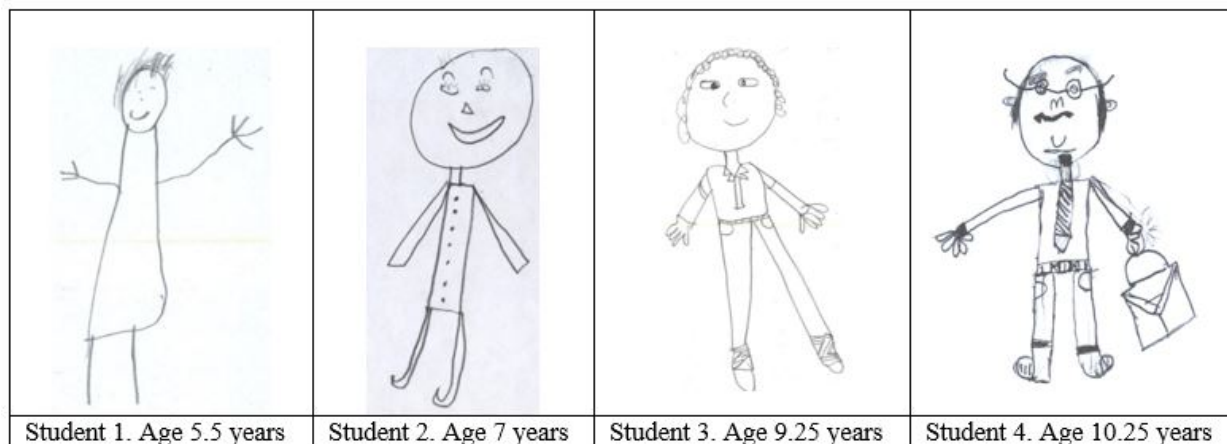

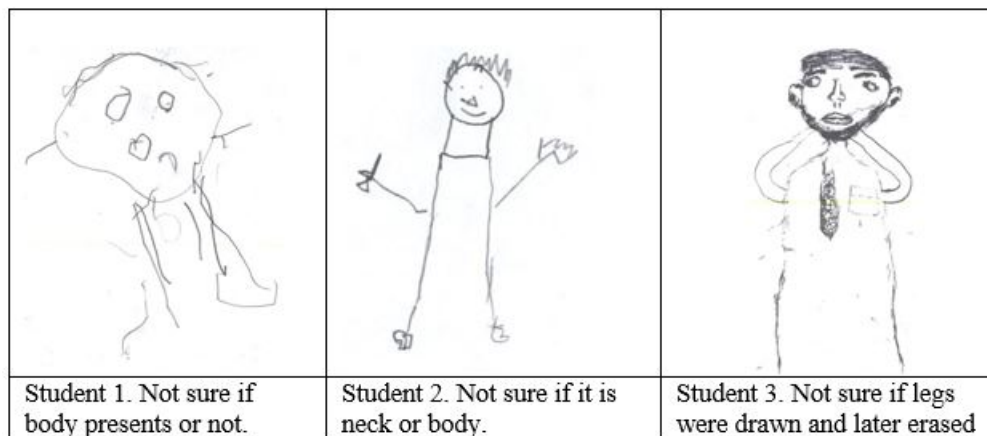Figure 1: Sample sketches of the students.

Figure 2: Some sketches which went through vote of majority.

RESULTS

Comparison of 1st and 2nd test on graph

Firstly, we observed the average scores in Test 1 and Test 2 for every grade and city. Figure 3 presents comparative graphs for each grade and school. For the aim of comparison, we put a thin straight purple line named 'Expected', which represents the expected mental score of mean age for the grade.

Observation 1: the 'expected' line tends to move up with every grade, reaching the highest point in Primary 4, where only one school managed to pass it on both the 1st and 2nd tests. The behavior of 'expected' clearly shows that the expected score of students tend to decline with every age.

Observation 2: While Nursery 1 and 2 managed to improve their scores significantly on the 2nd test, Primary 3 and 4 failed to do so. Primary 3 students underperformed on the 2nd test compared to the 1st test. For Primary 1 and 2, scores are somehow stabilized, with the fact that Primary 1 still tends to improve score on 2nd test.

Observation 3: The results of Kaduna Nursery 1, Lagos Primary 2 and Lagos Primary 4 show that it is absolutely possible that the average score of the students can either increase significantly or decline drastically as was the case for Abuja Primary 3 and Kano Primary 4.

Statistical test of the observations

Total comparison: Firstly, we tested Test1 and Test2 for all students. The results of the t-test for the 2 tests differ significantly, $t(447)=-5.277$, $p<0.001$ with $d=0.25$. While Test1 has (M=18.38, SD=8.2), Test2 has (M=19.62, SD=7.6) and correlation coefficient between them is 0.807. Thus, we have high correlation between tests, and Test2 significantly differs from Test1, with Test 2 having higher average, but with small effect size.

Gender comparison: secondly, we conducted t-test on different genders. We compared Test1 and Test2 for both boys and girls, separately.

Boys result: $t(241)=-2.702$, $p=0.007$ with $d=0.174$. While Test1 has (M=17.77, SD=8.4), Test2 has (M=18.67, SD=7.6) and correlation coefficient between them is 0.797.

Girls result $t(206)=-5.028$, $p<0.001$ with $d=0.034$. While Test1 has (M=19.1, SD=7.8), Test2 has (M=20.73, SD=7.5) and correlation coefficient between them is 0.818.

It is important to note that girls performed better than boys. They have better average score; their significance level is stronger compared to the boys'. Also, their effect size is closer to medium, while boys' is below small.

Comparison by grades: In the next step, we tested our Observation 2, which states that students at younger age improve their scores while the older ones do not. Table 2 (paired samples statistics), Table 3 (paired samples correlations) and Table 4 (paired samples t-test) shows the summary of all grades.

From table 2 it is easy to observe that all grades tend to improve their scores in the 2nd test, and all scores are highly correlated according to table 3. However, when we conducted the t-test on means, only Nursery 1, Nursery 2 and Primary 1 appeared to be statistically significant, as is shown in table 4. Yet, table 5 shows that effect size (d) for these mentioned grades are above medium, relative to other grades (Primary 2-3-4) which have small or lower than small effect size.

We should note that all three grades (Nursery 1, Nursery 2, and Primary 1), on average, improved their scores by 2 points or more (Table 2). These results led us to the hypothesis that the age group of 4-7 years (Nursery 1- 2 and Primary 1) will improve their scores significantly on retest, while the students aged 7-10 years will not (Primary 2-3-4). This raises a new question of whether this hypothesis will be affected by gender difference in those 2 age groups.

Comparison of genders in 2 formed age groups: The last test we conducted was aimed at determining whether boys and girls in the above two age groups (4-7 and 7-10 years) will behave differently. To achieve this, we tested both groups according to gender. While "Boys 4-7" and

"Girls 4-7" did not have any extra ordinary results, both supporting results of section 4.2.3, in here we present outcomes for "Boys 7-10" and "Girls 7-10" groups:

Boys 7-10: t (149)=-0.014, p=0.989 with d=0.001. While Test1 has (M=21.79, SD=7.8), Test2 has (M=21.08, SD=7.4) and correlation coefficient between them is 0.718.

Girls 7-10: t (124) =-2.803, p=0.006 with d=0.25. While Test1 has (M=22.95, SD=6.9), Test2 has (M=24.22,

SD=6.6) and correlation coefficient between them is 0.723.

These results show that boys within the ages of 7-10 significantly affect the results, while girls in this age group are still expected to improve their scores.

We can conclude that girls are expected to improve their scores for all age groups; this is not the case for boys. We expect that boys in that age group (7-10 years) will not improve their scores on retest.
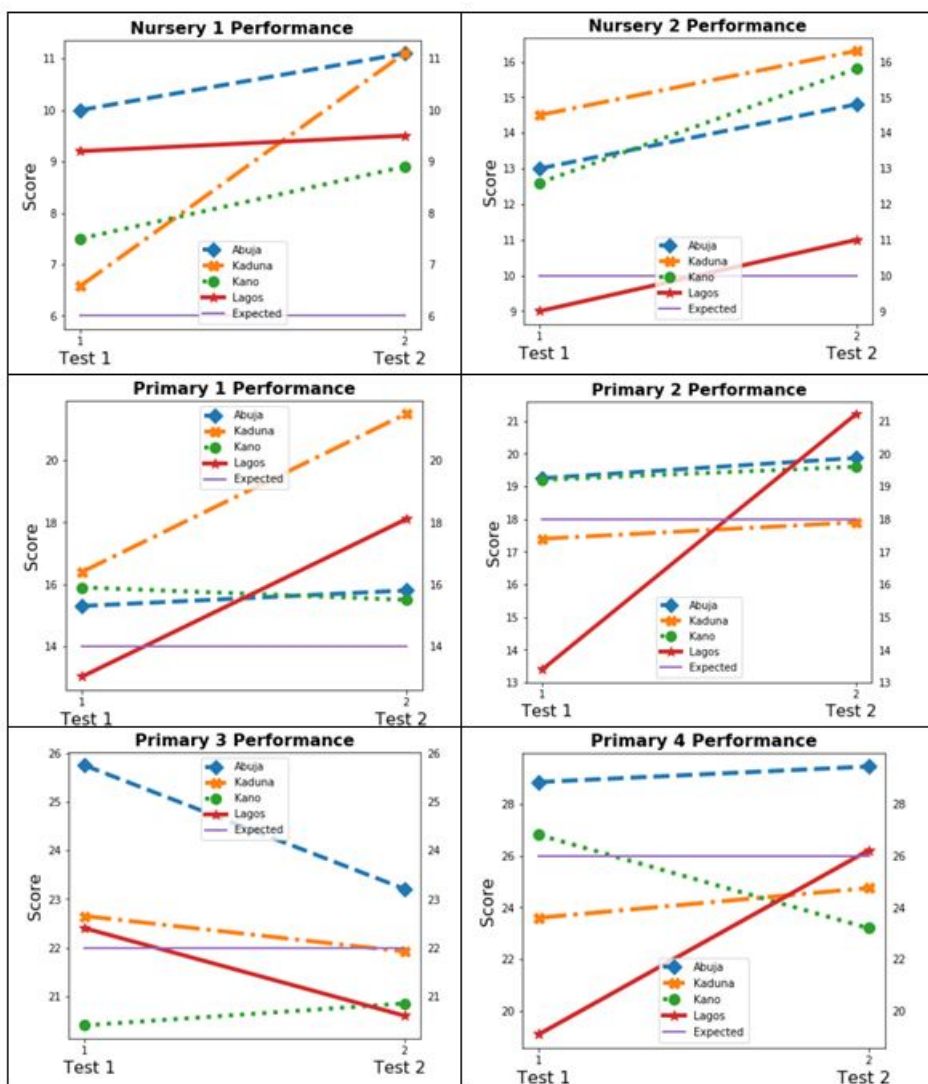
Figure 3: Comparative graphs for each grade and school.

Table 2: Paired samples statistics.

		Mean	N	Std. Deviation	Std. Error Mean
Nursery 1	Nursery1_Test1	8.7925	53	4.17575	0.57358
	Nursery1_Test2	10.766	53	4.18104	0.57431
Nursery 2	Nursery2_Test1	12.632	68	3.988432572	0.483668493
	Nursery2_Test2	15.118	68	4.166423756	0.505253095
Primary 1					
	Primary1_Test1	15.306	72	5.69566	0.67124

	Primary1_Test2	17.653	72	6.14647	0.72437
Primary 2	Primary2_Test1	18.43	86	6.131	0.66112
	Primary2_Test2	19.756	86	5.98417	0.64529
Primary 3	Primary3_Test1	23.777	85	6.36279	0.69014
	Primary3_Test2	22.906	85	6.15209	0.66729
Primary 4	Primary4_Test1	26.605	81	6.52434	0.72493
	Primary4_Test2	27.642	81	6.12027	0.68003

Table 3: Paired samples correlations.

		N	Correlation	Sig.
Nursery 1	Nursery1_Test1 & Nursery1_Test2	53	0.745	0
Nursery 2	Nursery2_Test1 & Nursery2_Test2	68	0.584	0
Primary 1	Primary1_Test1 & Primary1_Test2	72	0.594	0
Primary 2	Primary2_Test1 & Primary2_Test2	86	0.585	0
Primary 3	Primary3_Test1 & Primary3_Test2	85	0.763	0
Primary 4	Primary4_Test1 & Primary4_Test2	81	0.57	0

Table 4: Paired samples t-test.

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Dev.	Std. Error Mean	Lower	Upper			
Nursery 1	Nursery1_Test1-Nursery1_Test2	-1.77358	2.98484	0.41	-2.59631	-0.95086	-4.326	52	0
Nursery 2	Nursery2_Test1-Nursery2_Test2	-2.48529	3.72363	0.45156	-3.38661	-1.58398	-5.504	67	0
Primary 1	Primary1_Test1-Primary1_Test2	-2.34722	5.35279	0.63083	-3.60507	-1.08938	-3.721	71	0
Primary 2	Primary2_Test1-Primary2_Test2	-1.32558	5.52096	0.59534	-2.50928	-0.14189	-2.227	85	0.029
Primary 3	Primary3_Test1-Primary3_Test2	0.87059	4.30887	0.46736	-0.05881	1.79999	1.863	84	0.066
Primary 4	Primary4_Test1-Primary4_Test2	-1.03704	5.87674	0.65297	-2.33649	0.26242	-1.588	80	0.116

Table 5: Effect sizes for grades (for table 4).

Grades	Effect size
Nursery1_Test1 - Nursery1_Test2	0.6
Nursery2_Test1 - Nursery2_Test2	0.67
Primary1_Test1 - Primary1_Test2	0.44
Primary2_Test1 - Primary2_Test2	0.24
Primary3_Test1 - Primary3_Test2	0.2
Primary4_Test1 - Primary4_Test2	0.18

DISCUSSION

We started our study with three main goals, all mentioned in Section 2. We will try to answer to each of them separately.

Goal 1: To conduct DAMT to normal schooling students in Sub-Saharan region, Nigeria, to see if it is applicable in this region by comparing average scores.

We successfully conducted DAMT on children from Sub-Saharan region, Nigeria. We have selected students from quite different cultures and different regions of Nigeria. A total of 447 students from different age groups took part in this experiment. Figure 3 presents us some evidence that scores of students will be scattered around the 'expected' score. To strengthen our results, we checked the relationship between the average age and average score between each group. During calculation, we choose

the highest score for each student, from possible. While Table 6 presents these averages, Figure 4 presents the correlation and possible linear regression between them.

Table 6: Average of ages and scores for each age group.

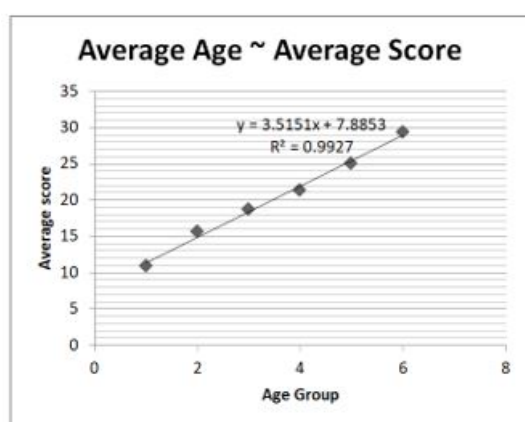|	Age Group	Age Avr.	Test Avr.
Nursery 1	1	4.65 years	10.94
Nursery 2	2	5.55 years	15.67
Primary 1	3	6.52 years	18.73
Primary 2	4	7.47 years	21.36
Primary 3	5	8.7 years	25.02
Primary 4	6	9.48 years	29.41

Figure 4: Correlation and linear regression for table 6.

Correlation coefficient is high $R^2=0.99$ and data obey to linear regression very perfectly, at least for our data in the age range of 4-10 years. This strong relation leads us to the conclusion that DAMT is a reliable way of assessing children's mental development in Nigeria.

Goal 2: Find, compare, and present results on reliability of DAMT.

Dunn, reached a correlation ratio of 0.93 during his test on the reliability of DAMT, and our findings are not too different. Even the results of the correlation check between tests was not as high as 0.93. In section 4.2.1 we presented results that a correlation between scores is high 0.807 and we have enough evidence to say that it is statistically significant. Thus, it can be concluded that DAMT is a reliable test for Nigerian school children [12].

Goal 3: Detect factors which can affect the reliability of DAMT (age and gender). One of main aims of the study was to check if age influences retest results. Picard, stated in his study that "the results did not support the hypothesis that graphic fluency would be higher in girls as compared to boys, as no sex difference was found" [15]. We also could not find a significant difference in the results of boys and girls within the age group of 4-7 years, but results from Section 4.2.4 shows that there is difference in results between girls and boys aged 7-10 years. Our observations show that girls tend to improve

their scores in all age groups, while boys lose this ability as their age increases. This is very important for the reliability of DAMT, as according to Anastasi [11], the reliability of the test means "consistency of the scores obtained by the same person, when re-examined by the same test on different occasions".

In this study, we aimed to conduct a DAMT on 447 schooling children aged 48-130 months. We found that DAMT is a useful test to assess children's mental development in Nigeria. The results of the test show that girls are statistically expected to improve their scores on retest, regardless of their age. Boys in the age group of 4-7 years are also expected to improve their scores on retest; nonetheless, this is not the case for boys aged 7-10 years. Thus, when discussing the reliability of DAMT, it is important to consider the age and gender of the student to avoid misinterpretation of the results. This finding can be connected to Pickard, who stated that girls from Grade 3 to Grade 6 will outperform boys [15]. Approximately, from same study, we conclude that girls are in the age range of 10 years to 12 years. Even though we investigated the study from different perspectives, the results supported Pickard's assertion. Our results also support the results from Brown's study, which showed that children in the age group of 7-11 years tend to draw a slightly different picture (in average 5 elements out of 105) on retest [16].

Compliance with ethical standards

This research was conducted solely for educational purpose. We have no conflict of interest to report. We didn't receive any financial support during research. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

REFERENCES

1. Baraheni N, Heidarabady S, Nemati S, et al. Goodenough-harris drawing a man test (GHDAMT)

as a substitute of ages and stages questionnaires (ASQ2) for evaluation of cognition. Iran J Child Neurol 2018; 12:94-102.

2. Laak J, de Goede M, Aleva A, et al. The draw-a-person test: An indicator of children's cognitive and socioemotional adaptation?. J Genet Psychol 2005; 166:77-93.

3. Goodenough FL. Measurement of intelligence by drawings. A text book 1926.

4. Harris DB. Children's drawings as measures of intellectual maturity: Revision of goodenough draw-a-man test 1963.

5. Koppitz EM. Psychological evaluation of children's human figure drawings. Grune and Stratton 1968.

6. Naglieri JA, McNeish TJ, Achilles N. Draw a person test. Tools of the trade: A therapist's guide to art therapy assessments. 2004; 124.

7. Cox MV, Howarth C. The human figure drawings of normal children and those with severe learning difficulties. British J Developmental Psychol1989; 7:333-339.

8. Fabry JJ, Bertinetti JF. A construct validation study of the human figure drawing test. Percept Mot Skills 1990; 70:465-466.

9. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/Miranda-Crusco-WP-CLS-20137-FINAL.pdf

10. Yama MF. The usefulness of human figure drawings as an index of overall adjustment. J Pers Assess 1990; 54:78-86.

11. Anastasi A. Psychological testing: Basic concepts and common misconceptions. Annual Meeting of the American Psychological Association, 1984, Toronto, Canada; This lecture was presented at the aforementioned meeting. Am Psychol Association 1985.

12. Dunn JA. Validity coefficients for the new harris-goodenough draw-a-man test. Percept Mot Skills 1967; 24:299-301.

13. Williams TO, Fall AM, Eaves RC, et al. The reliability of scores for the draw-a-person intellectual ability test for children, adolescents, and adults. J Psychoedu Assessment 2006; 24:137-144.

14. Reynolds CR, Hickman JA. Draw-a-person intellectual ability test for children, adolescents, and adults: Examiner's manual. Pro Ed 2004.

15. Picard D. Sex differences in scores on the draw-a-person test across childhood: do they relate to graphic fluency? Percept Mot Skills 2015; 120:273-287.

16. Brown EV. Reliability of children's drawings in the goodenough-harris "draw-a-man test". Percept Mot Skills 1977; 44:739-742.

17. Nigeria USE. Nigeria fact sheet. Economic Section, United States Embassy in Nigeria Abuja 2012.