1874-3501/22

Send Orders for Reprints to reprints@benthamscience.net

The Open Psychology Journal

Content list available at: https://openpsychologyjournal.com

CrossMark

RESEARCH ARTICLE

Inter-rater and Intra-Rater Reliability Test with Goodenough-Harris Drawing Test

Medianta Tarigan[1,*] and Fadillah Fadillah[2]

[1] *Indonesia University of Education, Bandung, Indonesia*
[2] *Bandung Institute of Technology, Bandung, Indonesia*

Abstract:

Introduction:

There are various methods to measure intelligence in children, one of them is by drawing method. Goodenough-Harris Drawing Test (GHDT) is the first cognitive measurement tool developed based on drawing. Giving quantitative scores in drawing methods has been a challenging task, especially in controlling the level of subjectivity. It requires psychometric reliability evidence to make sure the result is free from bias.

Methods:

Drawings of 799 children (kindegarten/4 to 6 y.o= 412, primary school/7 to 9 y.o= 387 ; boys= 388, girls= 411) were examined to investigate inter-rater and intra-rater reliability in GHDT scoring. Data are scored by two raters who have been given special training on the scoring system based on the manual book. Pearson correlation is used to analyzed drawing reliability.

Results:

Significant correlation between the scores of two rates were 0.840 (p<.001) for the man drawing, and 0.844 (p<.001) for the woman drawing. In addition, intra-rater reliability ranged from 0.880 to 0.963.

Conclusion:

The study showed that GHDT has a high-reliability score by raters who trained based on a manual book GHDT scoring system. In addition, these results also showed scoring criteria were applied consistently.

Keywords: Children test, Drawing test, Inter-rater, Intra-rater, Reliability, Intelligence test.

Article History	Received: November 30, 2021	Revised: March 8, 2022	Accepted: April 5, 2022

1. BACKGROUND

Measurement of intelligence in children generally aims to predict the success of children in learning and adapting [1, 2]. Various methods are used to assess children's intelligence, ranging from drawing, answering questions, or telling stories [3, 4]. Using drawing as a measurement method is considered to have many advantages. Not only is drawing an enjoyable activity, rarely children do resist [5]; drawing also has a high correlation with concept development and general intelligence [6]. It is believed that drawing provides rich insight into young children's thinking [7].

Human figures are considered the most familiar object to be drawn by children of any age and educational level [8]. Florence Goodenough developed a formal intelligence test by using human figure drawing as measurement items. In her book, *Measurement of intelligence by drawings* published in 1926, Goodenough describes the research and development process of what she called A Goodenough Draw-a-Man Test (GDAMT). GDAMT is a non-verbal test that is used to measure the level of cognitive maturity, by looking at the detailed execution of human figure drawings [9]. There were 51 strict indicators for rating each drawing, which was later developed and improved by Harris, Goodenough's student and colleague into 73 indicators for the man figure and 71 indicators for the woman figure [10]. This test is known as Goodenough-Harris Drawing Test (GHDT). Not only did Harris add score indicators, but he also added comprehensive instructions: drawing Draw-a-Woman (DAW), Self-Portrait-

[*] Address correspondence to this author at the Indonesia University of Education; Bandung, Indonesia; Tel: +6282124238492; E-mail: medianta@upi.edu

DOI: 10.2174/18743501-v15-e2207130, 2022, *15*, e187435012207130

Man (SPM; for boys), and Self-Portrait-Woman (SPW; for girls). This addition from Harris is expected to change the scoring system to be more objective than the original version. However, Harris only added a scoring system for Draw-a-Man and Draw-a-Woman, meanwhile, the Self-Portrait scoring system has not been specified yet [11].

Goodenough-Harris Drawing Test (GHDT) is the first cognitive measurement tool developed based on drawing [12]. This method has no right or wrong answer and no time limit, and is non-invasive and nonthreatening. Hence, children can respond sincerely and comfortably [13]. It is believed that a child draws what he knows, not what he sees [14]. This 'knowing' sense is defined as intellectual maturity tapped by the drawing test [13]. Through the GHDT scoring system, a total test score will be generated and described as the level of non-verbal cognitive development [15].

Until today, using the drawing method as an objective measurement method is still argued [11]. Several previous studies have shown that this method is a valid and reliable measurement method [12]. However, more studies are required to give proven and robust information related to the scoring system in achieving reliability and validity [11, 13]. Regardless of the existing set of guidelines and standards, there is still a salient problem in Human Figure Drawing test scoring and interpretation system [16]. One of the problems that raise enough pros and cons is the subjective bias of the rater in deciding to score. This happens due to various backgrounds of raters such as education, professionalism, experience, or point of view. Projection test, such as GHDT, is susceptible to misleading interpretation when the interpreter is influenced by their social stereotypes [17].

In giving a quantitative score in a drawing test, it has been challenging for the scorer to control the level of the rater's subjectivity. Therefore, it requires psychometric reliability evidence to make sure that this subjectivity bias does not affect the rating results. The classic analysis method which can be used to determine test reliability is by obtaining two scores from a group of participants and correlating the score with coefficient correlation [18]. In this context, the two scores are acquired from (a) A set of children's drawings (DAM & DAW) given scores by the same rater (intra-rater reliability), (b) two separate scores obtained from the same child's drawing rated by two different raters (inter-rater reliability), (c) scores obtained from the same child on two different times and assessed by the same rater (test-retest) [19]. Reliability testing is a technique that is generally used to measure the relationship between each score given by several raters. Consistency between the scores given by the two raters is a prerequisite to ensure the accuracy of the assessment. Inter-rater agreement and inter-rater reliability are the two indices used to ensure the assessment's consistency [20].

Several studies show that the Goodenough-Harris scoring system has good inter and intra-rater reliability [13, 21 - 23]. The reliability coefficient of inter-raters of GHDT ranged between .92 to .98 for man and woman figure drawings [13]. Research on the GHDT scoring system and cultural bias in Indonesia has already been studied. The results showed that the scoring system is not relevant to children's culture, especially in DIY Yogyakarta [24]. The study suggests conducting further research to develop an appropriate scoring system for Indonesian children. While other research that has also been conducted in Indonesia is testing the validity of items of the GHDT, where all indicators on GHDT are declared valid both on DAM and DAW. The Cronbach Alfa internal consistency reliability test showed a value of 0.92 (DAM) and 0.867 (DAW). This study suggests psychometric testing on a bigger sample size [23].

The method of inter-rater rating process in the drawing test is carried out to reduce the uncertainty of quantitative scoring. The inter-rater evaluation index is based on a comparison of the variance of scores between different raters. Meanwhile, the intra-rater evaluation index is based on a comparison of the variance of scores between two images of the same rater. The most popular method used to test inter-rater reliability is a correlation, which can be known through coefficients: Pearson, Kendall's tau, and Spearman rho [25]. A method that offers many advantages, especially simple procedures and enjoyable experiences for children, practitioners and researchers also need clear information about its psychometric strength. This study aims to determine the inter-rater and intra-rater reliability of GHDT quantitative scoring results.

2. METHODS

2.1. Participants

This study sample included 799 subjects from general Indonesian kindergarten and primary school student population (kindegarten/4 to 6 y.o= 412, primary school/7 to 9 y.o= 387 ; boys= 388, girls= 411). In Table **1**, an overview of participants' demographics is presented.

Table 1. Overview of Research Subjects.

Categorization of Research Subjects	Category	Total	Percentage
Gender	Male	388	48.56%
	Female	411	51.44%
Grade	Kindergarten	412	51.56%
	Primary School	387	48.44%
Age	4 – 6 years old	412	51.56%
	7 – 9 years old	387	48.56%

2.2. Instrument

For Goodenough-Harris Drawing Test, the subject was given 3 sheets of unlined paper (8 1/2 x 11 inches) and a pencil with an eraser. The subjects were asked to draw a man, a woman, and themself. There was no time limit set, but the subject was usually able to finish in less than fifteen minutes.

2.3. Procedure

Before data gathering, the relevant teachers were contacted for consent, permission, and discussion about assessment plans. The teachers and parents were informed as to the nature and procedures of the assessment. Five psychologists were recruited and familiarized themselves with the test procedures, class environment, and the participants days before testing was conducted. The participants were assessed in a group of classes that consisted of a maximum of 15 children per class size. Each psychologist was responsible for approx. 5 children. The standardized instruction was given in Bahasa. The duration of testing was approx. an hour per class period. Scoring of the test was done in two drawings (Draw a Man and Draw A Woman) by two raters according to the rules prescribed by the Harris scoring system. For inter-rater analysis, there were two different raters involved. Previously, the two raters had knowledge of scoring techniques. Each rater conducted scoring on all items based on the scoring manual book as a guideline for the two drawings. GHDT standard score was obtained by conversion table according to children's gender. Furthermore, raters gave a score and converted it, hence the score was 1 to 10 for each of the DAP and DAW. The intra-rater analysis also involved 2 raters to assess the drawings. Each of the raters gave scores on 43 drawing samples. The rater gave a score to two subject drawings based on the scoring manual book. Three months after, each rater gave another score for the same drawings as same as the manual book.

3. RESULTS

All data were analyzed using JASP 0.13.1.0 software for Windows. First, descriptive statistics (means, standard deviations, and frequency tables) were calculated to determine the overall data overview. Second, to study intra and inter reliability, Pearson's correlation analysis was conducted to measure the strength and direction of the rater scores relationship.

Table 2. Descriptive Statistics.

-	Rater 1 (R1)				Rater 2 (R2)			
	Draw A Man		Draw A Woman		Draw A Man		Draw A Woman	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
N	388	411	388	411	388	411	388	411
Mean	88.912	90.827	83.711	86.399	88.189	89.25	83.028	84.201
Median	87	90	81	84	86	88	81	83
Std. Dev	17.67	16.201	14.041	14.644	16.709	14.295	13.077	13.149
Sum	34409	37330	32396	35510	34041	36414	32049	34270
Minimum	56	60	52	62	56	58	52	56
Maximum	152	151	144	171	152	139	131	152

a. Inter-rater Reliability.

According to the results of the GHDT Scale Score (SS) in Table **2**, the average values for girls are greater than for boys for the two raters. In general, the average values for Draw A Man are higher than for Draw A Woman for the two raters. The highest score is a Draw A Woman by girls that scored by Rater 1. The minimum score is 52 in Draw A Woman score by boys in both raters.

Table 3. Inter-rater Reliability.

-	Variable	Reliability (*r*)
R1 vs R2	DAM	0.840
	α	< .001
	DAW	0.844
	α	< .001

The score range given by each rater is 0 – 73 for DAM and 0 -71 for DAW. The result of inter-rater reliability is shown in Table **2**. The result proved all values α < .001, with a positive correlation. Inter-rater correlation value in DAP (*r* = 0.844) is higher than that in DAM (*r* = 0.840). Both of the reliability coefficients are in the high correlation category (Table **3**).

The result of intra-rater reliability is shown in Table **4**. The result showed that both variables tested were significant, α < .001. The reliability value in R1 is higher than in R2, both in DAP and DAW. *The r-value* in DAM from R1 has a value of 0.942, while DAM from R2 is 0.880. In DAW, *the r-value* in R1 is 0.963, however, in R2, the *r-value* is 0.909.

Table 4. Intra-rater Reliability.

Variable	Reliability (*r*)	
	DAM	DAW
R1	0.942	0.963
Α	< .001	< .001
R2	0.880	0.909
Α	< .001	< .001

b. Intra-rater Reliability

4. DISCUSSION

Drawing representation made by children is a symbol of the important concepts understood about the object, and this understanding of objects develops when children start to enter primary school as their cognitive maturity gets stronger and logical. Goodenough assumes that children's drawings are complex and have a lot of meaningful marks that comprise a great deal of information about their intellectual component, which links theoretically to the psychometrics study of intelligence [26]. By making a measurement based on human figure drawing, Goodenough tried to capture information related to children's cognitive maturity levels. However, quantifying the results of children's drawings to make standardized scores for cognitive measurements is a complex challenge. There is a need for strong psychometrics evidence that proves score reliability.

The descriptive analysis showed that GHDT average IQ score for girls is higher than for boys, from both raters. This is in line with the results obtained by the previous research where GHDT average IQ score for girls is higher than boys [8, 27, 28]. It is also said that most Human Figure Drawing researchers such as Goodenough, Koppitz, Harris, and Machover found that the drawings of girls in the primary grades are superior to those of boys [29]. Meanwhile, it is found that there is no different score for both sexes in other countries such as India [30] and Greek [31]. Further research is needed to find out bias-free culture in GHDT. The descriptive analysis also showed that DAW scores are lower than DAM scores for both girls and boys. Although in the Harris scoring system, there are more scoring indicators for DAW (73) than for DAM indicators (71), this does not make the DAW score higher. It will be more difficult for children to execute female figures than male figures because female figures have a higher level of variety and demand for detail.

However, this study showed that the inter-rater reliability for Draw A Man (DAM) and Draw A Woman (DAW) has a strong positive correlation. The coefficient value is 0.840 for the DAM inter-rater reliability, and 0.844 for the DAW inter-rater reliability; all coefficients are significant (α < .001), indicating that the scoring system carried out by the two raters has good reliability. Coefficient index range from 0.70-0.89 indicated high correlation [25, 26]. The inter-rater correlation was lower than those reported by Yater, where the Pearson correlation between the scores of two raters was 0.951 for DAM and 0.954 for DAW [32]. The results of the DAM inter-rater correlation in other studies also show consistent results. The average interscorer correlation obtained when two examiners scored identical drawings was +.90 [29]. The coefficient of inter-rater reliability on DAM was shown around 0.88 [21]. The previous study also showed that the coefficient between raters using the GHDT scoring system was 0.95 (for the mentally retarded sample), 0.86 (for the learning disabled group), and 0.91 (for the normal sample) [22]. The test of the inter-rater reliability coefficient ranged from 0.83 to 0.93 in the sub-group, and 0.86 to 0.91 in the total sample study conducted on 131 children from various socioeconomic backgrounds in the UK [30]. Inter-rater agreement was good when identifying DAP scores in South African preschool children, especially in children with lower DAP IQ scores [33]. Based on these results, it can be concluded that the score given by each rater is consistent.

This study also showed that two raters gave scores consistently, both for DAM and DAW. Rater 1 showed 0.942 coefficient reliability for Draw a Man and 0.963 coefficient reliability for Draw a Woman. While Rater 2 showed 0.880 coefficient reliability for Draw a Man and 0.909 coefficient reliability for Draw a Woman. All these results were categorized as high and significant correlation (<.001). Dunn showed a 0.93 coefficient intra-rater reliability in previous research [21]. The study regarding GHDT assessment consistency by the same rater was also conducted by McCarthy, which shows a correlation of 0.95 [34]. Furthermore, the result of inter-rater reliability and intra-rater conducted by Yater also shows that the coefficient value range is from 0.750 to 0.917 [32]. This result showed that the rater consistency in rating, both for all Draw a Man and Draw a Woman data in this study, is consistent.

CONCLUSION

Inter-rater and intra-rater reliability testing in this study showed positive and significant coefficient correlation (<.001), which proves reliability. From inter-rater coefficient correlation, it can be concluded that the score given by the two raters is consistent. Meanwhile, from intra-rater correlation, it can be concluded that the rater gave consistent scores both for Draw a Man and Draw a Woman from all drawings. Furthermore, it requires a separate study to find out bias-free culture in GHDT.

LIST OF ABBREVIATIONS

GHDT	=	Goodenough-Harris Drawing Test
GDAMT	=	Goodenough Draw-a-Man Test
GHDT	=	Goodenough-Harris Drawing Test
DAW	=	Draw-a-Woman
DAM	=	Draw A Man

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was ethically approved by the Bandung Institute of Technology and the Indonesia University of Education, and the study was conducted according to APA ethical standards.

HUMAN AND ANIMAL RIGHTS

No animals were used in this research. All human research procedures followed were by the ethical standards of the committee responsible for human experimentation (institutional and national), and with the Helsinki Declaration of 1975, as revised in 2013.

CONSENT FOR PUBLICATION

Informed consent was obtained from the participants.

STANDARDS OF REPORTING

STROBE guidelines were followed.

AVAILABILITY OF DATA AND MATERIALS

The data analyzed during the current study are available from the corresponding author [M.T.] on reasonable request.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Deary IJ, Strand S, Smith P, Fernandes C. Intelligence and educational achievement. Intelligence 2007; 35(1): 13-21. [http://dx.doi.org/10.1016/j.intell.2006.02.001]

[2] Batty GD, Deary IJ, Gottfredson LS. Premorbid (early life) IQ and later mortality risk: Systematic review. Ann Epidemiol 2007; 17(4): 278-88. [http://dx.doi.org/10.1016/j.annepidem.2006.07.010] [PMID: 17174570]

[3] Crisp C. The efficacy of intelligence testing in children with physical disabilities, visual impairments and/or the inability to speak. Int J Spec Educ 2007.

[4] Hagmann-von Arx P, Lemola S, Grob A. Does IQ = IQ? Comparability of intelligence test scores in typically developing children. Assessment 2018; 25(6): 691-701. [http://dx.doi.org/10.1177/1073191116662911] [PMID: 27497247]

[5] Zlateva A. How To Read Childrens' Drawings. 2019.

[6] Raja S, John BM. An assessment of drawing age in pre-school children using 'Draw-a-man' test. J Nepal Paediatr Soc 2014; 34(1): 14-7. [http://dx.doi.org/10.3126/jnps.v34i1.9299]

[7] Anning A, Ring K. Making sense of children's drawings. Maidenhead: OUP/ McGraw-Hill Education 2014.

[8] Baraheni N, Heidarabady S, Nemati S, Ghojazadeh M. 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(4): 94-102.

[http://dx.doi.org/10.22037/ijcn.v12i4.11150] [PMID: 30279712]

[9] Brian CR, Goodenough FL. The relative potency of color and form perception at various ages. J Exp Psychol 1929; 12(3): 197-213. [http://dx.doi.org/10.1037/h0070967]

[10] Harris DB. Children's drawings as measures of intellectual maturity. New York: Harcourt, Brace & World 1963.

[11] Imuta K, Scarf D, Pharo H, Hayne H. Drawing a close to the use of human figure drawings as a projective measure of intelligence. PLoS One 2013; 8(3): e58991. [http://dx.doi.org/10.1371/journal.pone.0058991] [PMID: 23516590]

[12] Campbell C, Bond T. Investigating young children's human figure drawings using Rasch analysis. Educ Psychol 2017; 37(7): 888-906. [http://dx.doi.org/10.1080/01443410.2017.1287882]

[13] Harris DB. Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough Draw-a-Man test. New York: Harcourt, Brace & World 1963.

[14] Elsie Jones-Smith. Culturally diverse counseling : Theory and practice. Los Angeles: SAGE 2019.

[15] Amod Z, Gericke R, Bain K. Projective assessment using the Draw-A-Person Test and Kinetic Family Drawing in South Africa.Psychological Assessment in South Africa. South Africa: Wits University Press 2013; pp. 375-93. [http://dx.doi.org/10.18772/22013015782.31]

[16] I B Weiner and Ro L Greene, Handbook of personality assessment. New York, NY: Wiley 2008.

[17] Kummerow E, Kirby N. Organisational Culture: Concept, Context, and Management. London: World Scientific Publishing Company 2013. [http://dx.doi.org/10.1142/7146]

[18] Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychol Bull 1979; 86(2): 376-90. [http://dx.doi.org/10.1037/0033-2909.86.2.376]

[19] Rae G, Hyland P. Generalisability and classical test theory analyses of Koppitz's Scoring System for human figure drawings. Br J Educ Psychol 2001; 71(3): 369-82. [http://dx.doi.org/10.1348/000709901158569] [PMID: 11593945]

[20] Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. Ann Acad Med Singap 2010; 39(8): 613-8. [PMID: 20838702]

[21] Dunn JA. Inter- and intra-rater reliability of the new Harris-Goodenough Draw-a-Man Test. Percept Mot Skills 1967; 24(1): 269-70. [http://dx.doi.org/10.2466/pms.1967.24.1.269]

[22] Naglieri JA, Maxwell S. Inter-rater reliability and concurrent validity of the Goodenough-Harris and McCarthy Draw-A-Child scoring systems. Percept Mot Skills 1981; 53(2): 343-8. [http://dx.doi.org/10.2466/pms.1981.53.2.343] [PMID: 7312519]

[23] Evans R, Ferguson N, Davies P, Williams P. Reliability of the draw-a-man test. Educ Res 1975; 18(1): 32-6. [http://dx.doi.org/10.1080/0013188750180104]

[24] Partosuwindo S R, Kuwato T. Validitas tes menggambar orang dari goodenough sebagai alat pengukur kemasakan intelektual pada anak umur 5-10 tahun. J Psikol 1976.

[25] Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: A review of the literature. Med Teach 2004; 26(4): 366-73. [http://dx.doi.org/10.1080/01421590410001696434] [PMID: 15203852]

[26] Young C. drawings : an investigation using the Goodenough-Harris thesis. 2009.

[27] Wiltshire E B, Gray J E. Draw-a-Man and Raven's Progressive Matrices (1938) intelligence test performance of reserve Indian children. Can J Behav Sci / Rev Can des Sci du Comport 1969. [http://dx.doi.org/10.1037/h0082690]

[28] O'Keefe R, Leskosky RJ, O'Brien TG, Yater AC, Barclay A. Influences of age, sex, and ethnic origin on goodenough-harris drawing test performances by disadvantaged preschool children. Percept Mot Skills 1971; 33(3): 708-10. [http://dx.doi.org/10.2466/pms.1971.33.3.708]

[29] Ezell MP. Comparisons of Draw-A-Child Test Among Preschool Children. All Grad Theses Diss 1975.

[30] Gaddes WH, McKenzie A, Barnsley R. Psychometric intelligence and spatial imagery in two Northwest Indian and two white groups of children. J Soc Psychol 1968; 75(1): 35-42. [http://dx.doi.org/10.1080/00224545.1968.9712472] [PMID: 5713288]

[31] Georgas JG, Papadopoulou E. The Harris-Goodenough and the Developmental Form Sequence with Five-Year-Old Greek Children. Percept Mot Skills 1968; 26(2): 352-4. [http://dx.doi.org/10.2466/pms.1968.26.2.352] [PMID: 4871650]

[32] Yater AC, Barclay AG, Mc Gilligan R. Inter-rater reliability of scoring Goodenough-Harris drawings by disadvantaged preschool children. Percept Mot Skills 1969; 28(1): 281-2. [http://dx.doi.org/10.2466/pms.1969.28.1.281] [PMID: 4888093]

[33] Springer PE, Kalk E, Pretorius C, *et al.* Value of the Goodenough Drawing Test as a research tool to detect developmental delay in South African preschool children. S Afr J Psychol 2020; 50(1): 81-91. [http://dx.doi.org/10.1177/0081246319850683]

[34] McCarthy D. A study of the reliability of the goodenough drawing test of intelligence. J Psychol 1944; 18(2): 201-16. [http://dx.doi.org/10.1080/00223980.1944.10544119]