Validation of Test Construction Skills Inventory through the lens of Item Response Theory (IRT)

American Journal of Creative Education
Vol. 3, No. 2, 86-100, 2020
e-ISSN: 2706-6088

Francis Ankomah[1]
Regina Mawusi Nugba[2]

(Corresponding Author)

[1,2]*Department of Education and Psychology, University of Cape Coast, UCC PMB, Cape Coast, Ghana.*
[1]*Email: francis.ankomah@stu.ucc.edu.gh Tel: +233543351914*
[2]*Email: regina.nugba@ucc.edu.gh Tel: +233243570799*

ABSTRACT

The Item Response Theory (IRT) examines how testees at different ability levels respond to an item. The theory assumes that latent traits that are fewer in number than the test items account for responses to the items. This study validated the test construction skills inventory through the lens of IRT. Also, the study assessed teachers' level of skills in test construction. The descriptive survey design was employed for the study. Using the cluster sampling, 583 senior high school teachers in the Cape Coast Metropolis were engaged in the study. The 3-parameter logistic model (3-PLM) through the IRTPRO software was used to validate the test construction skills inventory. Eighteen out of 25 items were deemed valid and used to assess teachers' skills in test construction. It was found that the majority of the teachers had higher skills in test construction. Recommendations were made to the Ministry of Education, Ghana (MoE), Ghana Education Service (GES), and heads of the various SHSs in the Cape Coast Metropolis.

Keywords: *Item response theory, 3-parameter logistic model, Test construction skills, Test of maximal/optimal performance, Test construction skill inventory, Likert-type scale.*

DOI: 10.20448/815.32.86.100
Citation | Francis Ankomah; Regina Mawusi Nugba (2020). Validation of Test Construction Skills Inventory through the lens of Item Response Theory (IRT). American Journal of Creative Education, 3(2): 86-100.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
History: Received: 6 August 2020/ Revised: 8 September 2020/ Accepted: 25 September 2020/ Published: 12 October 2020
Publisher: Online Science Publishing

Highlights of this paper

- The study validated a test construction skill inventory using Item Response Theory (IRT).
- The study advances knowledge on the measurement of psychological constructs such as skills, knowledge, and competence, among others.
- The study addresses the issue of maximum and typical scores as used in the measurement of psychological constructs.

1. INTRODUCTION

Quality education is the outcome of effective teaching and learning, which assessment forms an integral part (Amedahe., 2014). Educational tests serve several purposes. The Standards for Education and Psychological Testing (American Federation of Teachers National Council on Measurement in Education & National Education Association, 1990) outlines three main purposes for which educational tests used. Educational tests provide information that is used to; make inferences that inform teaching and learning either at the individual level or curricular level; make inferences learning outcomes for a group of students or individual students; and inform decisions about students, such as promotion, certification, and placement, among others.

Teachers, therefore, have to be knowledgeable in assessment to effectively assess their students' learning. Cizek, Fitzgerald, and Rachor (1996) stressed the need for rigorous and effective training in assessment during preservice training. Teachers' competency in assessment and evaluation, specifically, competency in test construction is important to the quality of education provided (Jiraro, Sujiva, & Wongwanich, 2014).

According to American Federation of Teachers National Council on Measurement in Education & National Education Association (1990) teachers should demonstrate skills in selecting, developing, applying, using, communicating, and evaluating student assessment information and student assessment practices. Particularly, this study focuses on teachers' ability in selecting and developing assessment methods appropriate for instructional decisions. Test construction skill has to do with the abilities required to develop quality test items. Agu, Onyekuba, and Anyichie (2013) outlined these abilities to include objectivity, communication, skills in validating items, and application of certain mechanisms to determine the validity and reliability implications of test results. Skills in test construction enable teachers construct tests which are precise, objective, have an appropriate level of language, and can be used to grade appropriately. With the acquisition of these skills, teachers can construct items that: (a) yield precise answers; (b) match the age and grade level of students; and (c) will provide students with ample time to respond (Ali, 1999).

Owing to the importance placed on teachers in possessing certain competencies in test construction, numerous studies have examined the level of skills or competency of teachers both in Ghana (Amedahe, 1993; Anhwere, 2009; Quaigrain, 1992; Quansah, Amoako, & Ankomah, 2019) and outside Ghana (Agu et al., 2013; Chan, 2009; Ebinye, 2001; Hamafyelto, Hamman-Tukur, & Hamafyelto, 2015; Ololube, 2008; Onyechere, 2000; Simon, 2002). Majority of these studies have documented poor test construction skills among teachers. Poor test construction skills of teachers have adverse impact on the validity of results produced by those teachers. Gleaning from the literature, it is evident that in attempts to measure teachers' test construction skills, the majority of the studies asked questions which could yield facts on the rigours of test construction. The studies did so, but with the use of Likert-type scales (Adamu, Dawha, & Kamar, 2015; Afemikhe & Imobekhai, 2016; Agu et al., 2013). That is, the researchers used Likert-type scales in measuring competencies or skills in test construction. Methodologically, the use of Likert-type scales in measuring competencies or skills in test construction is flawed. In one of the test construction scales for example, one of the item reads: "A teacher should prepare a marking guide/scoring rubric while constructing the test for his/her class". With such a statement, if a teacher responds by checking, say 'Strongly Agree', will that mean that the teacher is knowledgeable or has that particular skill? On the hand, if another teacher also responds by checking, say 'Strongly Disagree', will that also mean that teacher is not knowledgeable or does possess that particular skill? Clearly, interpreting the responses of the teachers in the aforementioned scenarios as depicting knowledge/skill may be quite challenging, as they depict their perception in test construction rather. Strongly agreeing or disagreeing could not be judged as either right or wrong, but rather, a perception or feeling. The point we are making here is that, teachers merely agreeing or disagreeing to such factual statements about test construction does not, in any way, depict their skills or competency in test construction.

It must be noted that skills, knowledge, competence, intelligence, and so forth are tests of abilities which depict what an individual can do (Crocker & Algina, 2008; Miller, Linn, & Gronlund, 2009). In the measurement of constructs of such nature, individuals are to get a high score as possible, where high scores show more of what the individual can do in terms of test construction. Strictly speaking, Likert-type scales are mostly used to measure typical performance or score, which depict what an individual will do rather than what they can do.

Typical Likert-type scale scores are mostly used in the measurement of constructs such as feelings, interest, attitude, perception, personality traits, and reactions to issues, among others, but not knowledge, skills, competence, and so forth. With the typical constructs, the emphasis is on obtaining representative responses rather than high scores. Given this, the use of the Likert-type scales by earlier researchers in measuring teachers' skills in test construction was problematic, since they do not elicit the right information from the teachers on test construction. Based on this flaw, the current study, however, modified and validated the Test Construction Skill Inventory (TCSI) developed by Agu et al. (2013) since the responses are on a Likert-type scale. This inventory was particularly selected for this study because, it is one of the instruments predominantly used by most researchers and students in Ghana and some other African countries working in the area of test construction, and assessment in general. Specifically, the Likert-type response of the scale was changed to True/False response type, and then the scale was validated using the Item Response Theory (IRT). The key thrust of this study, therefore, was to validate a test construction skill scale using IRT, to come out with a quality instrument that can yield valid and reliable information on teachers' test construction skills.

The outcome of this study is a validated test construction skill inventory which can be adopted by various stakeholders in education such as the ministries responsible for education, school administrators, headmasters, teachers, and researchers to succinctly assess teachers' skills in test construction. This study also contributes to literature, particularly, in terms of methodology. Even though this study was particularly about test construction, the approach adopted in this study can be helpful, as it may be a guide to researchers in other behavioural and social sciences on how to measure constructs such as skills, knowledge, and competence in any area in their respective disciplines. In addition, this study provides a clear picture on the amount skills possessed by senior high school teachers in Cape Coast, Ghana. This finding may be of importance to those of other developing countries, since most of these countries have similar structures in terms of education.

1.1. Objectives

The study was driven by the following objectives:

1. To validate a test construction skill scale using IRT.
2. To assess teachers' level of skills in test construction.

2. LITERATURE REVIEW

2.1. Test Construction Skill Inventory (TCSI)

Test Construction Skill Inventory (TCSI) was developed by Agu et al. (2013). The TCSI was developed to ascertain teachers' skills in test construction. The study utilised a sample of 543 secondary school teachers in Onitsha Education Zone, Anambra State, Nigeria. Stratified proportionate sampling technique was employed to secondary school teachers from three sub-zones. TCSI was developed based on guidance on how to construct test items. TCSI initially had 30 items, which were on a 4-point Likert-type scale, namely, strongly agree (SA), Agree (A), Disagree (D), and Strongly Disagree (SD). Through exploratory factor analysis, 25 items satisfactorily met the .35 minimum factor loading requirement by Baker. (2003). These 25 items uniquely loaded on 4 factors, namely, test guidance, content coverage, language use, and item organisation. The scale had a reliability coefficient of .73. The scale was interpreted using a criterion mean of 2.50, where a mean score above 2.5 shows the presence of the skill, while a mean score below 2.5 shows the absence of the skill.

2.2. Test Construction Skills as a Measure Maximum/Optimum Performance

Generally, test construction skills fall under the broad spectrum of psychological tests. A test is a standard procedure for sampling behaviour in a particular domain (Crocker & Algina, 2008). Similarly, Nitko (2004) conceptualised a test as an instrument or a systematic procedure for observing and describing one more characteristic of an individual using either a numeric scale or a classification scheme. The former, however, applies in the context of the current study. TCSI is seen as an instrument used to sample teachers' behaviour in terms of their skills in test construction. The domain, as indicated by Crocker and Algina, is the spectrum of behaviours in test construction or the skills expected of teachers to demonstrate in their quest to develop test items to assess their students' learning. (Cronbach, 1949; Cronbach., 1990) classified tests dichotomously as either test of maximum/optimal performance or tests of typical performance.

Test of maximum/optimal performance, on one hand, refers to procedures used to determine an individual's developed skills, ability, aptitude, or achievement (Miller et al., 2009). In this regard, the focus and interest are on describing how well the individual performs when they put in their best. Tests of maximum/optimal performance tell what an individual can do. As such, scores from such measures are interpreted as the individual doing his/her best as far as the construct of interest is concerned. In principle, it is expected that tests measuring maximum/optimal performance elicit factual information, as much as possible. For instance, a test seeking to measure students' achievement in a particular area, say, algebra. In the construction of a test of this nature, the items or tasks on the test should be those that would elicit factual information in algebra, but not opinion or perception. From this, scores of two individual students can be compared, where one could be referred to as performed better or worse than the other, depending on their scores on the test. In another instance where the focus of the test is to measure teachers' skills in test construction, the test should contain factual items that border on test construction. Holding all things constant, a teacher with higher skills in test construction is expected to demonstrate the appropriate factual knowledge on each item. In both instances cited, the students and the teachers are expected to produce their best performance. For tests of maximum/optimal performance, scoring is done as either write or wrong.

Tests of typical performance, on the other hand, are tests used in measuring an individual's typical behaviour. These tests are used in determining feelings, attitudes, personality, temperaments, interests, and adjustments, among others (Crocker & Algina, 2008; Cronbach., 1990). Procedures for these type of tests are primarily concerned with what individuals will do rather than what they can do. Cronbach indicated that, in the measurement of typical behaviours, judgments such as 'good' or 'best' performance is not applicable. For example, in the measurement of personality, the interest is to describe the characteristics with no attempt to consider a particular characteristic as ideal. Similarly, in the case of interest, an individual with a high score in, say, engineering cannot be described as better than another who shows interest in politics.

Now, on the issue of whether or not the measurement of test construction skills is a test of maximum/optimal performance is an issue of concern in this study. To clearly define the category in which test construction skills can be placed, We would look at three main criteria, namely, the purpose of the test, the nature of items and its response format, and the interpretation of scores generated from the test (Ankomah, 2019). Test, in this regard, refers to the the validated test construction skills inventory by the authors of the current study. First, in terms of purpose, the purpose of the test is to sample teachers' behaviour in terms of their skills in test construction. Secondly, the items are factual and thus, elicit factual knowledge from teachers. The responses to such items are scored as right or wrong. Lastly, scores generated from the test are interpreted as what teachers can do as in terms of test construction. Therefore, interpretations such as good or poor can be applied, since it is a test of maximum/optimum performance, hence teachers can put in their best to be described as such.

In the case of the instrument under review (TCSI), the responses were on a 4-point Likert-type scale. Likert scale was invented and first used by Rensis Likert as a technique for the measurement of attitudes (Likert, 1932). Likert sees attitudes as dispositions toward overt action; it is a flexible element in personality in which such elements exist within a certain range of responses. Likert rated the five responses, namely, strongly approve, approve, undecided, disapprove, and strongly disapprove as 1, 2, 3, 4, and 5, respectively. Sum scores or means are then computed and used for interpretations thereof. Finally, the scores are interpreted as an individual's dispositions toward overt action, but not their best performance. Since the invention of Likert's scale, other scholars have adapted his technique as a way of measuring other typical behaviours such as interest, personality, feelings, and temperaments, among others. It is worthy to note that Likert-type scales are appropriate in the measurement of typical rather than maximum behaviours. For example, one of the items in TCSI reads, "A teacher should prepare a marking guide while constructing the test". Looking at an item of this nature, a teacher choosing any of the responses such as strongly agree, agree, disagree, or strongly disagree does not indicate whether the teacher knows that marking guide should be prepared while constructing the test. It rather presents an opinion or feeling rather than knowledge or skills. However, when the responses are either 'True/False' or 'Yes/No', and say, a teacher chooses 'False', this would imply that teacher is not knowledgeable on that particular item. In view of this, it can be said that Likert-type scales are not appropriate in the measurement of constructs such as knowledge, skills, intelligence, competence, aptitude, and so forth. Simply put, the use of Likert-type scales for such purposes is problematic. It must be noted that this article does not, in any way, intend to discredit the work of Agu et al. (2013), however, it simply calls attention to certain problems in the measurement of psychological constructs in general.

2.3. Item Response Theory

IRT is a test theory, which evolved as a result of the inefficiencies in the Class Test Theory (CTT). CTT is a test theory based on the fallible measures of human traits as against true objective values (Spearman, 1904). The CTT operates on the assumption that an individual's observed score on a construct is a combination of true score and error score. Some of the deficiencies in the CTT, among others, are; it is based on a composite score, it is based on a group, and it does not take into how testees with different ability levels on the trait have performed on the item (Crocker & Algina, 2008). Following these inefficiencies, among others, the IRT evolved. IRT, primarily, emphasises how individuals with different abilities on the trait respond to an item. IRT has been used for validating achievement and non-achievement tests (Baker, 2001; Morizot, Ainsworth, & Reise, 2007). IRT assumes that the ability scale is an interval scale with a zero midpoint, and scores ranging from positive infinity to negative infinity, but practically ranging from -3 to +3 (Baker, 2001). The negative part represents low ability, while positive represents high ability. The basic assumption of the IRT is that every testee (respondent) has some amount of the latent trait in question, which is referred to as ability (θ). Each testee has a score which places him/her on the ability scale. Any point on the ability scale has a corresponding probability that a testee with such ability will choose the correct response to the item. The probability is low for testees with low ability and high for testees with high ability.

IRT has several models which can be used to analysed data on polytomously and/or dichotomously scored items. The models, among others, include a three-parameter logistic model (3PLM), two-parameter logistic model (2PLM), one-parameter logistic model (1PLM), rating scale model (RSM), nominal model (NM), graded response model (GRM), partial credit model (PCM), and generalised partial credit model (GPCM). For this study, the 3PLM was used to validate the items on test construction skills. The 3PLM consists of three parameters, namely; item difficulty, discrimination, and the probability of guessing. These parameters (difficulty, discrimination, and the probability of guessing) in the CTT are not the same as in the IRT.

First, difficulty in the CTT is the proportion of testees getting an item correct. This ranges from 0 to 1. However, in IRT, it is the location on the ability scale where an item functions. The difficulty is a point on the ability scale where the probability of correct response is .5 for one- and two-parameter models and (1 + c)/2 for a three-parameter model. The difficulty denoted by 'b', ranges from positive infinity to negative infinity. A negative difficulty index means the item functions among the low ability group, and this means the item is easy. Positive difficulty index shows that the item functions among the high ability group; and thus, the item is difficult. Another parameter in 3PLM is discrimination. In CTT, discrimination is the difference between the proportion of testees in the upper and lower-scoring groups who had an item correct. Discrimination indices range from -1 to +1. In the IRT, however, discrimination refers to the slope or steepness of the item characteristic curve, which is denoted by 'a'. It tells how an item discriminates between low and high abilities. The discrimination indices range from 0 to positive infinity. Discrimination indices are interpreted based on the following criteria: none = 0; very low = .01 - .34; low = .35 - .64; moderate = .65 - 1.34; high = 1.35 - 1.69; very high > 1.70 (Baker, 2001). Guessing parameter is the lower bound for the item characteristic curve. It is denoted by 'c', and it is the probability of getting an item correct at all ability levels. It is the probability that connotes the tip of the item characteristic curve from below.

3. METHODOLOGY

The study employed a descriptive survey design. The study employed the cluster sampling technique to select a sample of 583 senior high school teachers in the Cape Coast Metropolis. The Test Construction Skills Inventory (TCSI) developed by Agu et al. (2013) was employed to gather data from the selected teachers within the Cape Coast Metropolis. The inventory has 25 items. The responses of the TCSI, which was on a 4-point Likert-type scale was, however, changed to True/False response type. Because the scale was changed to dichotomous (True/False) on the basis of methodological flaw, the logistic model was appropriate to use for the validation. Therefore, the three-parameter logistic model (3PLM) was specifically used for this study based on the fact that items on the test construction skill scale were scored dichotomously as either '1' or '0' for a correct or wrong response, respectively. Based on the scale, there was the possibility that teachers could simply guess to get an item correct. The use of 3PLM has the guessing factor incorporated. Crocker and Algina (2008) posit that 3PLM is advantageous than 1 and 2PLMs since the model can accommodate guessing. This, therefore, explains why difficulty, discrimination, and guessing were estimated for this validation. IRTPRO software version 4.2 was used for the data analysis.

3.1. Model Specification

The logistic function was employed to compute the Item Characteristics Curve (ICC). Specifically, the 3PLM was used. The 3PLM is a modified version of the 2PLM to incorporate the lower asymptotic behaviour (guessing parameter). This was specified as follows:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

The *P*(θ) in the function denotes the probability of a teacher with test construction skills at an ability level of θ. The '*b*' denotes the difficulty parameter. This shows the location on the test construction skills ability scale where the probability of correct response is (1 + c)/2. It indicates the location where the item functions among the ability groups. The 'a' in the model represents the discrimination parameter. This is denoted by the slope of the ICC. The 'c' is the chance of getting an item right by guessing. This is represented by the lower asymptotic probability. The 'e' is a constant set at 2.718.

4. RESULTS

4.1. Assumptions

Before the IRT analysis, assumptions such as unidimensionality, local independence, and speediness of the test were checked. The unidimensionality assumption was tested using confirmatory factor analysis with AMOS software version 21.0. The result suggested that scale was unidimensional, NNF = .92, IFI = .89, CFI = .86, GFI = .94, SRMR = .02, and RMSEA = .06. These imply that the model fits the data. All the items had factor loadings above .30, except for item 1, which had a factor loading of .24. Item 1 was, therefore, discarded. In all, 24 instead of 25 items were used to run the IRT. The second assumption, local independence, was tested. This assumption holds the premise that when the ability is held constant, the responses of the individual items should be statistically independent. This assumption was tested through IRTPRO. Table 1 presents the results.

Examination of the Standardized LD X^2 Statistics for the pair of items suggests that the Standardized LD X^2 Statistics were less than 10.0, except for the pair for items 22 and 23 which was greater than 10 Table 1. Based on this, it can be said that the local independence assumption was met. The implication of this is that the items are statistically independent of each other when the ability is held constant. Finally, the speed test assumption was also checked. During the administration of the inventory, respondents were given ample time to read and respond to the items. There was no specific time allocated for the return of the inventory. Respondents, however, submitted the inventory after they have duly responded to the items. This confirms that the test was not speeded.

4.2. Objective 1: Validation of Test Construction Skills Inventory (TCSI)

Having met the assumptions underlining IRT, the 3PLM was performed to validate the TCSI. Details of the results are presented in Table 2.

Table-1. Marginal fit (X2) and Standardized LD X² Statistics for Group 1.

Item	Label	X^2	1	2	3	4	5	6	7	8	9	10
1	NB2	0.0										
2	NB3	0.0	6.9									
3	NB4	0.0	1.1	-0.5								
4	NB5	0.0	0.7	1.5	-0.5							
5	NB6	0.1	-0.4	-0.6	1.1	-0.2						
6	NB7	0.0	-0.5	-0.2	-0.5	2.9	-0.6					
7	NB8	0.0	-0.2	-0.7	5.8	-0.6	0.8	-0.4				
8	NB9	0.0	7.3	1.0	7.4	-0.5	2.8	-0.4	3.8			
9	NB10	0.0	0.6	1.1	3.9	-0.7	-0.6	-0.6	-0.4	5.9		
10	NB11	0.0	-0.0	-0.3	-0.0	0.2	-0.6	1.0	1.8	2.6	-0.3	
11	NB12	0.0	-0.6	-0.7	-0.6	0.4	0.6	-0.6	1.4	-0.3	-0.5	6.2
12	NB13	0.0	1.2	0.4	-0.3	-0.7	-0.6	-0.7	-0.5	2.0	-0.5	-0.7
13	NB14	0.0	-0.7	-0.7	0.3	-0.7	-0.6	-0.7	0.9	-0.3	0.5	-0.2
14	NB15	0.0	-0.6	2.3	0.0	-0.4	-0.5	0.7	0.5	-0.4	4.7	-0.3
15	NB16	0.0	-0.3	-0.5	-0.3	-0.5	-0.6	-0.6	-0.4	0.0	-0.0	0.4
16	NB17	0.4	-0.4	-0.3	0.2	0.2	-0.4	-0.2	1.5	0.5	-0.3	-0.4
17	NB18	0.1	-0.6	0.2	0.2	0.3	-0.6	-0.3	0.1	0.3	-0.5	0.8
18	NB19	0.0	-0.6	-0.6	1.1	0.7	-0.6	-0.7	-0.4	-0.1	-0.6	-0.6
19	NB20	0.0	-0.3	-0.4	-0.3	0.5	-0.6	-0.7	-0.6	-0.6	-0.3	1.2
20	NB21	0.0	-0.2	-0.2	-0.2	-0.7	-0.6	-0.7	0.6	-0.5	-0.5	0.5
21	NB22	0.0	0.7	0.5	-0.6	-0.7	0.5	-0.1	-0.1	-0.2	-0.6	-0.6
22	NB23	0.0	-0.0	-0.1	-0.0	0.8	-0.3	-0.6	-0.1	-0.4	0.1	0.0
23	NB24	0.0	0.3	-0.6	-0.5	-0.5	-0.6	0.4	-0.6	-0.1	1.3	-0.7
24	NB25	0.0	-0.5	0.6	-0.5	-0.3	-0.5	-0.7	0.6	-0.1	-0.4	-0.7
Item	Label	X^2	11	12	13	14	15	16	17	18	19	20
11	NB12	0.0										
12	NB13	0.0	0.4									
13	NB14	0.0	0.2	-0.1								
14	NB15	0.0	-0.1	-0.5	4.1							
15	NB16	0.0	-0.3	2.4	0.4	-0.4						
16	NB17	0.4	0.1	2.1	1.3	0.6	1.1					
17	NB18	0.1	0.2	1.9	-0.4	0.1	2.3	3.5				
18	NB19	0.0	0.4	-0.4	4.3	-0.2	-0.6	-0.1	-0.6			
19	NB20	0.0	1.1	-0.6	-0.2	0.1	0.4	-0.4	-0.6	-0.7		
20	NB21	0.0	0.3	-0.0	0.0	-0.3	0.7	0.0	0.9	-0.7	0.3	
21	NB22	0.0	-0.6	-0.7	-0.1	-0.7	-0.7	0.4	-0.5	-0.5	0.6	-0.7
22	NB23	0.0	-0.3	-0.1	-0.7	-0.1	0.0	-0.2	-0.2	-0.6	-0.2	-0.4
23	NB24	0.0	-0.4	-0.7	-0.7	0.8	-0.7	-0.2	-0.6	-0.2	-0.7	-0.6
24	NB25	0.0	-0.6	-0.4	1.5	0.6	-0.5	0.0	-0.2	-0.4	-0.7	5.7
Item	Label	X^2	21	22	23							
21	NB22	0.0										
22	NB23	0.0	3.1									
23	NB24	0.0	0.1	20.2								
24	NB25	0.0	4.2	5.6	4.6							

From Table 2, the S- X^2 item fit statistics indicates that all the items apart from item 6 ($p = .025$) fit the model. The overall goodness of fit model, M_2 (228) = 348.70, $p < .001$, RMSEA = .07. This suggests that though the test was significant, the RMSEA value was closer to 0, hence, the model can be deemed fit. This implies that the hypothesised model fit the observed data.

It has earlier been indicated that item 1 had a factor loading below .30, hence, it was not included in the model. As presented in Table 2, the discrimination indices for all the items range from .65 to 3.29. Fourteen (14) of the items (items; 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, and 19) had moderate discrimination indices, thus, they ranged from .65 – 1.34. Also, 6 of the items (20, 21, 22, 23, 24, and 25) had low discrimination indices, thus, .35 - .64; 2 items (6 and 13) had high discrimination indices, thus, 1.35 – 1.69. Finally, 2 items (17 and 18) had very high discrimination indices, thus, 1.70 and above (Baker, 2001). In terms of difficulty, the indices for all the items ranged from -.01 to -6.03, and this indicates that all the items functioned among the low ability levels, hence they are considered as easy items. The difficulty of an item describes where the item functions along the ability scale. An easy item functions among the low ability levels, whereas a difficult item functions among the high ability examinees. The probability of guessing ranged from .19 to .20. This means that at all ability levels, the probability of getting the item correct by guessing alone is approximately .20. These guessing parameters (.20 and .19) were less than the natural random guessing of .5. This implies that for all the items, there is a less probability of a respondent with low ability to guess.

Table-2. 3PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b).

Item	Discrimination	Difficulty	Guessing	Factor Loadings	S- X^2 Item Statistics			New λ_2
	a	*b*	*c*	λ_1	X^2	*df*	*Sig.*	
2	.65	-3.89	.20	.36	5.73	2	.057	0.42
3	.91	-3.32	.20	.47	1.43	2	.490	0.53
4	.92	-3.10	.20	.48	4.71	2	.095	0.53
5	.95	-2.02	.20	.49	10.65	6	.100	0.49
6	1.35	-1.87	.20	.62	12.81	5	.025*	0.61
7	.76	-2.37	.20	.41	7.36	6	.290	0.39
8	.99	-2.36	.20	.50	7.98	4	.092	0.49
9	.71	-5.10	.20	.39	4.17	3	.245	0.50
10	1.03	-2.70	.20	.52	2.11	1	.147	0.57
11	.82	-4.14	.20	.44	5.57	4	.235	0.39
12	.78	-2.55	.20	.42	6.30	4	.178	0.40
13	1.43	-.01	.19	.64	2.96	5	.707	0.68
14	.59	-3.57	.20	.33	3.73	1	.054	0.34
15	1.02	-3.05	.20	.52	9.82	6	.132	0.54
16	.76	-1.94	.20	.41	7.11	3	.068	0.42
17	3.29	-1.39	.19	.89	3.50	5	.624	0.89
18	1.88	-1.27	.19	.74	5.02	5	.415	0.74
19	.70	-2.53	.20	.38	2.26	1	.134	0.39
20	.50	-5.91	.20	.28**	2.18	1	.141	
21	.61	-4.58	.20	.34	9.46	5	.092	0.88
22	.37	-3.98	.20	.22**	7.77	4	.100	
23	.37	-6.03	.20	.21**	4.63	6	.593	
24	.42	-3.04	.20	.24**	3.29	6	.772	
25	.36	-3.72	.20	.21**	5.73	2	.057	

Note: Items 2 – 25 = questionnaire items (see Appendix A); **items to be discarded; M_2 (228) = 348.70, $p < .001$, RMSEA = .07.

From the 3PLM, it is evident that more than half (14) of the items moderately discriminated between high and low ability groups. However, only 6 items had low discrimination. Again, all the items were relatively easy, but the probability for guessing was small. From these results, it can be concluded all the items, apart from item 1, are good.

In addition to the aforementioned three parameters, the Response Factor Analysis (RFA) indicated that five items (20, 22, 23, 24, and 25) had factor loadings below .30, hence, they are to be deleted as recommended by Pallant (2010) and Field (2009), since they are contributing less than 9% of the variances to the construct. Based on the results, in all, 6 items have to be deleted (1, 20, 22, 23, 24, and 25). However, when the aforementioned 6 items were deleted, the new factor loadings ranged from .34 to .89. Finally, the teacher test construction skills scale was made up of 19 items, which were ideal based on the analysis. Upon careful consideration item 22 (Consider the age of learners during item writing) was retained, however, 'age' was changed to 'class level', since age was not appropriate. Again, item 7 (Subject test items to item analysis) was deleted, since item analysis is mostly used by testing agencies but not classroom teachers. The items deleted included; Items 1, 7, 20, 21, 23, 24, and 25 (Appendix

A). Finally, 18 items were deemed valid based on the IRT results (see Appendix B). This, therefore, constitutes the final version of TCSI. Figure 1 and Figure 2 present the test characteristic curve and test information curves.

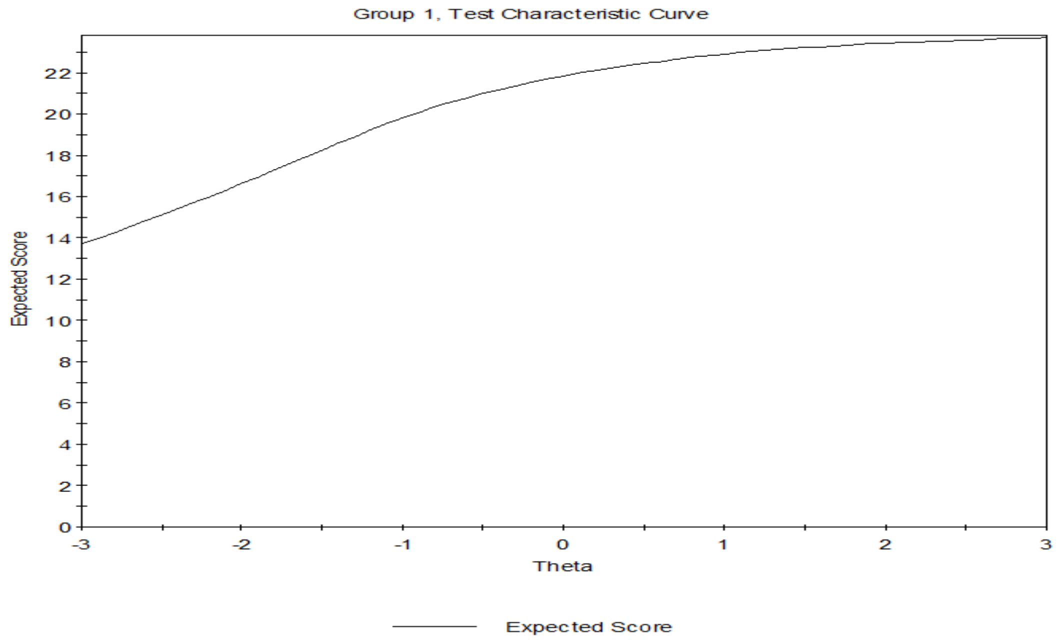

Figure-1. Test characteristic curve.

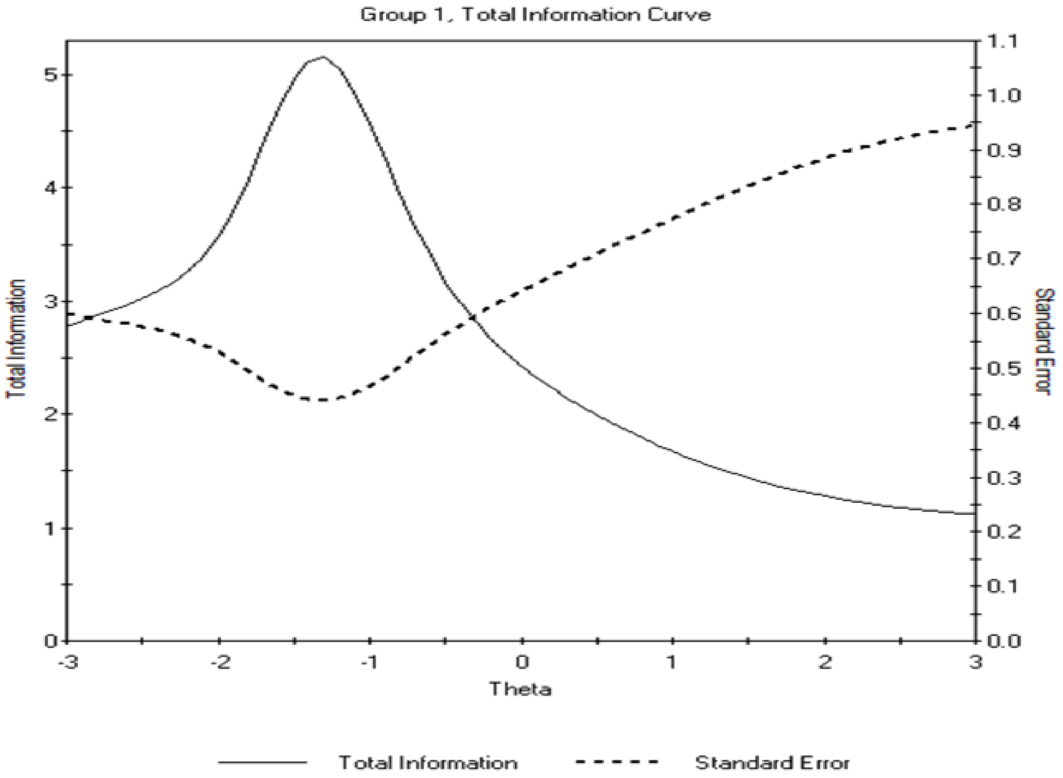

Figure-2. Test information curve.

The test characteristic curve presents a functional relation between the true score and the ability scale. At an ability level of 1.0, the true score is 22 Figure 1. The test characteristic curve increases monotonically. The test information curve reached its peak at an ability level of -1.2 with maximum information of 5.1 and a standard error of 1.06. Within ability range of -1.5 to -0.5, the test information was greater than 3.0. In all, the test measures the ability with unequal precision Figure 2.

4.2.1. Scoring and Interpretation

Each of the 18 items should be scored dichotomously as '1' or '0', where '1' will be awarded for correct response and '0' for an incorrect response. Afterwards, respondents' scores for all the 18 items should be summed. In all, the scores of each respondent will range from 0 to 18. Scores from 0 – 8.9 will be described as low level of skills (thus, below 50%), scores from 9 – 14.3 (thus, 50% to 79.9%) will be described as a moderate level of skills, while scores 14.4 – 18 (thus, 80% or more of the total scores) will be described as a high level of skills in test construction. Finally, respondents' scores should be compared with the various score ranges to determine the level of skills in test construction. The new instrument is named TCSI-18.

4.3. Objective 2: To assess teachers' Level of Skills in Test Construction

The second objective examined teachers' level of skills in test construction. The validated version of TCSI was used to assess test construction skills among senior high school teachers in the Cape Coast Metropolis. Table 3 presents the results on the second objective.

Table-3. Distribution of teachers by level of skills in test construction.

Level of skills	Score range (out of 18)	Frequency	Percentage (%)
Low	0 – 8.9	117	20.1
Moderate	9 – 14.3	197	33.8
High	14.4 – 18	269	46.1
Total		**583**	**100.0**

From Table 3, it can be seen that majority (46.1%) of the teachers had higher skills in test construction, 33.8% also had moderate skills in test construction, while 20.1% had a lower level of skills in test construction. It can, therefore, be said that respondents, generally, have higher skills in test construction. This result could emanate from the fact that the majority of the teachers reported they were professional teachers and had taken at least a course in assessment or measurement and evaluation.

5. DISCUSSION

The study revealed that teachers had high skills in test construction. Nearly half of the respondents had a score of at least 14.4 out of 18 in test construction. This result has implications on teachers' assessment of students' learning, and the effectiveness of teaching and learning, at large. High skills in test construction put teachers in a better position to deliver diligently up to the task. With this, teachers are capable of engaging in proper assessment of students' learning. Teaching/learning and assessment are complementary ways of describing the same activity. That is to say, teaching without assessment is not effective, likewise, assessment without teaching is baseless, and thus, teachers assess what they have taught. High skills of teachers in test construction presuppose that teaching and learning are effective, even though that was not the position of this current study, and also not within its mandates. As indicated earlier, for classroom teachers, the assessment provides them with much information tailored to the entire teaching and learning activities. Through assessment, teachers can determine whether a particular lesson has to be retaught, probably, with the same or different method. Information on assessment provides feedback to students on which segments/topics they have mastered, and where they are deficient.The findings of this study, that teachers have high skills in test construction is in harmony with a couple of studies

(Adamu et al., 2015; Afemikhe & Imobekhai, 2016; Agu et al., 2013). The findings of the aforementioned studies equally found a high level of test construction skills among teachers in Nigeria, even though their approach was quite different from the approach of the current study. The previous authors used Likert-type scales in measuring teachers' skills in test construction. In strict technical sense, a higher score on such a Likert-type scale depicts what teachers will do rather than what they can do. That is to say, in the case of Adamu et al., for example, the teachers strongly agreed to all the 26 items on test construction. In principle, agreement to such statements, in any way, does not tell what teachers know as far as test construction is concerned. Perhaps, agreement to such statements could be teachers' perception or opinion, which are not necessarily their skills in test construction. The finding of the current study is unique as it clearly articulates what teachers can do in terms of test construction.

The finding of the current study, however, disconfirms some earlier studies (Hamman-Tukur & Kamis, 2000; Koksal, 2004; Marmah & Impraim, 2013; Quansah et al., 2019). The authors found that teachers do not have adequate skills in test construction. Interestingly, the authors, in their quest to investigate test construction skills, evaluated samples of the test developed by teachers. Indeed, one would agree to such an approach as it directly deals with work samples of the teachers. The issues, however, are: first, whether those tests were truly developed by the teachers solely and not in collaboration with other teachers; and two, were the questions collated from past questions which have duly been crafted? In my candid opinion, such approaches, though appear promising, there are other factors which could explain results away from such approaches. In contrast, with the finding of the current study, teachers' skills were assessed and described appropriately. We do not, in any way, dispute the findings of the previous studies, we, however, evaluated their findings within the context of their study. This study, in a way, addresses the inefficiencies in assessing test construction skills.

6. CONCLUSION AND RECOMMENDATION

It is important to state that the measurement of psychological constructs such as intelligence, skills, ability, knowledge, competence, and aptitude, among others are tests of maximal/optimal performance, however, the use of Likert-type scales for such constructs are not well-placed. Likert-type scales are useful in the measurement of typical behaviours such as interest, personality, feelings, and temperaments, among others. It is worthy to note that Likert-type scales are appropriate in the measurement of typical rather than maximum behaviours.

The study revealed that teachers had a higher level of skills in test construction, and this could be due to the educational background of teachers. It is recommended that the Ministry of Education, Ghana (MoE), Ghana Education Service (GES), and heads of the various SHSs in the Cape Coast Metropolis, as part of their training programmess or workshops for teachers should continue and intensify the acquisition of skills by teachers in test construction to improve their skills.

7. LIMITATIONS

It is worthy to state the items on the scale were only a sample of behaviours in test construction process. Therefore, interpretation of the results should be done with caution. In addition, the validated test construction scale is not a perfect measure of test construction skills, other methods could be used to complement.

REFERENCES

Adamu, G. G., Dawha, J. M., & Kamar, T. S. (2015). A scheme for assessing technical teachers' competencies in constructing assessment instruments in technical colleges in Gombe State. *ATBU, Journal of Science, Technology & Education (JOSTE)*, *3*(2), 1-8.

Afemikhe, O. A., & Imobekhai, S. Y. (2016). *Nigerian teachers' utilization of test construction procedures for validity improvement of achievement tests*. Unpublished Paper, Institute of Education, University of Benin, Benin City, Nigeria.

Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Educational Research and Reviews*, *8*(8), 431-439.

Ali, A. A. (1999). *Basic research skills in education*. Enugu: Orient Printing and Publishing.

Amedahe, F. K. (1993). Test construction practices in secondary schools in Central Region of Ghana. *The Oguaa Educator*, *2*, 52-63.

Amedahe., F. K. (2014). *The issue of falling educational standards in Ghana: A perception or reality?* Cape Coast: University of Cape Coast Press.

American Federation of Teachers National Council on Measurement in Education & National Education Association. (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: Author.

Anhwere, Y. M. (2009). *Assessment practices of teacher training college tutors in Ghana*. Unpublished Master's thesis, University of Cape Coast, Cape Coast.

Ankomah, F. (2019). *Predictors of adherence to test construction principles: The case of senior high school teachers in Sekondi-Takoradi Metropolis*. Unpublished masters' thesis, University of Cape Coast, Ghana.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Original Work Published in 1985

Baker., J. O. (2003). *Testing in modern classrooms*. London: George Allen and Urwin Ltd.

Chan, K. K. (2009). *Using test blueprint in classroom assessments: Why and how*. Paper presented at the Paper Presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference, Brisbane, Australia.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159-179.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York: Cengage Learning.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper & Brothers, Publishers.

Cronbach., L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Ebinye, P. O. (2001). Problems of testing under the continuous assessment programme. *Journal of Quality Education*, *1*(1), 12-19.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE Publications Ltd.

Hamafyelto, R., Hamman-Tukur, A., & Hamafyelto, S. (2015). Assessing teacher competence in test construction and content validity of teacher made examination questions in commerce in Borno State, Nigeria. *Journal of Education*, *5*(5), 123-128.

Hamman-Tukur, A., & Kamis, A. B. (2000). Content analysis of B.sc Bio-chemesrty examination questions implication for testing, teaching and development. *African Journal Research in Education*, *1*, 59-62.

Jiraro, S., Sujiva, S., & Wongwanich, S. (2014). An application of action research for teacher empowerment to develop teachers' test construction competency development models. *Procedia-Social and Behavioral Sciences*, *116*, 1263-1267.Available at: https://doi.org/10.1016/j.sbspro.2014.01.380.

Koksal, D. (2004). Assessing teachers' testing skills in ELT and enhancing their professional development through distance learning on the net. *Turkish Journal of Distance Education*, *5*(1), 1-11.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 5-55.

Marmah, A. A., & Impraim, A. K. (2013). University lecturer's competence in the construction of multiple choice test items: A case study of Coltek-Kumasi. *The International Journal of Humanities & Social Studies*, *1*(4), 1-9.

Miller, M. D., Linn, R. L., & Gronlund, N. F. (2009). *Measurement and assessment in teaching* (10th ed.). Ohio: Pearson Education, Inc.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology (pp. 407-423). New York: Guilford.

Nitko, J. A. (2004). *Educational assessment of students* (4th ed.). New Jersey: Prentice Hall.

Ololube, N. P. (2008). Evaluation competencies of professional and non-professional teachers in Nigeria. *Studies in Educational Evaluation, 34*(1), 44-51.Available at: https://doi.org/10.1016/j.stueduc.2008.01.004.

Onyechere, I. (2000). *New face of examination malpractice among Nigerian youths*: The Guardian Newspaper.

Pallant, J. (2010). *A step by step guide to data analysis using the SPSS program: SPSS survival manual* (4th ed.). Crows Nest: Allen & Unwin.

Quaigrain, A. K. (1992). *Teacher competence in the use of essay tests: A study of secondary schools in the Western Region of Ghana.* Unpublished Master‘s Thesis, University of Cape Coast, Ghana.

Quansah, F., Amoako, I., & Ankomah, F. (2019). Teachers' test construction skills in senior high schools in Ghana: Document analysis. *International Journal of Assessment Tools in Education*, *6*(1), 1-8.Available at: https://doi.org/10.21449/ijate.481164.

Simon, G. M. (2002). *Testing to destruction: A problem in a small state.* Unpublished Seminar Paper.

Spearman, C. (1904). Measurement of association, Part II. Correction of ‘systematic deviations’. *The American Journal of Psychology 15*, 88-101.

Appendix A – Initial Scale

No.	Items	True	False
	A teacher should take the following steps in constructing tests for his/her class		
1.	Give clear instructions to guide the test takers.		
2.	Write test so that both high and low achievers can understand.		
3.	Avoid gender stereotypes in the test items.		
4.	Ascribe scores for each test item.		
5.	Avoid too long questions or phrases in item writing.		
6.	Outline the content covered for the term before setting test from them.		
7.	Subject test items to item analysis.		
8.	Add enough test items to cover all the requisite levels of cognitive domain.		
9.	Ensure that the items are measuring the determined objectives.		
10.	Set essay items that elicit creative and imaginative answers from the students.		
11.	Prepare a marking guide while constructing the test.		
12.	Add sufficient items to cover the appropriate instructional units.		
13.	Limit essay tests to high level objectives.		
14.	Organize test items in a logical manner.		
15.	Set tests with due regard to the time available for testing.		
16.	Avoid the use of clues in multiple choice questions.		
17.	Avoid the use of overlapping items.		
18.	Avoid overlapping alternatives in writing objective tests.		
19.	Prepare a test blueprint as a guide in the test construction.		
20.	Consult standard text books in the subject for guide.		
21.	Keep a resource bank of questions that can be used to when setting tests.		
22.	Consider the age of learners during item writing.		
23.	Submit items for vetting to the Head of Department or the principal.		
24.	Submit tests meant for promotional examinations for expert editing on time.		
25.	Review draft of the test at least twice in two days before administering.		

Appendix B – Final Scale (TCSI-18)

Please read carefully the following statements and indicate to the best of your knowledge whether the statements are true or false.

No.	Items	True	False
	A teacher should take the following steps in constructing tests for his/her class		
1.	Write test items so that both high and low achievers can understand.		
2.	Avoid gender stereotypes in the test items.		
3.	Ascribe scores for each test item.		
4.	Avoid too long questions or phrases in item writing.		
5.	Outline the content covered for the term before setting test from them.		
6.	Add enough test items to cover all the requisite levels of cognitive domain.		
7.	Ensure that the items are measuring the determined objectives.		
8.	Set essay items that elicit creative and imaginative answers from the students.		
9.	Prepare a marking guide while constructing the test.		
10.	Add sufficient items to cover the appropriate instructional units.		
11.	Limit essay tests to high level objectives.		
12.	Organize test items in a logical manner.		
13.	Set tests with due regard to the time available for testing.		
14.	Avoid the use of clues in multiple choice questions.		
15.	Avoid the use of overlapping items.		
16.	Avoid overlapping alternatives in writing objective tests.		
17.	Prepare a test blueprint as a guide in the test construction.		
18.	Consider the class level of learners during item writing.		

__Online Science Publishing__ is not responsible or answerable for any loss, damage or liability, etc. caused in relation to/arising out of the use of the content. Any queries should be directed to the corresponding author of the article.