CHAPTER ONE

INTRODUCTION

Background to the Study

There is a regular administration of public examinations done at the end of different levels of education, for example, administration of Basic Education Certificate Examination (BECE) which is carried out at the end of junior secondary school and Senior School Certificate Examination (SSCE) is administered to candidates at the conclusion of senior secondary school. These are conducted by different examination bodies in Nigeria. Candidates have preference for these examination bodies out of the notion that one may be superior to others. Employers of labour and educational institutions also behave similarly. These impressions made test experts find equating necessary.

In Nigeria, there are three main examination bodies that conduct public examination. They are: West African Examination Council (WAEC), National Examination Council (NECO) and National Business and Technical Examination Board (NABTEB). All these examination bodies present Senior School Certificate Examination (SSCE) to candidates. Their syllabi are prepared from the approved curriculum produced by the National Educational Research and Development Council (NERDC) for Senior Secondary Schools. This guides the teaching and learning processes in the classroom. Each of these examinations can be referred to as different test forms. The items of these test forms are required to be of almost the same content and should be able to examine uniform skill and knowledge. Despite the fact that all the forms are expected to cover the same content, test the same skill and knowledge, there is likelihood that one of the forms will be more difficult than the other due to differences that exist in abilities of examinees taking the different test forms. A statistical process called equating can, therefore, be used to

adjust for differences in difficulty across alternate forms resulting in comparable score scales and more accurate estimates of ability (Albano, 2011).

Kolen and Brenann (2004) mentioned that equating control score stability from one test administration to another if the test is given annually or in different forms. Mainly, standardized tests are conducted annually, they are prepared by experts and administered by an examination body. Examination, being a valid and reliable test, used for accurate measurement of students' performances and also to make determinations regarding certification (Kolawole, 2001), is administered by examination body before equating can take place.

The three examination bodies administer questions in most subjects at SSCE level in two parts; the objective paper and the essay paper, while science subjects involve practical examination as the third paper. Chemistry is one of the core science subjects that has being offered since the inception of SSCE. Chemistry at SSCE level consists of two papers, paper I and II. Paper I is the practical aspect of the examination while paper II is further divided into two: multiple choice format and the essay format. Candidates who register for Chemistry at SSCE must attempt these tests (i.e practical, essay and multiple-choice tests) for assessment purpose. Chemistry syllabus supplies information and skills needed to make students ready for higher cognitive learning of chemistry, it helps to prepare students who want to study science related courses at the higher institution. Chemistry related specialist studies like biochemistry, medicine, geology, pharmacy, industrial chemistry, pure chemistry and engineering require at least credit level pass in Chemistry. Taking into account the usefulness of Chemistry in the technological development of the nation, it is expected of students offering it to have more interest in the subject and perform better. Available reports of Nigerian secondary school students' performance in Chemistry are unsatisfactory (WAEC, 2009 & Udo, 2008).

Table 1: Number and Percentages of Students who Obtained Grades 1- 6 in West African Senior School Examination May/ June in Chemistry from 2007 - 2016

Year	Total entry	Credit passed (A1 – C6)	% Passed
2007	424,747	196,063	46.16
2008	456,980	202,762	44.37
2009	456,980	203,365	43.49
2010	465,643	263,059	50.70
2011	565,692	280,280	49.54
2012	627,302	270,570	43.13
2013	649,524	460,470	72.05
2014	652,809	399,062	61.88
2015	665,527	457,979	69.59
2016	645,740	531,360	82.92
2017	709,404	590,629	83.91
2018	732,508	423,451	58.17
2019	747,075	572,044	77.02

Source: West African Examination Council, Yaba, Lagos. 2016.

Results in Table 1 show that there is a continuous increase of the number of students who sat for WAEC Chemistry from 2007 – 2019 and the inconsistency of their performance. In 2012, the students had the least performance at credit level, out of 627,302 students that sat for the examination 270,570 (43.13%) passed at credit level and above while year 2019 had the highest performance of credit level pass. Those that sat for the examination were 747,075 students out of which 572,044 (77.02%) passed. There is inconsistency in the performance of students in WAEC Chemistry between years 2007 and 2019.

Chemistry SSCE syllabus prepared by the three examination bodies (WAEC, NECO and NABTEB) always influence the content area of what is taught in schools. This is derived from the curriculum produced by the National Educational Research and Development Council (NERDC). It reflects in the mode and type of examination questions set by the examination

bodies which are expected to measure the same construct. It is almost impossible to construct various test forms that are accurately parallel. In order to compare the scores obtained from the different test forms, a statistical process called equating can be applied.

Equating is a statistical process used in situations where multiple forms of a test exist, the test forms should be constructed according to the same content and statistical specifications and should be administered under similar conditions. Albano (2010) defined equating as a statistical procedure commonly used in testing programmes where administrations across more than one occasion and more than one examinee group can lead to overexposure of items, threatening the security of the test. It has to do with changing of the units of different test forms of the same test once the tests are parallel. The purpose of test equating is to compare test scores of various test forms administered to testees. Ryan (2011) also defined equating as a technical procedure or process conducted to establish comparable scores, with equivalent meaning, on different versions of test forms of parallel test. In equating, test scores are slightly changed on the basis of the difficulty of the test forms given to testees. Likewise, Wendy (2002) described test equating as a statistical method for measuring and controlling for variations in the difficulty of different tests so that scores from equated tests have comparable meaning.

Test equating process makes it possible for the test users to interchange multiple forms of a test (Sharon, 2005). The process of equating enables the scores from one test form to be expressed based on the scores from the other form (Dorans & Holland, 2000; Van der Linden, 2006). There are various methods of test score equating. Some of the methods include: Mean Equating, Linear equating, Levine linear equating, Tucker linear equating, Chained linear equating, Equipercentile equating, Frequency estimation equipercentile equating, Chained

equipercentile equating, One parameter logistic model (Rasch), Two parameter logistic model and Three parameter logistic model (separate and concurrent calibration).

There are two approaches that can be used in equating different forms of test: Item Response Theory (IRT) and Classical Test Theory (CTT). Linear and Equipercentile methods of equating are CTT equating approaches. Dorans (2000) described linear equating as a CTT equating approach that provides a transformation in such a way that scores from two or more tests will be considered equated if they correspond to equal standard score deviates. Linear equating is a linear relationship that exists between scores obtained from two test forms taken by a group of test takers, if they are the same number of standard deviations below or above the mean of the group. That is, a score on the new form and a score on the reference form can be said to be equivalent in a group of test-takers if they are the same number of standard deviations above or below the mean of the group.

Linear equating can be suitably used when the different test takers that will be given the different test forms to be equated have the same abilities. Levine and Tucker equating methods utilise linear methods of equating under the common items non-equivalent groups design. Kolen and Brennan (2004) suggested that Levine linear equating method will be more appropriate to use when the group of examinees attempting the different test forms to be equated have varying abilities. The assumption of Levine equating is that the test items in the different test forms are randomly parallel to the set of items to be equated and the students' abilities are statistically different (Kolen and Brennan, 2004). Levine equating method is preferred when it is known that populations differ substantially and if there is a justification that the test forms are similar (Kolen & Brennan, 2004). Also, it is often compared with other nonlinear methods because under

certain circumstances, it is more accurate than other linear equating methods (Mroch, Suh, Kane & Ripkey, 2009).

Livingston (2004) described equipercentile equating as an equating method used to equate scores obtained from testees on the new form to scores obtained from testees on the reference form. Scores from the new form are to be transformed to the scores on the reference form that have the same percentile rank in that group. The main concern in equipercenitle equating is to find a score y on form Y that has the same percentile rank as a score x on form X. Other common equipercentile-like equating methods are frequency estimation equipercentile and chained equipercentile methods (Hou, 2007). Chained equipercentile equating method has been found applicable under common items non-equivalent groups design. Chained equipercentile equating method involves conversion of scores which are chained together to yield a conversion of form X scores to form Y scores; form X scores are converted to scores on the common items in population 1 using equipercentile equating method. Then the set of scores from the common items are equated to form Y scores using test takers from population 2.

Data are required to be collected using specific designs before equating procedures can be carried out (Albano, 2010). The four designs that are commonly used in practice are single group design, random group design, equivalent group design and Non-equivalent groups Anchor Test design (NEAT) also called Common Items Non-equivalent Groups (CINEG) design. NEAT design includes common/ anchor items and unique items in each test form to be equated. Common items are items embedded in the different test forms that are to be equated and these are used to contrast the groups of testees assessing the same skills and abilities that the test measures (Chen, Huang & MacGregor, 2009). Sonya (2011) stated that scores of anchor items are used to make adjustments for differences in test form difficulty, taking into account

differences in group performance. Unique items are items that are peculiar to each test form to be equated.

This study therefore compared scores that were obtained from 2017 WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers. Hence the study equated scores from these different forms of tests, such that no testee was disadvantaged or advantaged as a result of the form of test taken. As stated by Peterson, Marco, Stuart and Ord (1994), the process of equating is used to ensure that scores resulting from the administration of the multiple forms can be used interchangeably.

Statement of the Problem

There are various forms of certification examination bodies that make use of multiple forms of tests. Each of the test forms has different items, but each form is supposed to measure the same thing. The major examination bodies responsible for certification at the senior secondary level in Nigeria have curriculum content that are similar in nature. Scores obtained from these tests should be comparable so as to ensure uniform standard, consistency and fairness.

Researchers have different views about each examination body. It has been observed that candidates, educational institutions, employers of labour and other end users have preference for certificates of certain examination bodies when compared to others (Olatunji, 2015). For some time now, there has been prevalent criticism of examination bodies in Nigeria among some institutions and employers of labour as some of them prefer candidates with credit passes in the SSCE conducted by WAEC to those conducted by NECO and NABTEB. There are misconceptions about the quality of examinations conducted by the three examination bodies. Kpolovie, Ololube and Ekwebelem (2011) mentioned that some Universities in Nigeria and

abroad denied candidates with NECO certificate based on speculations about their integrity. Bandele and Adewale (2013) on the other hand submitted that WAEC, NECO and NABTEB are comparable and equivalent when the coefficients of reliability and validity of Mathematics achievement examination conducted by the three examination bodies were compared. Therefore, it is imperative to conduct a study by conversion of units of WAEC, NECO and NABTEB so that scores obtained from them could be directly compared.

In comparing the quality of these examination bodies, Levine linear and chained equipercentile equating methods are highly recommended, especially when the groups of test takers that the different test forms will be administered to are from population dissimilar in their ability level (Kolen & Brennan, 2004; Holland, von Davier, Sinharay & Han, 2006). Using these equating methods helped to determine the equivalence of scores obtained from the different test forms (WAEC, NECO and NABTEB) in this study.

There are several studies that have been carried out in this regard. Some of the studies include that of Adewale (2016) who equated two years BECE results in Basic Science and Technology in Oyo State using linear and equipercentile equating methods. Olatunji (2015) equated candidates' scores in 2009 WAEC, NECO and NABTEB SSCE Economics among students in Kwara State using linear and equipercentile equating methods. Findings from the studies showed that linear equating is more robust. Adokoniyi (2015) also carried out a study on equating of the multiple-choice Kwara State joint senior secondary school Economics mock papers using mean, linear and equipercentile equating methods under non-equivalent groups anchor test design. None of these researchers compared both Levine linear and chained equipercentile methods in their studies.

Alfred (2011) carried out a research on assessment of the equivalence of 2007 SSCE multiple choice Economics test items in Ilorin and found out that there is a significant difference in the difficulty levels of Economics multiple choice items conducted by WAEC, NECO and NABTEB. Wang, Lee, Brennan and Kolen (2006) compared chained equipercentile and frequency estimation method with common-item nonequivalent group (CINEG) design using a 60-item mathematics test data from four test forms. Also, Liou, Cheng and Li (2001) carried out a study on equating of two forms of a Geography test administered to nonequivalent groups with common items using Tucker, Levine, chained equipercentile, and frequency estimation equipercentile methods. These researchers took chained equipercentile and Levine linear methods into account in their studies but a search through literatures revealed that no work has been done on equating scores on SSCE Chemistry multiple-choice papers using Levine linear and Chained equipercentile equating methods under non-equivalent groups anchor test design which are best used for substantially different populations.

Most previous researchers who had equated students' scores had used linear and equipercentile methods which have been found appropriate when population of testees have the same abilities. Studies have suggested that the two equating methods (Levine and Chained equipercentile) that were used in this study are better choices for equating when the population differs substantially (Kolen & Brennan, 2004; Holland, et al, 2006 and Wang, Lee, Brennan, & Kolen, 2008). These constitute the gap this research filled. WAEC, NECO and NABTEB SSCE of Chemistry students' scores in South–west Nigeria were equated, giving room for scores obtained from these examination bodies to be compared. The study was also able to find out which of the two equating methods, that is, Levine linear equating and Chained equipercentile equating methods, is better to use for equating.

Purpose of the Study

The general purpose of this study was to equate the multiple-choice Chemistry papers of WAEC, NECO and NABTEB Senior School Certificate Examination (SSCE) using Levine linear and Chained equipercentile equating methods.

Specifically, this study investigated:

a. profile of students' performance on the common items of SSCE Chemistry multiple-choice papers
b. profile of students' performance on the unique items of SSCE Chemistry multiple-choice papers
c. results of Levine linear equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using standard score deviates
d. results of chained equipercentile equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using percentile ranking
e. invariance of equated scores of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers across equating methods

Research Questions

The following research questions were answered in the study:

a. What is the profile of students' performance on the common items of SSCE Chemistry multiple-choice papers?

b. What is the profile of students' performance on the unique items of SSCE Chemistry multiple-choice papers?
c. What are the results of Levine linear equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using standard score deviates?
d. What are the results of chained equipercentile equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using percentile ranking?
e. How invariant are the equated scores of WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers across equating methods?

Scope of the Study

This study was carried out among public Senior Secondary Schools in South–West Nigeria and the study sample comprised all Senior Secondary three students in the geo-political zone. The study equated WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers. This paper is compulsory for all science candidates at the senior secondary school level in Nigeria. The main aim of this study was to investigate the analysis of two equating methods, that is, Levine linear equating and chained equipercentile equating methods to equate WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers. The 2017 WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers were used as instrument to obtain data for the study. Data obtained was analyzed using descriptive statistic, that is, mean, t-score, percentile ranking and coefficient of variation statistics.

Operational Definition of Terms

The following are the major terms and variables as they are used in this study in order to avoid ambiguity.

Test Forms – these are different versions of Senior School Certificate Examination (SSCE) (WAEC, NECO and NABTEB) whose scores were considered interchangeable, they are intended for the same purpose and were administered in the same way.

Chemistry Multiple Choice Test- this refers to 2017 WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers that were responded to by science students of Senior Secondary Schools in South-West, Nigeria.

Levine linear Equating – this is a process of transforming students' scores in WAEC, NECO and NABTEB Chemistry Multiple Choice items in order to equate the scores as they conform to equal means and standard deviations.

Chained Equipercentile Equating – in this study, chained equipercentile equating is a process of equating students' scores in WAEC, NECO and NABTEB Chemistry multiple choice items through scores that were obtained from common items as they conform to the same percentile rank

Unique Items – these are test items that belong exclusively to each test form in this study. They appeared on the upper and lower parts of each test form

Common Items – The set of items that commonly appeared on the test forms that were used in this study, which serves as link to connect the test forms together onto a common scale. They are also anchor items which appear in about the same positions and exactly same way in all the test forms.

Significance of the Study

This research helped to analyze how equivalent Chemistry Senior Secondary Certificate Examination is. The findings from this study therefore, are of great importance to examination bodies, evaluators, teachers, students, parents and examination administrators. This study must therefore provide information that would help in comparing the quality of the three examination bodies in Nigeria – WAEC, NECO and NABTEB and could further encourage higher standards across the examination bodies.

It is hoped that the results of this study might help to indicate the overall quality and standard of certificate examination bodies in Nigeria at the senior secondary level. Teachers, therefore, will be able to counsel students appropriately and correct their wrong impression on the quality of certificate examinations, so that they might have more confidence and trust in the grades obtained from Senior Secondary Certificate Examination in Nigeria. Parents might also benefit from the outcome of this study. It might enable them to ascertain the quality of WAEC, NECO and NABTEB so as to encourage their children or wards to take these certificate examinations serious. The outcome of this study might also give information to students about the equivalence of Senior Secondary Certificate Examination and boost their confidence in the quality of the certificates that can be obtained from examination bodies in Nigeria so as not to see any of these examination bodies as inferior.

This study might be useful to examination administrators who need guide on equating test forms using Levine linear and Chained equipercentile equating methods and might also provide test developers with empirical information on equivalence of senior secondary certificate examination and also to be able to address the issue of comparability of quality among the examination bodies. It is hoped that this study might help educational policy makers more in planning of educational programmes and taking important decisions like equating test forms to

complement test development and score reporting processes. It might also help in supporting item bank development. Also, it will be of benefit to the researcher who intends to publish the results in a professional education journal as well as providing feedback to the participating schools. Finally, as the importance of certificate examination cannot be over emphasized the outcome of this study might be of great assistance to other educational researchers who would want to carry out a similar study in other subjects and in other states of the nation. Apart from providing point of reference to future research, the findings or results of the study would be complementary to existing knowledge in education.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

This chapter reviewed related literature and was discussed under the following sub – headings:

a. Standardization of Tests
b. Test Scores Equating and Methods of Equating
c. Levine Linear Equating
d. Chained Equipercentile Equating
e. Equating Designs
f. Test Equating in Public Examinations
g. Equating of Senior School Certificate Examination Chemistry Scores
h. Theoretical Framework
i. Appraisal of the Reviewed Literature

Standardization of Tests

When test equating is used as a means for comparing students' score, standardized test is required (Agah, 2013). For equating procedure to be acceptable, testing conditions should be standardized (Kolen & Brennan, 2014). A standardized test as defined by Burrows (2016) is a test that is administered to students in a very consistent manner, that is, the questions on the test items are all the same, the time allocated to each student is the same, and the manner scoring the test is uniform for all students. It helps to give a standard and fair evaluation to all students. Roles of standardized test have been summarized by Barett (2011):

i. They must measure the same skill every time, which means using test items that are designed according to certain standards, patterns, and rules

ii. Questions with correct answers that are beyond dispute must be included, and they must be able to apply exactly the same grading standards to a potentially infinite number of test-takers.

iii. They must make use of "norm-referenced" scoring procedure

Standardized tests are basically used for certification and are conducted by established examination bodies. In Nigeria, WAEC, NECO, NABTEB and JAMB are the main examination bodies that conduct standardized tests. They can be referred to as national examination. It is a test conducted annually to communicate valuable information about students' achievement status to decision maker (Stiggins, 2008). Data obtained from a standardized test should be of high quality in relation to its validity and reliability. Wragg (2001) mentioned that the data or information obtained from national examinations should be accurate, valid, reliable and of high quality. For a test to be said to be reliable, it means its scores are precise and consistent. Reliable test scores are precise and they contain little measurement error (Suen & McClellan, 2003). Validity is the extent to which evidence supports interpretation of test scores and decisions (American Educational Research Association, 1999)

In order to ensure that the test outcome is as valid as possible, there should be standard administration of the test. The test should therefore be;

i. Written at the same time and same day for every student,

ii. Administered with consistent instructions,

iii. Allowed the same amount of time for each student to write the test, and

iv. Scored in the same manner (Poulsen & Hewson, 2013).

Another way to ensure that a test meets the criteria for validity and reliability is through procedure for test construction. When constructing a test, there are a number of principles that

should guide a test writer. Among authors who have written on steps on constructing a standardized test, Abiri (2007), Pandeya (2015) and Sharma and Poonam (2017) identified different steps in test construction. They are:

1. Determine the purpose of the test – objectives of the test should be determined and the test content, extent of coverage, item types and level of difficulty of the test should be determined
2. Planning the test
    i. Determine the test format to be used – the test format to be used should be decided, items of different format should be mixed together but items that have similar format should be grouped together
    ii. prepare a table of specifications – table of specification also known as test blue print is a two-way table that keeps in view the content area and objectives of learning as per Blooms taxonomy of educational objectives. It also shows the proportion of test content that are to be sampled by the test items.
3. Writing the test items – writing of test items requires the following:
    i. The writer must have complete mastery over the subject matter
    ii. The content area and the objectives of learning should be adequately covered as laid down in the table of specification
    iii. Vocabulary used in the items should be simple enough to be understood by the intelligence level of those the test is meant for
    iv. The items should be arranged properly from the easiest items to the most difficult ones
    v. Instructions, time limit and score allotment should be clearly stated

vi. Test items should be given to a group of experts in the subject matter in order to remove vagueness, ambiguity and language difficulty

4. Carry out a preliminary/ trial run of the test – after considering experts' suggestions, the trial form of the test should be administered to a fairly representative sample of the target population (those for whom the test is aimed at) and the test will be scored. This helps to find out the weaknesses and inadequacies of the items, non-functional distracters in multiple-choice tests and very difficult or very easy item.
5. Carry out an item analysis – results obtained from trial run of the test are used for item analysis in order to establish item difficulty, item discrimination and effectiveness of the distracters in multiple-choice items. Before these can be determined, the scores of testees should be placed in order of magnitude and then divided into three equal groups. The upper and lower groups are the ones involved in the analysis ignoring those in the middle group.

    i. Item difficulty - this is the percentage of testees who correctly respond to a test item, it ranges between 0% and 100%. The higher the value of an item the easier is the item. Items with values above 90% are very easy items while items with values above 20% are very difficult items. Moderately difficult items are items whose values are between 20% and 90% and they are most preferred to the very easy and very difficult items.

    ii. Item discrimination – this is the extent to which students with varying levels of achievement perform differently on an item. It discriminates between weak and strong testees. Discrimination power of an item ranges from -1.00 to 1.00, any item with negative discriminating power should be disqualified. The higher the

value the more discriminating the item. Items with discriminating power of 0.40 and above are quite satisfactory while below those below 0.40 will need amendment.

iii. Effectiveness of the distracters – all options (both the key and incorrect options) are expected to function effectively in each multiple-choice item. That is, each option should be picked by at least one person from each group. It is expected that each incorrect option should be picked more frequently by those in the lower group which is contrary to the case with the key

6. Compiling the final test items – results obtained from item analysis will help in amending items that need amendment and refuse items that have been found to be unsuitable. Test items that have satisfied the requirements for constructing standardized test can then be selected and compiled for final use. Scoring methods, time limit for the test and other instructions regarding the test should be clearly written.

Test Scores Equating and Methods of Equating

In educational measurement, test scores are as important as the test itself. Test scores are set of figures that express students' performance on a test and provide information on which important decisions can be made. Decisions can be made by students as to what course to be studied, or by school, such as choosing a cut off mark for admitting students to study particular courses. Test scores are, therefore, expected to be as accurate as possible. Accurate test scores help in making fair and consistent decisions on examination results especially in standardized examinations. Standardized tests often occur in more than one edition, that is different test forms

with similar statistical characteristics and content are constructed. Ryan (2011) defined test form as a collection of test questions or tasks assembled, published, and administered to examinees.

Test equating allows interchangeable use of alternate test forms that have been built to the same content and statistical specifications (Haertel, 2004). Despite the efforts made by test developers to construct identical or similar tests in terms of statistics and of content, differences are bound to occur in test difficulty as long as different test forms and different test items are used in a session (Tanguma, 2000). Test developer tries as much as possible to adhere strictly to test specification so as to produce different forms of test that are similar in difficulty, this is almost never possible, as each test form contains different questions. Also, different circumstances make it a necessity for different students to be measured with different instruments which are intended to measure the same trait. One of the reasons why different test forms are administered at different times is due to security problem (Asiret & Sunbul, 2016). Some tests are specifically designed for application with a particular population of respondents. For example, a population of respondents with high proficiency level will be administered test with more difficult items. The psychometric and statistical characteristics of tests differ depending on the characteristics of the population they were designed for. Scores obtained from different tests and test forms cannot be directly comparable. These and many other reasons bring about the use of a statistical procedure called equating to adjust scores on different test forms so that the scores can be used interchangeably.

von Davier, Holland and Thayer (2016) defined test equating methods as method used to produce scores that are comparable across different test forms. The process of equating is used in situation where there are multiple forms of test and examinees taking different forms are compared to each other. Despite the fact that test forms are designed based on the same

specification, to cover the same content, at the same level of difficulty, the test forms turn out not to be only identical. Albano (2011) opined that in this case, ability differences for examinees taking different forms of tests are confounded by differences in form difficulty. Equating methods can, therefore, be used to adjust for differences in difficulty across forms, resulting in comparable score scales and more accurate estimates of ability. In order to obtain unbiased and reliable scores from a test result, the scores need to be comparable, that is, scores obtained from different scores of a test should indicate the same level of performance regardless the test form that is administered to an examinee.

Equating adjusts for differences in difficulty among forms of test that are built to share some qualities such as test difficulty and similar content (Kolen & Brennan, 2004). In equating, test scores are adjusted based on difficulty of the test forms administered. There could be unfair treatment of examinees who take more difficult test form if equating is not carried out. According to Muraki, Hombo and Lee (2000), equated scores on alternate test forms can be compared, and differences in examinee scores after equating can be attributed to differences in ability instead of differences in difficulty between test forms. Equating, therefore, helps to assure examinees fairness when test difficulty varies. Doran, Moses and Eignor (2010) emphasized that fair and equitable treatment of testees should be commensurate with their actual performance on the test they took. Examinees that are administered more difficult test form are not disadvantaged while those that wrote the easier test will not have undue advantage over those who wrote the more difficult test.

Equating aims at adjusting for differences in difficulty across alternate forms of tests, so that scores obtained from different tests can be used interchangeably since the score scales that will be produced are comparable. As stated by Kolen and Brennan (2004), equating refers to a

statistical procedure that is usually used to adjust scores on different forms so that scaled scores can be used interchangeably. Scaled scores are used to make a given score indicate the same level of knowledge or skill, no matter which form of the test the test-taker took (Livingston, 2004). The scholar further explained that scaled scores are adjusted to compensate for differences in the difficulty of the questions. A scaled score is the total number of raw score obtained by testee that have been converted to a consistent and standardized scale. It is very useful in reporting scores from certificated examination in order to manage possible difficulty across test forms.

Occurrence of different test forms with the aim of measuring the same construct yearly brings about comparison of test scores. This happens in Indonesia, United Kingdom and also here in Nigeria. As stated by Cao (2008), when multiple test forms are used, an equating process should be applied so that examinees' proficiencies obtained across forms and across occasions can be compared on the common scale, which further addresses the fairness concern. Administering of different test forms to different testees, built to the same content and statistical specification but containing different collection of test questions is often used to address test security problems and to compare changes in performance across time. In practice, equating can be used in college admissions using entrance examinations where different forms of the test are administered to groups of examinees, teacher-made tests using different test forms to reduce cheating and in testing companies using standardized tests where different forms are used to provide and report scaled scores. The test forms are usually referred to as alternate forms. A typical example of this in Nigeria is Unified Tertiary Matriculation Examination (UTME). Multiple forms of UTME are administered yearly, such that testees at the same examination centre cannot copy each other due to different set of questions in the alternate forms. This

reduces the chances of examination malpractices to the barest minimum. Alternate forms also help to limit item exposure.

In Indonesia, Keeves and Watanabe (2003) pointed out that three to seven test forms are yearly prepared for the final examinations in Primary, Junior secondary and senior secondary schools. Each of these test forms is constructed to be as similar as possible in content areas and in difficulty levels by using detailed test specifications prepared at the national level. Five test forms out of the seven test forms prepared are used as the main test forms while the remaining two test forms are kept as reserves. Equating is, therefore, necessary to adjust for test difficulty difference so that essential differences in performance are reported. When multiple test forms are administered to different groups of examinees and there is need for comparison, equating is used to adjust test form difficulty (Shin, 2015).

Basically, test equating methods have been classified into traditional (conventional) equating and Item Response Theory (IRT). The earlier mentioned is based on Classical Test Theory (CTT) often called the true score model while IRT is a more modern theory. CTT is concerned with the relationship among observed score, true score and error. Equating under CTT uses methods such as mean, linear and equipercentile equating whereas IRT models that are commonly used are the 1-Parameter Logistic Model (sometimes denoted as “1PL” or “the Rasch Model”), the 2-Parameter Logistic Model (sometimes denoted as “2PL”) and the 3-Parameter Logistic Model (sometimes denoted as “3PL”).

CTT procedures have been greatly practiced over the years. Despite its long usage, yet it is faced with the problems of non-correlation of true and error scores, group dependence item statistics, that is, item difficulty and item discrimination, assumption of equal errors of

measurement among all testees (Enu, 2014). This led to the use of item response theory (IRT), the modern test theory. IRT makes information available on how examinees at different levels of ability on a trait have performed on an item unlike CTT, that uses examinees' raw scores as basis for determining test taker ability. IRT comprises mathematical models that assume the way test takers with different ability levels will perform on test items. IRT is a set of models which, by relating the likelihood of a particular reaction by each person with a given trait level to the characteristics of the item designed to elicit the level to which the individual possesses that trait (Nenty, 2003). In CTT, raw scores are the basis for determining ability of testees while IRT on the other hand, though uses raw scores as well, considers the characteristics of the particular set of individual items on a test form.

CTT and IRT have been compared by many researchers, examples include Ojerinde and Onyeneho (2012) who carried out a research by comparing classical test theory and item response theory using 2011 pre-test in the use of English language of UTME in Nigeria. The study was aimed at evaluating the use of English pre-test data so as to compare the indices obtained using 3-parameter model of IRT with those of the classical test theory (CTT) and hence verify their degree of comparability. One version of the pretest use of English was administered to 1075 testees. The instrument contains 100 item use of English items developed by UTME and the data was analyzed with the use of Microsoft excel programme for the CTT analysis. While XCALIBRE software was used for the IRT analysis. Results from the findings of the study showed that the 3PLM was found to be more suitable in multiple choice ability test. Generally, the indices obtained from both procedures gave valuable information with comparable and almost interchangeable results. It was therefore recommended that both IRT and CTT parameters should be used in empirical

determination of validities of dichotomously scored items to ensure common bases of test analysis, enhance interpretability and objectivity of test agencies in Africa.

Also, Wilberg (2004) examined classical test theory versus the modern test theory (IRT) in an evaluation of the theory test in the Swedish driving – license test which is made up of a theory test and a practical road test. The study examined which among the one (1PL), two (2PL) and three (3PL) parameter logistic IRT models that is the most suitable to use in the Swedish driving – license test. A sample of 5404 test-takers who sat for one of the test versions of the Swedish theory driving license test in January 2004 was used to evaluate the test results. 43.4% of the test takers were women and 56.4% were men with average age of 23.6 years.

The theory test had 65 multiple choice items and it was criterion referenced. The test–taker receives one point for any item answered correctly. Test-takers that had a score higher or equal to the cut-off score 52 (80%) passed the test. The conclusion of the evaluation was that 3PL model is preferable to use when the theory test was evaluated. The study as well compared the IRT model that was selected with the indices in classical test theory (CTT) and concluded that both indices from CTT and IRT gave valuable information and should be included in an analysis of the theory test in the Swedish driving license test. Silvestre -Tipay (2009) investigated the behaviour of item and person statistics derived from two framework of a Biological Science test design for college fresh students. The outcome of the study showed that the degree of difference of item and person statistics across sample appeared to be similar in CTT and IRT.

According to McCallon and Schumacker (2002), IRT measurement models when compared to classical models, offered several distinct benefits. These included the following:

a. Item statistics are independent of the sample from which they were estimated.

b. Examinee scores are independent of test difficulty

c. Item analysis accommodates matching test items to examine knowledge level.

d. Test analysis doesn't require strict parallel test for assessing reliability.

e. Item statistics and examinee ability are both reported on the same scale.

Equating being a process carried out to establish similar scores on different versions of test forms of the same test having the same meaning, allows the scores to be used interchangeably (Ryan, 2011). The goal of equating is to adjust for differences in difficulty that exist across alternate forms of tests so that comparable score scales are produced. Equating is unavoidable when multiple forms of test exist no matter the amount of resources spent in constructing the forms to be parallel in item type format, subject matter and timing, the test difficulty will certainly vary. Equating is, therefore, essential so as to get the best of test construction. Also, where multiple test forms occur there will be multiple score scales that measure the construct of interest at different levels of test difficulty. Albano (2016) stated that equating defines a functional statistical relationship between multiple test score distributions and thereby between multiple score scales. The functional statistical relationship can be described as equating function if the test forms have been built to the same specification and have similar statistical characteristics.

According to Chen et al (2009), equating is the statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences based on the difficulty in test form, not for differences in content. Regardless of the effort made by test developers in constructing similar test forms, multiple test forms differ invariably in difficulty level and in variability. Equating, therefore, is a statistical process which

is used to transform scores from one test form to the scale of another (Asiret & Sunbul, 2016). Several test experts have proposed different properties of equating which are used as the principal basis for developing equating procedures. These properties as suggested by Chen et al (2009) are:

a. **Symmetry:** This property requires that the function used to convert a score on Form X to the scale of Form Y be the inverse of the function used to transform a score on Form Y to the Form X scale. This rules out regression as an equating method.
b. **Same Specifications:** the test forms must have same construct and similar reliability. Imperfect reliability affects the equitability of scores. Adequate reliability is needed to ensure that the results associated with an equating are informative enough to be acceptable for practical use with individuals.
c. **Equity:** It must be a matter of indifference to each examinee whether Form X or Form Y is given to them. ***Lord's equity property (1984)*** implies that examinees with a given true score would have identical observed score means, standard deviations, and distributional shapes of converted scores on Form X and scores on Form Y.
d. **Group/Population Invariance:** The equating relationship is the same regardless of the group of examinees used to conduct the equating. E.g. males, females. However, group invariance does not necessarily hold for these methods when observed scores are substituted for true scores. Dorans and Holland (2000) developed procedures and statistics for investigating group invariance.

Kolen and Brennan (2014) suggested a list of steps that can be used for implementing equating:

a. Decide on the purpose for equating
b. Construct alternate forms. Alternate test forms are constructed in accordance with the same content and statistical specifications
c. Choose a design for data collection. Equating requires that data be collected for providing information on how test forms differ statistically.
d. Implement the data collection design. The test is administered and the data are collected as specified by the design.
e. Choose one or more operational definitions of equating. Equating requires that a choice be made about what types of relationships between forms are to be estimated. For example the choice might involve deciding on whether to implement linear or nonlinear equating methods.
f. Choose one or more statistical estimation method. Various procedures exist for estimating a particular equating relationship.
g. Appraise the results of equating. The results obtained after equating is conducted need to be evaluated. The test development process, test administration, statistical established method and properties of the resulting equating are all components of the evaluation.

When two tests forms are said to be equated, it means they measure the same contents and cognitive processes and support the same inferences on students' knowledge and ability (Ryan, 2011). Test forms are successfully equated if testees' performance does not depend on whether form A or form B is administered to them. Equating is said to be successful to the extent that form taken by each examninee is a matter of indifference (Kolen and Brennan, 2004, p.

430). This is because equated test forms are meant to be interchangeable and examinees can be expected to have equivalent scores no matter what form each person takes.

In educational testing programmes, alternative scoring procedures are becoming more commonly in use. That is, tests are scored number-correct, with scores varying between zero and number of test items. This score can be referred to as raw score which does not need any form of adjustment, it is simply the number of questions answered correctly. Tan and Michel (2011) defined a raw score as the entire score points a testee obtains by answering questions correctly on a test. Raw score expresses the performance of testee on a test.

Test scores can be presented in two ways: raw scores and scaled scores. A raw score as defined by Federal Ministry of Education, Science and Technology (2004), is a direct numerical report of a person's test performance such as number of questions answered, time required and count of right answers. Raw scores can be transformed to scaled scores through statistical procedures. Scaled scores are obtained as a result of some transformation or adjustments that have been carried out on raw scores. Raw scores do not always represent a fair comparison in cases where multiple forms of test exist, for example, if two testees have equivalent raw scores on two different test forms (harder test form and easier test form), the test taker who took the harder test form will show a higher level of performance even with lower score than the one who took the easier test form if they each got the same number of questions correct. The use of raw scores may therefore not be useful. They give very little information about the performance of a student. Scaled score give room for direct and fair comparison of results from one test form to another.

Scaled score serve a purpose of reporting scores for all testees on a consistent scale. Tan and Michel (2011) defined scaled scores as scores that have been mathematically transformed from one set of numbers (that is the raw scores) in order to make them comparable in some way. Percent-correct score is another way of transforming raw scores, it also provides a numerical summary of testees' performance, though it does not present a fair comparison of different test forms and does not give further information to test scores users. It represents the percentage of questions a testee answered correctly on a test. Both raw scores and percent-correct score are not suitable as primary scale for reporting assessment results in most standardized testing programmes because these scores are not comparable across forms. Tan and Michel (2011) pointed out that in order for standardized testing programmes to have consistency in score interpretation when there are different editions of test, programmes often transform test scores (summed raw score points assigned to different questions) in to a set of values different from the raw score points obtained directly from a test.

In order to have meaningful information about test scores, test scores are interpreted either with norm-referenced interpretation or criterion-referenced interpretation, or sometimes with both. A norm referenced interpretation is when test scores of a test taker's performance is being compared to the performance of other people in a specified reference population (Ryan, 2011), while a criterion referenced interpretation is when an individual's test performance can be described without referring to the performance of others (Aleluya, 2012). Norm-referenced interpretation helps test score users to give meaning to an examinee's ability in connection with their standing among other examinees. In other words, it compares individual examinee's score with the performance of others in the class who have taken the same test. For example, an examinee's score of 55 can be compared with a mean score of 49 and a median score of 48 for

the class, then one could say the examinee's score is above average (with the class mean score of 49) and the examinee's performance could be placed in the upper half of the class (with the class median score of 48). Criterion-referenced interpretation on the other hand is when score conveys information about an examinee in connection with a particular subject matter, regardless of other examinees' performance. The different ways of interpreting test scores mentioned above helps in giving test scores the kind of meaning they need in being useful instruments of measurement. This is imperative in standardized testing programmes.

There are various methods available to researchers who want to carry out study on test score equating. Some are Classical Test Theory (CTT) based and others are based on Item Response Theory (IRT). CTT based equating methods include: mean equating, linear equating (Levine linear and Tucker linear equating) and equipercentile equating (Frequency estimation and chained equipercentile equatings). IRT based equating methods include: One parameter logistic (Rasch) model equating (Concurrent calibration, Fixed based procedure, Equating constant procedure), Two parameter logistic model equating (2PL concurrent calibration, 2PL partial credit model) and Three parameter logistic model (separate and concurrent calibration).

In mean equating, differences in difficulty of test forms (X and Y) are estimated by the mean difference, that is, $\mu Y - \mu X$ (Albano, 2015) and the assumption is that the test forms are equally difficult or easy for all level of students (Chen et al, 2009). In linear equating, the mean and standard deviation of the equated scores of Form X are equal to the mean and standard deviation of the target Form Y scores. Scores are equated in equipercentile equating by converting all scores on the new form to scores on the reference form that has the same percentile rank. The procedure in frequency estimation method considers two scores from two different test scores (X and Y) to be comparable if the two scores have the same percentile rank.

Chained equipercentile equating involves equating of test forms using a chain that proceeds from the new form to the common-item scale and then to the reference form (Wolf, 2013).

Classical test theory assumes that, there are no perfect measures of ability, observed score (X) of each testee is comprised of True Score (T) and random error (E) (Wiberg, 2004; Schumacker, 2005). CTT makes use of traditional item and sample dependent statistics which include item difficulty and item discrimination estimates, which are focused on testee assessment at the test score level (Schumacker, 2005). Item Response Theory (IRT) is founded on the assumption that, a mathematical function exists that describes the relationship between an examinee proficiency and probability that an examinee will answer an item correctly (Chong, 2007; Hambleton & Swaminathan, 1995).

Different researches have been carried out on equating of test scores with the use of different designs of equating and comparing different methods of equating. It has been discovered from different research results that many factors such as content difficulty, content format, length of test, population ability contribute to the differences of the methods (von Davier & Chen, 2013). Sinharray and Holland (2009) investigated the missing data assumptions of three methods of equating (poststratification equating method, chain equipercentile equating method, and item response theory observed score equating method under the NEAT design. Different assumptions about the missing data in the NEAT design were made under each of these equating methods. After the study described the missing data assumptions of the three equating methods then a fair comparison of the three methods were carried out using data from three different operational tests. For each of the data set, the researchers examined how the three equating methods will perform when the missing data satisfy the assumptions made by only one of these equating methods. Findings from the study showed that chain equipercentile equating method to

some extent is more satisfactory than the other methods. It was recommended that equating practitioners should seriously consider the chain equating method when using the NEAT design.

Skaggs (2005) investigated how effective it is to equate very small samples using the random group design. Two identical forms and two non identical forms which differs by one-tenth of a standard deviation in overall difficulty were equated using mean equating, identity equating, unsmoothed equipercentile equating, linear equating and equipercentile equating using two through six moments of log-linear presmoothing with samples of 25, 50, 75, 100, 150, and 200. Data for the study was obtained from the Social Studies Test of the Tests for General Educational Development (GED) in the United States. The test consisted of 50 multiple-choice items. Outcome of the study showed that identity equating was preferable to any other equating method especially when samples were as small as 25. As sample size increases standard error decreases. For samples of 50 and above, linear equating is the most accurate when the passing score is near the mean while equipercentile equating with 2 and 3-moment presmoothing were the best equating methods when passing score is greater than the mean. This finding is in agreement with Aiseret and Sunbul (2016) who compared different methods of equating for random group design, factors such as sample size, difference in difficulty between forms, and guessing parameter were considered using small samples.

The equating methods of identity, mean, linear, circle-arc, and 2- and 3-moments pre-smoothed equipercentile for different sample sizes (10, 25, 50, 75, 100, 150, 200) were used to equate two simulated test forms through 100 replications. Findings from the study showed that sample size of 50 or more had difference of 0.4 level of difficulty between the test forms, it was concluded that equating the forms gives better results than not equating. Circle-arc and mean-

equating were the two methods that produced lower equating errors for small samples under most of the conditions considered.

Likewise, Heh (2007) under random groups design examined the accuracy of small sample equating when tests mean difficulties are at variance at several levels. Nine different equating methods (identity, mean, linear, unsmoothed equipercentile, and 2-6 moments pre-smoothing polynomial log-linear equipercentile) were carried out on simulated tests with 6 different levels of mean difficulty differences (0, 0.15, 0.30, 0.45, 0.60 and 0.75) for 6 sample sizes (25, 50, 75, 100, 150 and 200) using Monte Carlo simulations with1,000 replications per cell. Findings from the study showed that for the test lengths and equating designs considered, small sample equating accuracy depended on difficulty differences between the test forms, range of scores over which equating was evaluated, sample size, and equating methods employed. Two- and 3-moments polynomial log-linear presmoothed equipercentile equating methods were therefore found to be most error free for small sample equating under most of the conditions researched.

Wolkowitz and Davis-Becker (2015) investigated the need for a set of common items to have the same content representation as the total test and still produce accurate equating. The study obtained data from four credentialing exams with seven different common item blocks. Tucker Linear as the CTT method and Rasch true score equating as the IRT method of equating were used to perform equating procedure. Findings from the study showed that all the four exams were unidimensional and produced more accurate equating for the Tucker Linear procedure when more items were used in the common item block. This points out that the Tucker Linear method of equating with a common item block of 50% of the total items on the exam that were close to but not proportional to the content representation of the total test, performed better

than the six other equating blocks. For IRT equating, all common item block performed equally well, reason being that the two groups that participated in the study had similar abilities and approximately equal difficulties of the content areas for the exams. It was concluded in the study that common item block that is strictly proportional in content or difficulty to the entire exam may not be needed if the exam is unidimensional.

Baghaei (2010) equated two forms of a reading comprehension test and a pass or fail decision consistency was investigated under two conditions of with or without equating. Equating of the two test forms was carried out using concurrent common item equating with one parameter item response theory model. Results from the study showed that when equating is lacking there will be unfair pass or fail decisions. Wang (2013) investigated how various test characteristics and examinee characteristics influence common item non-equivalent group (CINEG) mixed-format test score equating results. The study made use of simulated data and simulees' item responses were generated using items selected from one multiple choice (MC) item pool and one constructed response (CR) item pool which were constructed based on the College Board Advanced Placement examinations from various subject areas.

Five main factors including item-type dimensionality, group ability difference, within group ability difference, length and composition of the common-item set and format representativeness of the common-item set were investigated in the study. Also, the performance of two equating methods, that is, the presmoothed frequency estimation method and the presmoothed chained equipercentile equating method, were compared under various conditions. The study concluded that the presmoothed frequency estimation method was more sensitive to group ability difference than the presmoothed chained equipercentile equating method. The two methods performed nearly the same in terms of random error.

Sunnassee (2011) carried out a research and examined the factors that affect the accuracy of classical method of equating for small samples, a simulation study was used under the NEAT design. The equating methods that were put into consideration in the study in general, are used under non-equivalent anchor test (NEAT) designs with observed score. The equating methods are: (1) identity method (IDEN); (2) circle-arc method (CARC); (3) chained linear method (CLIN); (4) smoothed chained equipercentile method (SCEE); (5) smoothed frequency estimation method (SFRE); (6) the Tucker method (TLIN); and (7) the Levine-observed score method (LLIN).

Levels of test difficulty and measurement precision, various test lengths and 20 different sampling conditions related to sample size and the magnitude of ability differences between the samples under a non-equivalent anchor test design (NEAT) equating design are part of the 60 test characteristic conditions in the simulation study design. The main purpose of the study was to establish a set of guidelines that help testing practitioners to have a better knowledge of which methods of small sample equating will work better under particular conditions, as well as when small sample equating may not be appropriate.

Suggestion made from findings in the research is that caution is needed when equating small samples under the NEAT design where any of these six conditions occur: (1) small sample size; (2) the magnitude in the differences in group ability; (3) difference in mean item difficulty between alternate forms; (4) lower average item discrimination of any alternate test forms; (5) equated test forms with too few items; and (6) low average item discrimination. Absence of these conditions suggests that small-sample equating is indeed possible.

Agah (2013) carried out a study to determine the relative efficiency of test score equating methods in the comparison of students' continuous assessment measures in Mathematics. Three equating methods, that is, linear equating, separate calibration and concurrent calibration that are based on classical test theory (CTT) and item response theory (IRT) frameworks were investigated. The design used for the study was Non-Equivalent Anchor Test (NEAT) group design. All senior secondary school III students of Crossrivers and Rivers State were the population study with a sample of 2,905 students drawn through multi-stage sampling procedure. Two parallel forms of Mathematics Achievement Test (MAT) that contains 40 items multiple - choice with reliability of 0.83 and 0.89 respectively were used as instruments for data collection in the study. Data collected were analysed using BILOG-MG and SPSS.

Some of the major findings of the study are: (1) the average root mean square error (ARMSE) obtained for the three equating methods (separate calibration, concurrent calibration and linear equating) were 0.09, 0.05 and 0.04 respectively which shows that linear equating yielded the least error and therefore seems to be more efficient in the study (2) there was no significant difference in the ability estimates of students in both states when their scores are scaled using separate calibration method (3) ability estimates of students in state A and B had no significant difference when their scores are scaled through concurrent calibration and (4) there was a significant difference in the ability estimates of students in both states for test scores equated through linear equating. It was therefore recommended by the study that linear equating method should be used to standardize students' continuous assessment (CA) scores, and also that score assigned to students' responses for every cognitive based continuous assessment should be reported in person-by-item response pattern as this will allow better CTT or IRT analysis to be performed.

Powers (2010) investigated the degree to which equating results are population invariant, the effect of group differences on results obtain from equating, and the impact of group differences on the degree to which statistical equating assumptions hold, whether matching techniques provide more accurate equating results, and whether matching techniques reduce the extent to which statistical equating assumptions are violated. Data for the study was obtained from one administration of four mixed-format Advanced Placement (AP) Exams to create pseudo old and new forms sharing common items. Single group (SG) equating design was used to analyse population invariance based on levels of examinee parental education. Frequency estimation, chained equipercentile, IRT true score and observed score were the common item nonequivalent group (CINEG) design equating methods on which equating was conducted. Examinees were sampled based on their level of parental education by creating old and new form groups with common item effect sizes (ESs) that ranges from 0 to 0.75. Groups with ESs greater than zero were matched using matching techniques including exact matching on parental education level and propensity score matching including other background variables.

Results from the study showed that there was little population dependence of equating results, despite large subgroup performance differences. As ES increased, CINEG equating results tended to become less accurate and less consistent. As group differences increased, the degree to which frequency estimation and chained equipercentile statistical assumptions held decreased. This is contrary to the work of Wang, Won-Clan, Brennan & Kolen (2006), as the study showed that with the presence of group differences, frequency estimation method with smaller Standard Error of Equating (SEE), tends to have larger bias than the chained equipercentile method.

Powers (2011) compared four different curvilinear equating methods; chained equipercentile, frequency estimation, IRT true score and observed score equating and investigated the impact of group differences on equating results and assumptions. It was noted that as group differences increased, equating results became more and more biased and dissimilar across equating methods. Research result showed that the cause of equating inaccuracies may be violations of equating assumptions and also found out that IRT and chained equipercentile equating methods seem likely to be less sensitive to group differences when compared to the frequency estimation method.

Liu, Zu, Curley and Carey (2014) investigated the impact of discrete anchor items when compared with passage-based anchor items when test scores were equated using empirical data obtained from an SAT critical reading section. Only observed score equating methods were used in this study. The study made further investigation on Zu and Liu (2010) study whose study was based on simulation. A discrete item stands alone and is unique, while on the other hand a passage-based item is usually administered with other items based on the same stimulus. Two test forms, X and Y, were spiraled and administered to testees in one administration. Also, two anchors that are content representative were constructed. The mean and standard deviation of the item difficulty in both anchors are comparable. Apart from the effect of anchor type, the effect of equating method and ability differences were also investigated. Findings from the study revealed that the anchor with more discrete items almost always leads to more accurate equating functions than does the anchor with more passage-based items. The discrete anchor produced more accurate equating than the passage-based anchor, regardless of the equating method used, meaning anchor type does not interact with equating method. Whereas, anchor type seems to interact with ability difference: For two similar groups, the discrete anchor produced slightly

better results and much better results when the two groups were substantially different. These results confirm the findings of Zu and Liu's (2010) study.

A study which equated two year BECE results in Basic Science and Technology in Oyo State Nigeria was carried out by Adewale (2016). In this study, two sets of scores were equated using linear and equipercentile methods, item by item performance of the students were compared for the two years using the two equating methods. Results from the study revealed that candidates' performance in the Basic Science and Technology for 2013 and 2014 multiple choice items tend to be equivalent. There was no significant difference in students' performance in the two examinations, the two examinations could therefore be used interchangeably. Also, the results revealed that linear equating method has lower coefficient of variation which makes it more robust and preferable to equipercentile equating method. This is in agreement with Olatunji (2015) who equated scores of SSCE Economics multiple choice paper using linear and equipercentile equating methods and found out that results of linear equating method is different from that of equipercentile equating method because it has lower coefficient of variation, thus it is found to be robust than equipercentile method.

There are various established equating methods that are commonly used under different equating designs. Equating methods basically are used to adjust for any test form difficulty across test administration over the years (Kolen & Brennan, 2004). Equating methods can be employed when equating alternate test forms and they can be classified into linear and non-linear methods (Olgar, 2015). Commonly used equating methods include: mean equating, linear equating, Levine linear equating, Tucker linear equating, equipercentile equating, frequency estimation and chained equipercentile equating. These methods are subsumed into traditional equating which is a Classical model theory, an equating approach (Felan, 2002). Another way in

which equating can be done is is Item Response Theory (IRT) which include Rasch model (one-parameter logistic model) based equating, two-parameter logistic model based equating and three-parameter logistic model based equating (Ajah, 2013). CTT and IRT are the two major test theories.

Mean equating is an equating method that needs only the population means to be estimated from the data obtained, its assumption is that the population distributions to be equated differ only in their means (Livingston & Kim, 2010). In mean equating, scores on test form X is considered to be different from scores on test form Y by a constant unit (Kolen and Brennan, (2004). The linear equating method include: Levine linear equating method and Tucker linear equating method. Tucker linear equating is a form of linear equating method in which the relationship between total test scores and common-item scores is defined in terms of regression slopes (Albano, 2012). Tucker method is used for two groups of examinees that do not have significant difference in their ability levels. It can be used for both equally reliable test forms and unequally reliable test forms as it is the only method that does not take reliability into consideration. The first assumption of the method is linear regression – the regression of form X on V is assumed to be same linear function for populations 1 and 2. Conditional variance is the second assumption– the conditional variance of test form X given V is assumed to be the same for both populations 1 and 2 (Kolen & Brennan, 2004).

This method can be carried out by estimating scores from common items on population 1, mean and standard deviation are used to estimate scores on new form which are linearly transformed by the known association of scores on test form Y (old form) and common items from population 2 (Livingston, 2004). Tucker method has been found to yield inaccurate

equating results when group differences exist in means and variances, and its inaccuracy increased with large sample size (Topczewski, Cui, Woodruff, Chen and Fang, 2013).

Topczewski, et al (2013) investigated four linear equating methods - Tucker, Angoff-Levine, Congeneric-Levine and a variant of the Congeneric-Levine method under the common item non-equivalent groups design. The study carried out an investigation on the accuracy of each of the linear equating methods on how they can estimate the indirect means and variances that are needed in computing the equating relationship. When group differences exist in means, variances or both means and variances, the choice of method used has a significant effect. A simulation data that is centered on item parameters of 75 multiple choice items from a nationally published English test and 60 multiple choice items from a nationally published Mathematics test was used in the study.

Item parameters for the two tests were estimated with the three-parameter logistic (3PL) item response theory (IRT) model and a random sample of 10,000 examinees, using the BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003) computer programme. Results from the study showed that with small sample size and little group difference, Tucker method is more accurate than the other methods. With large sample size and moderate group difference, Tucker linear equating is less accurate than the other methods. It is therefore suggested that when group differences exist and the Tucker method is contraindicated, any of the three other methods should be used.

Demir and Guler (2014) tested the statistical equivalence of different forms of a test which are administered at the same time, using an equating design with shared items for non-equivalent groups. The data collected for the study from 761 students who answered third and tenth booklets of the science studies literacy test was analyzed through Tucker Linear equating,

Levine linear equating, frequency prediction and Braun- Holland linear equating methods. Result from the study showed that the Braun-Holland linear equating method was the most appropriate for the equating of booklets 3 and 10 in the PISA 2009 Science Studies literacy test.

Levine linear equating is an equating method under common items non-equivalent groups design, it was originally developed by Levine (1955). This method is in two forms – Levine observed score and Levine true score methods. In this method, test form X is administered to population 1 and test form Y is also administered to population 2, while the two populations take a common set of items V, also known as anchor items which are present in X and Y. The observed scores on X that have been transformed have equivalent mean and standard deviation as the observed scores on Y. The assumptions for this method apply to true scores which are assumed to be related to observed scores according to CTT model, though only observed scores are used (Albano, 2010). True scores for X, Y and V are T_x, T_y and T_v respectively. The assumptions as stated by Kolen and Brennan (2004) are:

i. T_x and T_y correlate perfectly for both populations, and the same condition holds for T_y and T_v
ii. the linear function of T_x on T_y is the same for both populations, and the same condition holds for T_y and T_v; and
iii. measurement error variance for X is the same for both populations, and the same condition holds for Y and V.

In this method, a sample of testees take new test form Y and common items A, another sample of testees take old test form X with common items A (Gao, 2004). Classical test theory model are then used to estimate the means and variances of the test forms. The means of test

forms X and Y on T under Levine estimates are $\mu_{XT(L)}$ and $\mu_{YT(L)}$ while standard deviations are $\sigma_{XT(L)}$ and $\sigma_{YT(L)}$. the two test forms X and Y can then be equated to obtain equivalent scores.

Adokoniyi (2014) equated Kwara state joint senior secondary school mock multiple Economics papers from the year 2010 to 2012 using fifty multiple choice items with ten (20%) common items and forty unique items. Results from the study showed that groups involved performed averagely and differentially on the common items. All the linear equating methods employed in the study produced comparable results that could be used interchangeably. Levine linear equating method generated fair equivalent score compared to other linear equating methods. Equipercentile equating generated the most fairly accurate results of all equating methods, this indicated its robustness.

Livingston and Kim (2009) compared these equating methods: Chained mean, Chained equipercentile, the Identity equating, Levine linear and Chained linear. Data was obtained using a teacher certification examination. After the test forms were equated, the results showed a substantial difference in their difficulties levels. The Chained equipercentile method showed a better performance when compared to other equating methods used in the study.

Equipercentile equating is a non-linear relationship that exist between score scales by setting the percentile ranks equal for each score point (Albano, 2010). Frequency estimation method is an equating method under the non-equivalent groups design. Albano (2016) stated assumptions of frequency estimation method of equating as:

i. the conditional distribution of overall scores on X for each score in V is the same throughout the populations, and
ii. the conditional distribution of overall scores on Y for a given score point in V is the same across populations

Hou (2007), under these assumptions, described the following steps that frequency estimation method involves:

i. Data collection to get the score distributions of total scores from forms X and Y and their common item score;
ii. Evaluation of the marginal score distributions for the synthetic population based on the invariance assumption of conditional distributions;
iii. Using the percentile rank function to continuize the estimated discrete score distributions from step 3;
iv. Performing equipercentile equating using the percentile rank functions distribution from step 4.

von Davier et al., (2006) and Holland et al., (2006) from their studies have shown that when two groups differ substantially, the frequency estimation method may not be the best choice to carry out the equating procedures. Chained equipercentile equating method which is another equipercentile equating method can rather be used.

Chained equipercentile equating is an equating method which can be applied to equipercentile equating method under non-equivalent groups anchor test design. This method involves equating of form X scores to anchor items scores using testees from population 1, scores on the anchor items are then equated to form Y scores using testees from population 2. These conversions are therefore chained together to produce a conversion of form X sores to form Y scores. Chained equipercentile equating method is based on the assumptions that:

1. the equating of X to V is the same for populations 1 and 2, and
2. the equating of V to Y is the same for populations 1 and 2 (Albano, 2016).

Hou (2007) also affirms that this method assumes that the statistical relationship of the scores from the test forms with the common-item scores are population invariant. Kolen and Brennan (2004) mentioned that chained equipercentile method does not require that both populations be very similar such that the method can be found useful when the two groups differ. Difference between Chained equipercentile and Frequency estimation method has been observed because the bias that occur with Frequency estimation method may not exist with the Chained equipercentile method.

Several researches have been carried out on comparing different equating methods and determining their accuracy. Sinharay and Holland (2006) in their study discovered that large group differences yielded a substantial bias for both Frequency estimation and Chained equipercentile methods, but in general, Chained equipercentile method had less bias when compared with the Frequency estimation method. It was also discovered that the statistical characteristics of the common item set have little impact on equating function performance.

Wang, et al (2006) using common-item nonequivalent groups design in a simulation study, compared frequency estimation and chained equipercentile methods in order to determine errors associated with them. Four forms of a 60-item Mathematics test were used to obtain data for the study with randomly equivalent groups of about 3000 examinees per form who took the test. Three-parameter logistic (3PL) IRT model was used to estimate the item parameters. The four test forms used are parallel test forms with different test lengths. Results from the study showed that the frequency estimation method have larger bias than the chained equipercentile method when there are group differences.

Item Response Theory (IRT) is another approach to equating, in which examinee responses are modeled at the item level rather than the test score level (Wolf, 2013). Agah (2013)

mentioned that IRT models are mathematical functions that indicate the probability of discrete outcome, such as a correct answer to an item, in terms of person and item parameters. Item parameters are difficulty level, discrimination index and pseudo guessing. Equating methods under IRT include Rasch model also called one-parameter logistic model based equating, two-parameter logistic model based equating and three-parameter logistic model based equating (Agah, 2013). Like other equating methods, some statistical assumptions must be met in order to have a valid and reliable IRT measurement results. The assumptions are:

i. Monotonicity: this asserts that the likelihood of successful performance is a non-decreasing function of a testee's proficiency
ii. Local independence: performance of an item is provisionally independent given a testee's trait level
iii. Dimensionality: is the quantity of latent aptitudes needed to capture the construct of interest (Embretson & Reise, 2000).

Two equating methods, separate and concurrent methods under IRT, were compared by Osho (2019) who equated 2016 BECE Mathematics multiple choice test. Data collected for the study were analysed using the marginal maximum likelihood estimation (BILOG-Mg), mean/mean statistic and PIE for IRT true score equating. Results from the study showed that concurrent equating method is more efficient than separate equating method in equating of mathematics multiple choice tests. Results also showed that scores from the 2016 BECE Mathematics multiple choice test forms were not comparable.

Wolf (2013) used four equating methods - IRT True Score, IRT Observed Score, Frequency Estimation, and Chained Equipercentile to examine preservation of equity properties. This study was carried out under a common-item nonequivalent groups (CINEG) design using a

mixed-format test. Traditional equating methods (frequency estimation and chained equipercentile) were found to perform similarly when groups were equivalent while IRT equating methods had better performance when compared to conventional method of equating in terms of equity preservation across all conditions.

von Davier and Wilson (2008) investigated population invariance for gender groups for the Advanced Placement Programme Calculus AB exam using an internal anchor test data collection design. A multiple-choice test and a test composed of both multiple-choice and free-response questions were equated. Item response theory, chained linear and Tucker linear equating were the equating procedures that the study used. Outcome showed that the two administration groups did not differ much in ability, unlike the gender groups that had large differences in ability. It was discovered that all equating methods produced satisfactory and comparable results for both tests for equating based on gender or total administration groups.

Levine Equating

Levine linear equating method is an equating method classified within Linear equating method in NEAT design. Linear equating method is the most straightforward of the equating methods (Gao, 2004). Ryan (2011) described linear equating as a tool used primarily under Classical Test Theory for determining equivalent scores between two parallel test forms. It is basically used when two test forms X and Y to be equated are equally reliable and the standard score deviates of both forms X and Y can be considered to be equal. Linear equating as described by Albano (2010) defines a linear relationship that exists between scores obtained from test forms X and Y, based on the mean and standard deviation of each. It is expressed as linear

equating because the relationship between scores obtained from tests X and Y can be shown as a straight line on a graph (Ryan, 2011).

Linear equating method is best implemented when the groups of examinees taking the different test forms are equivalent or have equal ability, but can also be used in a non-equivalent anchor test group design (Kolen & Brennan, 2004; Tanguma, 2000). When the proficiency of the students who are administered the different test forms are not equal, Levine linear equating method has been indicated as a likely method to be more applicable. It is one of the equating methods under NEAT design (Kolen & Brennan, 2004). This procedure is based on the assumption that the test items in the different test forms are randomly parallel to the set of equating items and the abilities of examinees are statistically different. Levine equating method is classified into Levine Observed Score and Levine True Score methods.

Levine Observed score method being an equating method that connects observed scores on form X to the scale of observed scores on test form Y. This method is one of the equating methods under NEAT design (Kolen & Brennan, 2004). Albano (2010) mentioned that though only observed scores are used, assumptions for Levine observed score method are stated in terms of true scores across population of testees. The first assumption is that the correlation between true scores on test form X and anchor, V is 1, as is the correlation between true scores on test form Y and anchor, V. Secondly, the coefficients, which is as a result of a regression of true scores for form X on V are the same, as with true scores for form Y on V. And thirdly, variance of measurement error is the same for test forms X, Y and anchor, V. Levine observed scores equating method is sometimes more accurate than other linear equating methods when computed in practical applications for comparison purposes (Mroch et al., 2009).

According to von Davier & Kong (2003) and Hou (2007), this method is based on three statistical assumptions: correlational, linear regression and error variance assumptions. In correlational assumption, it is assumed that test forms X, Y and anchor test V all measure the same thing, true scores of test forms X and Y (T_x and T_y) as well as true score of the anchor V (T_v) correlate perfectly in both populations. In linear regression assumption for Levine method, the regressions of T_x on T_y and that of T_y on T_v are linear and the same for both populations 1 and 2. Also, Levine method assumes that the measurement error variances for X and Y are the same in the two populations. As observed score method equates observed scores on X to the scale of observed scores on Y, likewise, true score method equates true scores.

In addition to Levine equating method, Tucker equating method are two of the more popular methods of linear equating method. von Davier and Kong (2003) in comparing the two methods concluded that when two populations seem to be dissimilar, Levine method is more preferable to use. von Davier and Han (2004) also examined the relationship between Tucker and Levine equating methods by investigating the population sensitivity of linear equating methods that are mostly used, that is, Tucker, Levine observed-score, and chain linear methods, in NEAT design. It was concluded that Levine method seems to vary less across subpopulations while Tucker method seems to be the most varied. One of the Levine methods is better when it is suspected that the population differs.

The procedure for this method according to Gao (2004) is as follows: sample P takes new test form Y and anchor items A, sample Q are administered old test form X with a anchor items V, Levine linear method then uses a classical test theory model for test form X, test form Y, and set of anchor items A to estimate the means and variances of test forms X and Y on target population T (von Davier & Chen, 2013). The means of test forms X and Y on T under Levine

estimates are $\mu_{XT(L)}$ and $\mu_{YT(L)}$ while standard deviations are $\sigma_{XT(L)}$ and $\sigma_{YT(L)}$. The assumptions of Levine linear methods have made it able to obtain formulas for means and standard deviations of test forms X and Y on T which are then used to define the Levine linear observed score equating function, $Lin_{XY\,T(L)}(x)$.

It was also found by Chen, Livingston and Holland (2011) when three equating methods (Levine linear, Tucker and chained linear) were compared, that Levine observed score method had the highest equated scores for X when the means of two groups of examinees differ significantly, which was determined through their anchor test items score test. Larger difference in the means shows a better performance of the method. Ozdemir (2017) equated Trends in International Mathematics and Science Study (TIMSS) mathematics subtest scores obtained from TIMSS 2011 to scores obtained from TIMSS 2007 test form and also determined the better equating method to use. The two test forms had almost identical reliability coefficients of 0.892 and 0.892 respectively. The study revealed that Levine equating method with bias value of 0.744 outperformed chained equipercentile equating method which has higher bias value of 0.984.

Chained Equipercentile Equating Method

Equipercentile equating method is another equating method that is used in non-equivalent groups anchor test (NEAT) design. It is an equating procedure under classical test theory. In this method, a score on the new form and a score on the reference form are equivalent in a group of examinees if they have the same percentile rank in the group. When equating scores on the new form to scores on the reference form in a group of test-takers, each score on the new form is transformed to the score on the reference form that has the same percentile rank in that group (Livingston, 2004). According to Kolen and Brennan (2014), the equipercentile equating

function is established if the distribution of scores on form X that is transformed to form Y scale is found equivalent to the distribution of scores on form Y in the population. The equipercentile equating function is developed by identifying scores on form X that have the same percentile ranks as scores on form Y. two test scores X and Y can be put on the same scale when they share the same percentile in equivalent groups. In order to have accurate result, it is required that both test X and Y measure the same ability.

Under non-equivalent anchor groups design, there are two equipercentile equating methods- chained equipercentile equating and the frequency estimation methods. Chained equipercentile is described as an equipercentile equating that involves test forms X and Y with anchor A, and populations P and Q taking test forms X and Y, respectively, the chained equipercentile from X to Y is made up of two equipercentile equatings from X to A with population P and from A to Y with population Q (Chen & Holland, 2009). Hou (2013) mentioned that chained equipercentile method has to do with equating a long test (total test) to a short test (common items) that may be quite disimilar to the features of the long test. Shin (2015) stated specific steps in chained equipercentile equating: Scores on Form X are equated to scores on the common items using Population 1; common items scores are then equated to scores on Form Y using Population 2.

The assumptions for chained equipercentile equating method are (von Davier, et al 2004):

1. For a given population, the link from X to V is group invariant.

2. For a given population, the link from V to Y is group invariant.

The assumptions involve the linking relationship between total test scores and common item scores in both groups of test takers (Powers, 2011).

Kolen and Brennan (2004) stated that chained equipercentile equating is less computationally intensive because it does not require the joint distribution of total and common-item scores as needed in the frequency estimation method. Also, chained equipercentile equating method produces less bias when groups of testee taking the different test forms differ substantially (Kolen & Brennan, 2004; Wang, 2008). Many researchers who have carried out studies on chained equipercentile equating often compare this equating method with frequency estimation method, or sometimes with other types of equating methods. Holland, et al (2006) and Wang, et al (2008) compared the two methods and discovered that they produced quite different equating results. The scholars opined that chained equipercentile equating method could work better especially when the two groups differ.

Power (2011) examined the impact that group differences could have on equating accuracy using four equating methods among which was chained equipercentile equating method. The author discovered that chained equipercentile equating methods appear to be less sensitive to group differences when compared to other equating methods. Ricker and von Davier (2007) also found that frequency estimation method may have less bias than chained equipercentile method when the common item set is relatively short compared to the total test length but chained method appears less sensitive to large group differences when given sufficient numbers of representative common items. Also, in a research study carried out by von Davier, Holland, and Thayer (2003), the equating results produced when chained equipercentile method was employed were less population dependent when compared with the results from the frequency estimation method.

Equating Designs

An equating study requires a group of examinees that will be reasonably representative of those who will be tested in real life world administration so that accurate data can be collected for equating. Variety of equating designs can be used in collecting data. Livingston (2004) stated that an equating design is a plan for collecting the data you need for equating. According to Albano (2011), an equating design can be related to the essential component of an equating study, just as a research design is related to the structure of a research study. A properly done equating procedure requires an appropriate equating design to be employed so as to be able to gather a suitable data. Also, test equating practice requires methods to distinguish the effects of examinee abilities from difficulty differences of the tests to be equated. Factors such as examinees population, test security and the degree to which statistical assumptions are expected to hold influence choice of equating design.

There are varieties of equating designs depending on the needs of a testing programme. Majorly, equating designs are categorized as equivalent groups or nonequivalent groups. In equivalent group design it is assumed that the groups are drawn from the same population and so the groups involved have identical statistical properties. It is also referred to as horizontal equating or common subject equating. Forms of equivalent design include single group design, single group design with counterbalancing and random group design. Single group design requires that same test takers take both test forms X and Y to be equated, that is, the reference form and the new form. It is the simplest data collection design. This design has a major advantage which is its effective statistical ability because the same testees take both forms of test. This helps to control for differential examinee proficiency (Brennan, 2006). Order effects – practice effects and fatigue effects are main concern of this equating design. For example, if

reference form was administered first to all examinees followed by new form, fatigue effects could set in because reference form could appear more difficult due to examinees' tiredness when taking new form. Also, familiarity with the test can make performance better and thereby making new form appear easier, this is practice effects.

Table 2: Single Group (SG) Design

Population	Sample	X	Y
P	I	@	@

@ indicates examinees in sample for a given row take tests indicated in a given column.

Single group counterbalanced design reduces order effects. In this equating design, one group of examinees takes reference form first and the new form next while the other group does vice versa. These groups can be obtained by randomly dividing a sample of examinees into half and each group takes the test in different order.

Table 3: Single Group Counterbalanced (CB) Design

Population	Sample	X_1	X_2	Y_1	Y_2
P	1	@			@
P	2		@	@	

From table 3, X_1 means that form X is taken first, X_2, form X is taken second, Y_1, form Y is taken first, Y_2, form Y is taken second. @ indicates that examinees in the sample for that row take test while absence of @ indicates that no data was collected. In order to have accurate equating results, which is the main advantage of this design, the testees should take the two test form close together in time so that there will be no real change in their level of knowledge or

skills that the test measures. Its disadvantage is that conducting two independent tests to the same group of examinees is impractical.

In random group design, two equivalent samples are taken from a common population P and are randomly assigned the form to be administered. The test forms are distributed using a 'spiraling' process, that is, first examinee takes form X, second examinee takes form Y, third examinee form X, and so on.

Table 4: Random Group Design

Population	Sample	X	Y
P	**1**	**@**	
P	**2**		**@**

@ indicates that examinees in the sample take test while absence of @ shows that no data was collected.

In random group design groups of examinees are randomly equivalent (Kolen & Brennan, 2004), and there is random administration of the test forms to equivalent groups. In this design, the variance in ability between the groups of students who have taken different forms reveals the difference in difficulty between the test forms (Kolen & Brennan, 2004). Some of the practical advantage of random group design is that each examinee takes only one form of the test, therefore test time is minimized. This makes this design preferable to single group design in which students have access to take more than one test form, which is unfeasible to create enough testing time for every examinee to attend to more than one test. Also, it is fairly easy to administer, though it requires large numbers of testees in order to produce accurate results unlike single group design that needs relatively small sample sizes.

Non-equivalent group design is described as non-equivalent groups with anchor test (NEAT) design (von Davier, Holland, & Thayer, 2004) or common-items non-equivalent groups (CINEG) design (Kolen & Brennan 2004). It is also known as vertical equating. This design involves two populations P and Q with a sample of examinees from P taking test form X, and a sample from Q taking test form Y. Forms X and Y have items that are common to each of them, this is referred to as anchor items or common items. Different groups of examinees take the test forms, with the assumption that the examinees that the test forms are administered to are not equivalent in proficiency. This design therefore helps in equating examination because the proficiency level of examinees changes from time to time (Beguin 2000).

Table 5: Nonequivalent Groups Anchor Test (NEAT) Design

Population	Sample	X	A	Y
P	**1**	**@**	**@**	
Q	**2**		**@**	**@**

Anchor item or common item is a miniature of the total test form, which should be comparably representative of the total test forms in content and statistical characteristics. Shin (2015) mentioned that Common item sets should adequately reflect test specifications as well as form difficulty, that is, anchor sets should be content and statistically representative. They are items that measure the same skills and knowledge the actual test is measuring, the more similar the anchor is to the test, the better (Livingston 2004). Dorans, Moses and Eignor (2010) mentioned that the major features of the anchor tests are its stability over occasions when being used and it must have high correlation with the scores on the two tests being equated. Research

has verified that when the anchor test meets these properties, it can be efficiently used to equate test forms with minimal bias under the NEAT design (Mbella, 2012).

Anchor tests should behave in similar ways in the test forms by being placed at same position in both forms. The role of anchor items is to determine what is attributable to differences in total scores. When using anchor item design to equate test forms, it helps to distinguish whether any differences between the two groups of population's overall result is due to difference in students' ability, the test items being different or both (Ryan 2011). There are two variations to this design, it is either the set of anchor items are part of each of the test forms or considered as a separate test. A set of anchor items can be said to be internal anchor when the scores of examinees on the tests is being influenced by scores obtained from the set of anchor items, but when the scores obtained from the set of anchor items does not contribute to the examinees' scores on the test, it is known as external anchor. Internal anchor usually have higher correlation with the test being equated because it contributes to the total score. Anchors with longer items are usually more reliable and more highly correlated with the tests (Dorans et al 2010).

There are generally acceptable specifications for number of anchor items that should be present in test to be equated in order to have best results from the equating procedure. Kolen and Brennan (2004) noted that "experience suggests the rule of thumb that a common [anchor] item set should be at least 20 percent of the length of a total test containing 40 or more items, unless the test is very long, in which case 30 common items might suffice". Ryan (2011) also suggested 10 – 15 anchor items for test forms that contains as long as 50 items. Positioning of anchor items in test forms to be equated is another factor to consider in order to ensure proper equating. Ryan (2011) stated that the difference in item position may be great enough to affect student

performance on an item. For example, if an anchor item in test form A is located at the opening of a test and is positioned as the last item in test form B, student's performance on that particular item could be affected. Therefore, anchor items are specified to be placed at fixed positions on the two test forms.

Under each equating design several studies have compared and considered different methods of equating. Some have seen some equating methods better than others in the existence of various conditions. Hou (2007) evaluated how efficient hybrid Levine equipercentile (Hybrid LE) and modified frequency estimation (MFE) equating methods can be when accuracy of equating is to be improved. This is done in comparison with the percentile rank frequency estimation (FE), percentile rank chained equipercentile (CE) and kernel frequency estimation (Kernel FE) equating methods under the NEAT design. Comparison of the equating methods were done under various simulated conditions showing differences in the size of the sample, group proficiency, length of test, ratio of common items and the similarity of form difficulty with log-linear pre-smoothing. Results from the study revealed that the Hybrid LE and MFE methods had best performance under most simulation conditions. Also, Chen, Cui, Zhu and Gao (2010) compared classical equating methods, which involved the Levine observed score and the Tucker methods, varying differences in ability and form difficulty. The scholars concluded that when differences in the two conditions are small, both methods produce similar results.

Von davier and Chen (2013) in their study mentioned three different ways of using the information provided by the anchor scores to equate the scores of a new form to those of an old test form considering observed score equating methods under the NEAT design - one of the methods is when the anchor scores are used as a conditioning variable, such as Tucker method and poststratification equating. The second method is to employ scores from anchor as the

middle link in a chain of linking relationships, such as chain linear equating and chain equating, and the third way is to use the anchor scores in combination with the classical test theory. The study demonstrated that with real data hybrid Levine equipercentile equating and poststratification equating based on true anchor scores outperform both poststratification equating and chain equating.

Wan, Lee, Brennan and Kolen (2008) used simulation to compare two methods of test equating under the NEAT design: the frequency estimation and the chained equipercentile methods. The results from the study showed that when there is substantial group difference, the frequency estimation method has larger bias than the chained equipercentile method since the difference in bias increased as group differences increased. The study therefore concluded that frequency estimation method almost always has a smaller standard error of equating than the chained equipercentile method. The recommendation from the study is that frequency estimation can be used when group differences are small while chained equipercentile method is recommended when group differences are large. Researchers such as Holland, et al (2006) and Wang, et al (2008) suggested that chained equipercentile method could work better especially when the two groups of examinees differ. Also, a study by Holland, et al (2006) compared frequency estimation and chained equipercentile methods of equating under common-item nonequivalent groups design. They concluded from their findings that chained equipercentile method performed slightly better than the frequency estimation method in making accurate predictions based on their assumptions.

Vertical and horizontal equating are other types of equating. Vertical equating is concerned with equating test forms of different grades or levels, it is also referred to as across-grade-scaling. It gives room for comparison to be made between students at different levels and

also comparison of their growth over time. For example, this method places students' scores on two tests of different levels, such as English Language for JSSI and JSSII on the same scale, so that scores obtained from students in both tests can be compared (Lee, 2003; Leugn, 2003; Lissitz & Huynh, 2003). Ryan (2011) conversely defined horizontal equating as equating of test forms within the same grade level. It positions students' scores on two tests at the same level, for the same content area and for the same population so that their scores can be directly compared. It is also referred to as within-grade-scaling. Specifically, horizontal equating is used to compare two or more groups of examinees that are on equal level of ability employing two or more different test forms of the same content area and difficulty (Leung, 2003; Lissitz & Hunyh, 2003). Horizontal equating is very suitable for high stake test as multiple test forms are required to maintain test security. All CTT based equating methods and IRT based equating methods can be employed in both vertical and horizontal equating.

Under classical test theory equating, each testee's observed score (X) comprises True Score (T) and random error component (E) (Wiberg, 2004; Schumacker, 2005). Majorly, linear equating, chain equating and equipercntile equating are the test equating methods used under CTT. While test score equating methods used under IRT are one-Parameter Logistic Model (1 PLM), two-Parameter Logistic Model (2 PLM) and three-Parameter Logistic Model (3 PLM) all based on equating.

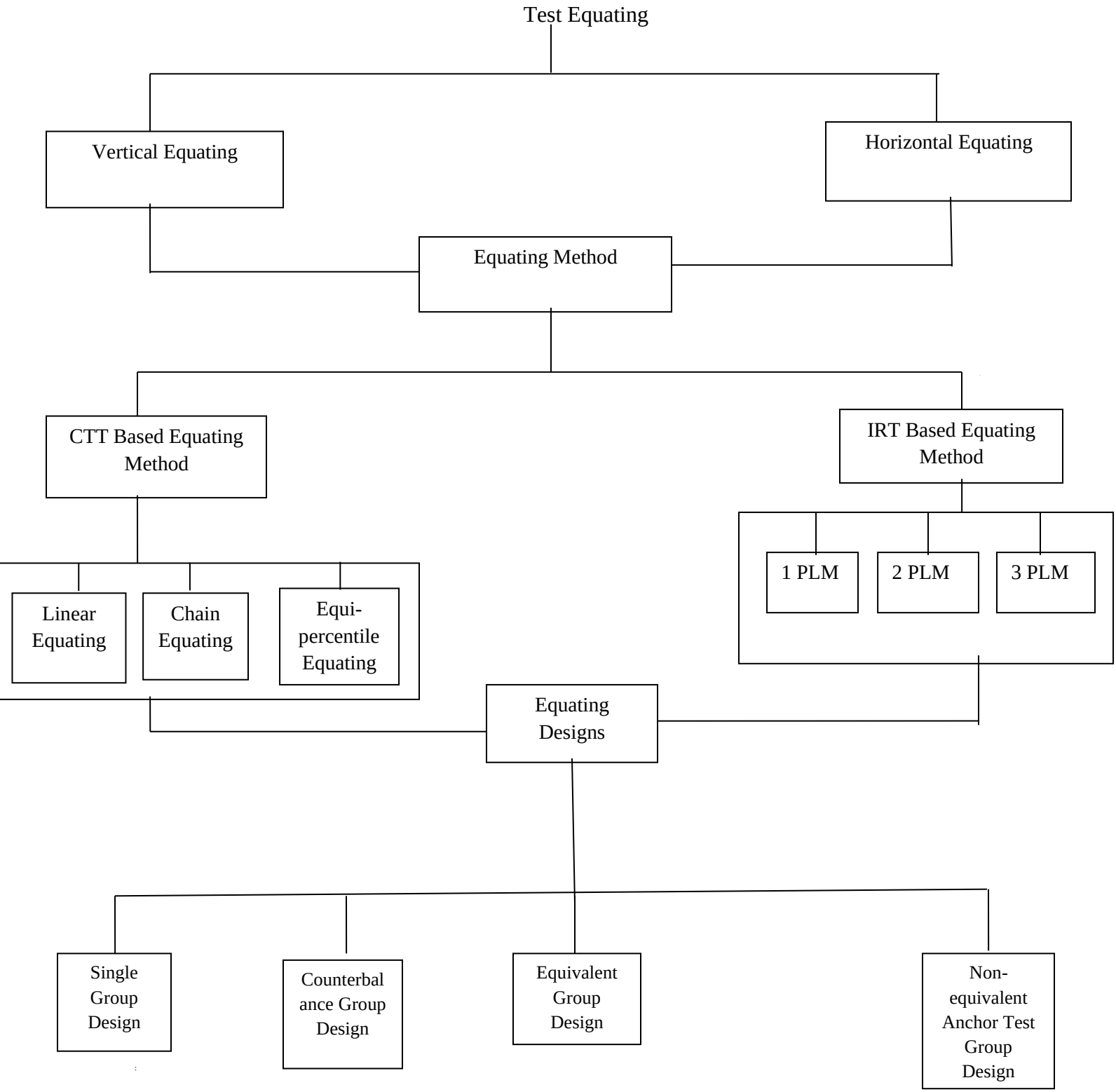

Fig. 1 Diagrammatic Representation of Test Score Equating Methods and Designs

Source: Agah, 2013

Test Equating in Public Examinations

Public examination is necessary and required for ensuring that uniform standard is maintained in the conduct of examination by the examination bodies that certify candidates, that is, West Africa Examination council (WAEC), National Examinations Council of Nigeria (NECO) and the National Business and Technical Examination Board (NABTEB). Comparing these examination bodies in Nigeria has always been a concern. It is required to compare between different examinations in the same subject, set by different boards. Results from these examinations are best compared through equating of test scores. Since the examination bodies in Nigeria vary, there is possibility of claiming that it is not possible to directly compare test scores from these different forms, on this basis, test equating is necessary. When different tests aim to measure the same construct from year to year, the issue of comparability of test scores is raised (Agah, 2013). Test equating research and development has attracted so much attention because of prevalent use of high stakes public examinations across the world, and there is pressure on psychometricians to be able to interpret results from administrations of different tests (Lamprianou, 2007). Test equating makes it possible for test scores from different items to be put on a common scale, this allows interchangeable use of alternate forms of test to be built to the same content, and statistical specifications (Haertel, 2004).

Equating of multiple test forms can be done by linking or connecting all the tests with the aim of ensuring that the different measures each test implies are all shown on a single common scale. A set of common items can be used to link two or more test forms of nearly the same difficulty. When equating test forms, it must be ensured that the tests items are built to the same test specifications and the students who will be the target population, that is examinees that the

tests will be administered to must have been taught based on the same curriculum. Asiret and Sunbul (2016) opined that based on the property of same specifications, test forms to be equated are required to have the same content and statistical characteristics. These criteria can be said to be met by the three public examination bodies that are responsible for candidates' certification at the senior secondary school level in Nigeria. Recent studies have examined and evaluated the equivalence of the alternative forms of SSCE conducted by West African Examination Council (WAEC), National Examination Council (NECO) and National Business and Technical Examination Board (NABTEB).

Bandele and Adewale (2013) examined and compared the item difficulty levels of WAEC, NECO and NABTEB Mathematics Achievement Examinations. Results from their findings showed that the differences in the difficulty levels are not significant and recommended that none of this examination should be seen as inferior, therefore there should be no discrimination among their certificates. Also, Moyinoluwa (2015) analysed the psychometric properties of mathematics in public examinations in Nigeria (2008- 2009) and discovered that basically all the examination bodies have the same structure of syllabus and the objective questions were set in relation to their syllabi. It was also discovered that they all developed tests (between the years under study) covering not less than 80% of their syllabi content. Likewise, Okoye and Nwafor, (2009) compared the content coverage of WAEC to that of NECO. The study showed similarity in the distribution of questions in Biology, Chemistry and Economics tests for both testing agencies.

Salako, Adegoke and Ogundipe (2017) compared the performance of WAEC and NECO SSCE in Mathematics and Physics. A quantitative comparative analysis on the candidates'

performances in the two subjects was carried out. t-distribution and correlation analysis were used to investigate the significance difference of means and correlation coefficients respectively between the successes documented in WAEC and NECO in the two subjects. The study showed that WAEC and NECO successes in Mathematics and Physics are not correlated. Likewise, Udofia and Udoh (2017) carried out a research on comparative analysis of WAEC and NECO SSCE Mathematics from 2008 - 2012. Chi-square and t-test were used to analyse the research questions generated at .05 level of significance. The study showed that WAEC and NECO are similar and comparable.

Other studies that had evaluated the equivalence of the examination bodies include that of Obinne (2011), who compared the Biology examination conducted by WAEC with that conducted by NECO to see if both tests are equally reliable, their standard error of measurements were estimated as a way of measuring their reliabilities. From the study it was found out that both tests were equally reliable. Also, Obinne, Nworgu, and Umobong (2013) analysed the Biology test scores of WAEC and NECO by estimating and comparing the differential item functioning (DIF) of the Biology test across students in urban and rural areas using the IRT methods. The study revealed that there is no significant difference in the DIF of Biology tests that the two examination bodies organized.

Olutola (2011) also analysed the item parameters of 2008 SSCE multiple choice Biology test items conducted by WAEC and NECO. Their difficulty index, discrimination power and power of distracters were determined and reliability tested. Reliability coefficients for WAEC and NECO SSCE Biology multiple choice items were 0.91 and 0.93 respectively. The research result showed a mean difficulty index of 0.42 and discrimination power of 0.43 for WAEC SSCE

Biology multiple choice items has higher values than that of NECO with mean difficulty index and discrimination power of 0.40 and 0.39 respectively. The study also showed that 95% functional options and 5% non-functional options are present in WAEC SSCE Biology multiple-choice items while NECO SSCE Biology multiple-choice items had 93% functional options and 7% non-functional options.

As a result of these findings, public examination bodies in Nigeria (WAEC, NECO and NABTEB) conducting standardized tests and issuing out certificates for final year senior secondary school (internal) and external candidates can be subjected to test equating. Public examination bodies in Nigeria include:

i. West African Examinations Council (WAEC)

ii. National Examinations Council (NECO)

iii. The National Business and Technical Examination Board (NABTEB)

iv. Unified Tertiary Matriculation Examination (UTME)

v. The Interim Joint Matriculation Board (IJMB)

vi. National Teachers Institute (NTI) and others

In this study, the first three examination bodies listed which are of importance to this study are discussed.

West African Examinations Council (WAEC): West African Examinations Council was established in 1952 and has contributed to education in Anglophonic countries of West Africa Ghana, Nigeria, Sierra Leone, the Gambia and Liberia. Prior to the establishment of the council, Dr. G. B. Jeffry who was the director of University of London Institute of Education was invited

by the British Secretary of State for the Colonies in 1949 to visit some West African countries to look into the general education level and requirements in West Africa. He visited Ghana, the Gambia, Sierra Leone, and Nigeria and strongly supported the need for a West African Examination Council, thereafter made detailed recommendations on the composition and responsibilities of the Council. The Legislative Assemblies of the West African countries passed an ordinance authorizing the West African Examination Council and agreed to the organization of exams, and giving out of certificates to students in individual countries by the Council.

WAEC has the sole function of organizing and conducting secondary school and public examinations in West African Countries such as Gambia, Ghana, Liberia, Nigeria and Sierra Leone (Durotolu, 1999). The council conducts four different types of examinations, they are:

a. International examinations which consists of WASSCE (West African Senior School Certificate Examination), SC/GCE O' levels, and HSC/GCE (Higher School Certificate/ General Certificate of Education) A' levels

b. National examinations taken in individual countries, these comprise the Junior Secondary School Certificate for Nigeria and the Gambia, Junior and Senior High School Certificate Examinations for Liberia, National Primary School and Basic Education Certificate Examinations for Sierra Leone, Basic Education Certificate Examinations for Ghana, and Senior School Certificate Examinations for Ghana

c. Examinations conducted in collaboration with other examining bodies, these include City and Guilds of London Institute, Royal Society of Arts

d. Examinations conducted on behalf of other examining bodies which include: University of London GCE, Scholastic Aptitude Test and Graduate Record

Examinations for Educational Testing Service, Princeton, USA, and JAMB (Joint Admissions and Matriculations Board) examination in countries outside Nigeria (WAEC diary, 2004).

Furthermore, Olutola (2011) pointed out the main examinations administered by the Test Administration Division of the Council. These are:

i. General Certificate of Education Examinations for Schools (Senior Secondary) conducted in May/June every year at ordinary and Advanced Levels. These examinations are handled by the school examination section.

ii. General Certificate of Education Examination Ordinary and Advanced Levels for private candidates and continuing education centers taken in November/December.

iii. London University General Certificate of Education level for foreigners only and is conducted in January and June.

In addition to preparation of syllabuses, constructing of questions, administration of examinations and issuing certificates to candidates, WAEC also conducts research and organizes seminars (Abiri, 2007). WAEC has different departments/ units like objective testing unit, Nigeria Aptitude Testing unit, Test Development and Research Division, Research and Aptitude Tests department, school examinations department and vocational examinations department and the like. All these departments/ unit help WAEC to meet up to her responsibilities. One of the council's responsibilities according to Olutola (2011) is conducting of series of research leading to national improvement of testing and examination procedures. In West Africa, WAEC as an international organization has obviously played a unique role in maintaining international academic standard which facilitates common views, interchange of manpower and ideas and

promotes mutual understanding among the people of member countries. WAEC among other things enables candidates to be eligible for admission into tertiary institution.

National Examination Council (NECO): due to some challenges face by the West African Examination Council (WAEC) and also the steady upturn in the number of candidates who register for SSCE year on year in Nigeria, the Federal Government inaugurated the National Examination Council (NECO) in 1999 to also be conducting SSCE in Nigeria. Also, the wide spread leakage of question papers for West African Certificate Examination (WASCE) conducted by WAEC is another reason why NECO came on board. This incidence prompted the Federal Government to set up the Etsu Nupe Panel in 1997 to re-examine the Nigerian education system and the vision 2010 committee to propose a path of development for the country. National Examination Council having the same standard as WAEC was set up because of the coherent report made by the two committees which was published in 1998. The National Council on Education supported this recommendation at its 46th meeting held in Abeokuta, Ogun state in March 1999. Hence, the Federal Government in April, 1999 declared a decree to start up the National Examination Council (NECO). The council has her headquarters in Minna, Niger state.

The functions of NECO amongst others include:

i. revising and considering, annually, in the public interest the examinations to be held for admission into Federal Government colleges and other allied institutions;
ii. the general control and conduct of the National Common Entrance Examinations for admission into Federal Government colleges and other allied institutions;

iii. developing and administering selection examinations into the Suleja Academy in accordance with such guidelines as maybe approved, from time to time, by the Minister;

iv. developing, administering and conducting aptitude tests for all candidates into Federal Government colleges and other allied institutions;

v. the general control of the conduct of the Junior Secondary School Certificate Examinations in all Federal Government colleges, and other allied institutions and in the Suleja Academy;

vi. conducting a Standard National Assessment of Educational Performance at junior and senior secondary school levels;

vii. conducting researches leading to national improvement of testing and examination procedures at junior and senior secondary school levels;

The council after her establishment has been accrediting schools for the SSCE and JSCE and has also been appointed as a consultant to conduct employment examination for organizations. NECO's maiden June/July SSCE was conducted in the year 2000 and had since continue to conduct senior school Certificate Examination (SSCE) twice in a year, June/July for internal candidates and November/December for external candidates (that s students who are not enrolled in the school system) alongside with the West African Examinations Council.

National Business and Technical Examinations Board (NABTEB): The National Business and Technical Examinations Board came into existence in 1992 to domesticate craft subjects examinations which were then conducted by Pittman's and Royal Society of Arts of London and City and Guilds of London Institute in conformity to the provisions of the National Policy on Education. It was established by Act 70 of 23rd August 1993. This was as a product of Osiyale

Committee's Report, who aimed at cutting down the amount of work carried out by WAEC and to give room for high level of productivity in the conduct of public examinations in Nigeria. Establishment of NABTEB was a key moment of an evolutionary process that spanned from 1977- 1992 and it was at different times that four Government panels were set up to review the place and structure of public examination in our educational system.

Each of these Government panels recommended and justified the reason why the number of examination bodies should be increased, and in particular, a separate body to perform the functions which NABTEB now performs. The process started with the findings of Justice Sogbetun Commission of Enquiry (1978), this was set up due to strong reaction of the public on perceived inefficiency and unchecked leakages of public examinations. Next was Angulu commission which was set up as a result of WAEC's presentation to the House of Representative Committee on Education in 1981, WAEC was in support of setting up other examination bodies in Nigeria in order to reduce her burden. Okoro panel was also set up in 1981 to review the Angulu report. Likewise, Professor Akin Osiyale's Task Force was set up in 1991 "to evolve a strategy to reduce the burden of WAEC and bring about greater efficiency in the conduct of public examinations". All these brought about the establishment of NABTEB.

NABTEB since then has been charged with the following mandate:

a. to conduct examination leading to award of:

i. National Technical Certificate (NTC)

ii. Advanced National Technical Certificate (ANTC)

iii. National Business Certificate (NBC)

iv. Advanced National Business Certificate (ANBC)

v. Modular Trade Certificate (MTC)

b. Take over the conduct of technical and business examinations hitherto conducted by the Royal Society of Arts of London, city and Gould's of London and the West African Examinations Council;

c. Issue results, Certificates and make awards in examinations conducted by the Board;

d. Conduct other specified examinations on behalf of or in collaboration with other examination bodies or agencies such as the London Chamber of Commerce or the Institute of Chartered Accountants of Nigeria etc;

e. Conduct common entrance examinations into Technical Colleges and allied institutions;

f. Monitor, collect and keep records of continuous assessment in Technical Colleges and allied institutions towards the award of certificates in National Business and Technical Examinations;

g. Conduct research; publish statistics and other information in order to develop appropriate examinations, tests and syllabi in technical and business studies;

h. Prepare and submit to the

i. secretary an annual report on standards of examinations and other related matter, and

j. Carryout such other activities as are necessary or expedient for the full discharge of all or any of the functions conferred on it under the Decree.

The board also conducts advance level versions of NTC and NBC examinations in these trades/ discipline: General education, Business Trade, Engineering/construction Trades and miscellaneous trades. Enrolment for NABTEB examination has increased greatly over the years when its certificate was listed by JAMB as a prerequisite for admission into higher institution.

The examinations are taken twice a year, in May/June for internal students, that is, School-based candidates, and in November/December for external students (private candidates).

Equating of Senior Secondary School Chemistry Scores

Chemistry is one of the core science subjects offered by science students at the senior secondary class. Knowledge and skills developed in learning of chemistry contribute enormously to the advancement of science and technology in any society. Mulemwa (2002) remarked that when a country does not have a sound science and technology base risks being alienated from the global village. Chemistry is a very important field for development of science and technology, this makes the nation to be directly involved in its teaching and learning processes. The use of Chemistry as a requirement for technological achievement cannot be over emphasized. The pivot of modern technology in Chemistry, its roles and the uses has expanded greatly in diverse occupation such as in the field of technology, industry, teaching service, health service, food processing, petrochemical industries, forestry and others (Ababio, 2000). Okeke (2005) also described chemistry as one of the pivot subjects for technological development. Chemistry has provided solutions to certain problem; it has also improved the world's economic status. Chemistry is one of the pure sciences that deal with every single material thing in the universe, the ability to understand and skillfully control these materials.

Chemistry can therefore be defined as a branch of science concerned with the nature of substances and how they can react with each other. The definition of chemistry keeps changing as new discoveries are made. Ojokuku (2012) also define chemistry as a branch of science that takes care of the investigation of matter: its structure, composition, properties and the changes it undergoes. Ojokuku (2012) further said that it involves the study of material substances that

occur on earth and in the universe. Among other science subjects chemistry has been identified as a very important subject. It is highly relevant and imperative to the improvement of science and technology of any nation. Due to its high relevance it is made a core science subject among other science related subjects in the Nigeria secondary school educational system. Chemistry is a science discipline whose primary objective is on the nature and properties of the non-living matter which surround us and preparation of new substances from the materials which nature has provided.

Chemistry is a big part of our everyday life, it is linked to virtually everything on earth. It features in almost every area of human endeavors. Chemistry among other science subjects features eminently in the areas of agriculture, health, oil and gas, environment, solid minerals, textile, cosmetics, water supply and sanitation, crime detection, pulp and paper, waste management just name it (Zuru, 2009). It is therefore important for any Chemistry student who wants to study any of the science related courses to understand Chemistry because all of the sciences involve matter. Students who have the ambition to become doctors, pharmacists, nurses, engineers, geologists and other science related careers all study Chemistry. Chemistry is offered at the senior secondary classes in order to help students learn important parts of scientific concepts that would enable them live well in their immediate environment (Jimoh, 2001). The relevance of chemistry and its education value which is pertinent to the development of technology and economics of a nation makes it an acceptable subject in the school curriculum.

Olorundare (1998), defined curriculum as a planned structure and sequential set of learning outcomes organized and carried out under the auspices of the school. Likewise, Ugwu (2008) defined curriculum as the experience a school system provides for its students. Also, curriculum is defined as the planned experiences which are offered to the learners within the

formal educational institutions individually or collectively under the institution's control for the inculcation of worthwhile knowledge, character development and skill acquisition (Okunloye, 2004). The curriculum is planned in such a way so as to produce and equip graduates for higher education relevant operational trade and entrepreneurial skill necessary for poverty eradication, job creation and wealth generation.

The study of chemistry as a science subject has great relevance to man as the application of its principles has helped in modern inventions (Giginna and Nweze, 2014), this makes the nation to be directly involved in its teaching and learning processes. The Federal Ministry of Education (2007) therefore has revised chemistry curriculum for Senior Secondary education. The curriculum of Senior Secondary School Chemistry is expected among other things to enable students

i. Develop interest in the subject of Chemistry
ii. Acquire basic theoretical and practical knowledge and skills
iii. Develop interest in science, technology and mathematics
iv. Develop reasonable level of competence in ICT applications that will engender entrepreneurial skills
v. Apply skills to meet societal needs of creating employment and wealth
vi. Be positioned to take advantage of the numerous career opportunities offered by Chemistry
vii. Be adequately prepared for further studies in chemistry

The senior secondary school chemistry curriculum was revised because it became imperative to update existing chemistry curriculum to cater for contemporary needs of the nation as a country aspiring to be amongst the first twenty economics in the world by the year 2020

(Fahmy, 2000). It is therefore very important to aim at ensuring that programmes in the revised curriculum are always relevant to the technological development of a nation and the universe at large. Eya (2015) pointed out some objectives of chemistry curriculum: to show chemistry and its link with industry, everyday life benefits and hazards; to provide a course which is complete for pupils not proceeding to higher education while it's at the same time a reasonably adequate foundation for a post secondary chemistry course. Agusiobo (2003) referred to curriculum as an organised framework that sets out the content that children are to learn and the process through which children achieve goal which the curriculum sets for them. This is applicable in the four walls of a classroom under the guidance of a teacher in a school system. Teachers therefore play important role in achieving objectives of chemistry curriculum for high academic performance.

One out of the roles of teachers is to effectively teach learners to strife to attain high academic performance. Students' academic performance is of necessity in order to produce quality graduates who will contribute meaningfully to the nation's economic growth. Ehegbulem (1992) referred to academic performance as the level of individual's attainment on learning tasks. It measures the extent to which learners have accomplished after a period of instruction. Students' attitude is a major factor that can affect academic performance. Adewumi (1998) opined that students' academic performance can be considered in relation to their attitude, aptitude, ability or mental capability with their colleagues. Also Yusuf (2004) defined student's academic performance as observable and measureable behaviour of a student in particular situation.

As important as chemistry is to a nation and despite its relevance to needs of learners and technological development of the nation, students' academic performance in the subject have not been encouraging. Adesokan (2002) stated that inspite of realization of the recognition given to

chemistry among the science subjects, students' negative attitude towards the subject is obviously seen, thereby leading to poor performance. Some researchers have investigated into reasons why students perform poorly in chemistry. Ojukwu (2016) attributed teachers' poor qualification, poor method of teaching, lack of teaching experience, and failing to use the instructional materials as perceived reasons why students performance poorly in chemistry. Korau (2006) reported factors such as student factor, teacher factor, societal factor, the governmental infrastructural problem, curriculum related variables, test related variables, textbook related variables and home related variables as factors affecting students' poor performance in chemistry. Saage (2009) identified specific variables such as poor primary school background in science, lack of incentives for test, little or no concentration on the part of students, students not interested in hard work, incompetent teachers in the primary school, large classes and fear of the subject psychologically.

Chemistry as a subject has two components, the theory and the practical aspects which make the teaching and learning of science real (Achor, Agogo & Orokpo, 2011). WAEC Chief Examiner's Report (2002) attributed the poor performance especially in practical aspect of Chemistry to students' non-familiarity with the use of simple laboratory equipment, spelling errors, inadequate exposure to laboratory techniques, lack of observational skills, omission of units when calculating values and inability to write symbols properly among others. While in the theory aspect of the exam inability to represent simple reaction by balanced equations, going against the rules of IUPAC nomenclature, poor spellings, definitions and diagram, non-familiarity with some contents of the syllabus, inadequate understanding of the fundamental principles in Chemistry, inability to distinguish between physical and chemical properties and incompetence in basic Mathematics and other factors have been attributed to poor students'

performance in Chemistry. Igwe (2015) in his study concluded that if government and other stakeholders in education (school administrators, chemistry teachers, students, parents and philanthropists) should come together towards addressing all the contending issues concerning secondary education chemistry curriculum, students academic performance will certainly improve. The relevance of chemistry which makes it very relevant to the progress of science and technology of a nation has led few researchers to study and estimate chemistry students' performance through different equating methods.

Casselman, Ohisen and Atwood (2016) used IRT equating method to identify topics that each student struggles with on practice tests in general chemistry in order to improve success rates of students. Students' test scores for the years 2013 to 2015 were used. Each exam included 20 or 25 items, five or ten items were reused from form year to the next while the remaining questions differed. IRT equating was used to compare the exams and to place students' scores on the same scale. This process helped in predicting how students would have performed if they had been given the previous year's exam. Results from the IRT equate showed that the implementation of practice tests with IRT feedback significantly improved students' test scores when compared to the previous year.

Hagge (2010) examined the effect of equating method and format representation of common items on the adequacy of mixed – format test equating using nonequivalent groups by carrying out analyses on three mixed format tests from the advanced placement Examination programme on three subjects – Chemistry, English language and Spanish language. The scholar considered operational examinee item responses for two classes of data, that is, operational test forms and pseudo-test forms. Factors of investigation that were considered for the operational test form analyses were difference in proficiency between old and new form groups of examinees

and relative difficulty of multiple – choice and constructed response items. Similarly, for the pseudo-test form analyses two additional factors of investigation were considered, these are format representativeness of the common item set and statistical representativeness of the common - item set. For each of the study condition, two traditional equating methods, that is, chained equipercentile and frequency estimation, and two IRT equating methods – IRT true score and observed score methods were used.

Five main findings were obtained from the operational and pseudo-test form analyses. As the difference in proficiency between old and new form groups of examinees increased, bias likewise tended to increase. Secondly, increases in bias were typically large for frequency estimation and small for IRT equating methods when compared to the criterion equating relationship for a given equating method. Another finding from the operational and pseudo-test form analyses is that standard errors of equating tended to be small for IRT observed score equating and large for chain equipercentile equating. Fourthly, results for the analyses were similar when the pseudo-tests were constructed to be similar to the operational test forms. Lastly, results were mixed relating to which common – item set composition resulted in the least bias. And finally, the outcome of the research suggested that the test (from the Advanced placement Examination programme; Chemistry, English language and Spanish language), examinee and common – item characteristics investigated do impact equating results.

Pido (2012) also compared item analysis results that were gotten when IRT and CTT approaches were used. The study aimed at analyzing, determining and comparing the item parameters of multiple choice questions of Uganda certificate of education (UCE). The sample population for the study was selected through multistage sampling procedure. 480 students' scripts in dichotomously scored Physics, Chemistry, Biology and Geography part of the UCE

were used for the study. XCALIBRE 4.1.7.1 software was used to carry out the data analysis and items parameters based on the CTT and IRT approaches were determined. The output of the data analysis contained the item characteristics curve (ICC), item difficulty indices (b), item discrimination indices (a) and differential item functioning with respect to gender. The b and a indices based on CTT and IRT approaches were compared using two methods of correlation coefficients. Findings from the study showed that there is a high correlation between b and a in IRT and CTT approaches. It was therefore recommended that both CTT and IRT should be used for item analysis since they produce similar results.

CTT and IRT were also compared by Magno (2009) who used the chemistry test data of junior secondary school students in Philipines to establish that there is difference between the two theories and employed Rasch model and Cronbach's alpha to analyse data. Results from the findings showed that IRT estimates of item difficulty did not change across samples as compared with CTT which was inconsistent and the difficulty indices were more stable across forms of test in IRT than CTT approach.

Theoretical Framework

The theories that support the basis for equating in this study are reviewed here. Equating is one of the basic applications of measurement theory which is critical to any testing programme that involves the use of multiple test forms/ administration. It is approached in two different measurement frameworks. One of the frameworks that is relevant to this research work is the classical framework which is based on number-correct (raw) scores. Equating under this framework requires converting of raw scores on one test form to the scale of raw scores on the

other test form. The following are the conditions that must be met before equating can be carried out under classical test theory framework

i. the test forms must measure the same construct
ii. the test forms have to be equally reliable
iii. conversion of test scores should be symmetric
iv. equating function should be invariant across subpopulations of testees at different distribution of performance

These requirements may not be possible to be entirely met in practice. Therefore, scores from different test forms will be adjusted to make up for lack of equivalence of test forms.

When equating under classical framework, non equivalent anchor test (NEAT) design is a possible data collection design which is applicable to this study. In this design, two different groups of testees are administered a test form each, and both groups write a common set of items along side, this is also called anchor items. The anchor items help to determine if the score differences can be separated from test difficulty. For classical framework, anchor items should be a mini version of the main test and longer anchor items will yield better results. One of the limitations of this framework is that it is sample dependent and this reduces its utility (Schumacker, 2010), that is, the testee sample should be similar to the testee population for whom the test is being administered. If this happens otherwise, sampling problem can be sorted out through the use of anchor items in the test forms. NEAT design involved the use of anchor items in this study.

The test score equating methods that are employed under classical test theory include mean, linear, Levine linear, Tucker, equipercentile, frequency estimation, chained equipercentile

and chained linear equating (Von Davier, 2008; Von Davier & Kong, 2005; Chong & Sharon, 2005; Skaggs, 2005; Felan, 2002; Tanguma, 2000). This study was limited to the use of two equating methods under NEAT design, they are Levine linear and Chained equipercentile methods. Hou (2007) described Chained equipecentile equating as a method that equates test form X to V (anchor item set) in population 1 and V (anchor item set) to test form Y in population 2 through a chain of two equipercentile equating using percentile rank function (Kolen & Brennan, 2004). Equipercentile functions $e_{v1}(x)$ and $e_{y1}(v)$ are used for populations 1 and 2. The equipercentile relationship that exist between test form X and anchor items V is calculated using data from the new form group ($e_{v1}(x)$) and the equipercentile relationship that exist between anchor items V and test form Y is calculated using data from the old form group ($e_{Y2(v)}$) Lastly, the equipercentile relationship between test forms X and Y can be calculated by chaining the two preceding results ($e_{Y(chain)} = e_{Y2[eV1(x)]}$) (Powers, 2010). In summary,

i. data from the new test form are used to calculate the equating relationship between test form X and anchor test V
ii. equating relationship between anchor test V and test form Y data from the old test form are used to calculate equating relationship between anchor test V and test form X
iii. the two equipercentile transformations are chained to equate the scores on the different test forms (Kolen & Brennan, 2004)

Hou (2007) stated that the main assumption for Chained equipercentile method is that the statistical relationships that exist between the two test form scores and the common-item scores are population invariant. von Davier, et al (2004) also mentioned these assumptions on chained equipercentile method

i. for a specified population, the link from test form X to anchor test V is group invariant

ii. for a specified population, the link from anchor test V to test form Y is group invariant

Theoretically, Chained equipercentile method has been found to produce less bias and seen as a better choice among other equipercentile methods, when there is substantial group difference (Kolen & Brennan, 2004 and Wang, et al, 2006). Its theoretical shortcoming is that it requires equating a long test (total test) to a short test (anchor items) that may not reflect the characteristics of the long test.

Levine linear method was proposed by Levine (1955), it is an equating method under NEAT design that is based on the CTT model of the true scores on the different test forms to be equated and the common item/ anchor test (von Davier & Chen, 2013). A classical test theory model for test forms X and Y and anchor test V was assumed as shown in (1)

$$X = t_x + E_x,\ Y = t_y + E_y \text{ and } A = t_A + E_A \quad (1)$$

Where; X, Y and A are test forms X, Y and anchor test A respectively

E_x, E_y and E_A are error terms of test forms X, Y and anchor test A respectively (they have zero expected values)

t_x, t_y and t_A are true scores of test forms X, Y and anchor test A

A significant assumption of Levine's method is congenericity (Topczeniski, et al, 2013) which can be formulated as the two population invariance assumptions, that for any target population, the true scores of the three tests (X,Y and A) perfectly correlate (von Davier & Chen,

2013). This is the way classical test theory asserts that scores of the three tests measure the same thing but may not be in the same scale or with the same reliability. Levine linear method is governed by three assumptions according to von Davier and Kong (2003) and Hou (2007), they are:

i. correlational assumption: test forms X, Y and anchor item set V all measure the same thing meaning that T_x and T_v, T_y and T_v correlates perfectly in both populations 1 and 2

ii. linear regression assumption: regression of T_x on T_v is assumed to be the same linear function for both Populations 1 and 2, and a similar assumption is made for the regression of on T_y on T_v.

iii. error variance assumption: measurement error variance for test form X is the same for Populations 1 and 2, same assumption applies to both test forms Y and anchor item set V.

Gao (2004) described procedure for Levine linear method as follows: testees from population 1 takes new test form Y and anchor items A, old test form X with a set of anchor items V will be taken by testees from population 2. This method uses a classical test theory model for test form X, test form Y, and set of anchor items A to estimate the means and variances of test forms X and Y on target population T (von Davier & Chen, 2013). The means of test forms X and Y on T under Levine estimates are $\mu_{XT(L)}$ and $\mu_{YT(L)}$ while standard deviations are $\sigma_{XT(L)}$ and $\sigma_{YT(L)}$. The assumptions of Levine linear methods are used to obtain formulas for the means and standard deviations of test forms X and Y on T which are then used to define the Levine linear observed-score equating function, $Lin_{XY\ T(L)}(x)$. This method has been found

suitable to use when the group of testees that are administered different test forms to have varying abilities (Holland, et al, 2006).

Appraisal of the Reviewed Literature

The relevant literature reviewed in this study is focused on test scores equating, Levine linear and Chained Equipercentile equating. Demir and Guler (2014) tested the statistical equivalence of different forms of a test using non-equivalent anchor test design data collected for the study from 761 students who answered third and tenth booklets of the science studies literacy test was analyzed through Tucker Linear equating, Levine linear equating, frequency prediction and Braun- Holland linear equating methods. Braun- Holland linear equating method was found to be the most appropriate equating method. The researcher did not use chemistry as a school subject and the exam was not a standardized exam.

Skaggs (2005) investigated the effectiveness of equating using very small samples under the random group design. Data for the study was obtained from the Social Studies Test of the Tests for General Educational Development (GED) in the United States. Results from study showed that as sample size increases standard error decreases and that linear equating is the most accurate when the passing score is near the mean while equipercentile equating with 2 and 3-moment presmoothing were the best equating methods when passing score is above the mean. Though test used in the study consisted of 50 multiple choice items but the study did not use chemistry as a school subject and did not use non equivalent data collection design. Wang (2013) investigated how various test characteristics and examinee characteristics influence common item non-equivalent group (CINEG) mixed-format test score equating results and found out that the two methods used, that is, presmoothed frequency estimation and presmoothed chained

equipercentile equating methods performed nearly the same in terms of random error. The researcher used mixed format test and not only multiple choice test which has higher reliability as was used in present study.

Agah (2013) carried out a study to determine the relative efficiency of test score equating methods in the comparison of students' continuous assessment measures in Mathematics, using Non-Equivalent Anchor Test (NEAT) group design. Linear equating, separate calibration and concurrent calibration based on CTT and IRT frameworks were the equating methods under investigation. Two parallel forms of Mathematics Achievement Test (MAT) that contains 40 items multiple-choice were the instrument used to collect data. The researcher did not use Chemistry as a school subject and large number of multiple choice items and failed to use standardized examinations like WAEC, NECO and NABTEB.

More recent studies by Adokoniyi (2014), Adewale (2015) and Olatunji (2015) were reviewed in this research study. Adokoniyi (2014) equated Kwara state joint senior secondary school mock multiple Economics papers using mean, linear and equipercentile equating methods but did not equate WAEC, NECO or NABTEB and also did not use chemistry as a school subject. Adewale (2015) and Olatunji (2015) equated two year BECE results in Basic Science and Technology and scores of SSCE Economics multiple-choice paper respectively using linear and equipercentile equating methods. Results from their studies both showed that Linear equating method has lower coefficient of variation which makes it more robust than equipercentile equating method. Both researcher did not consider other equating methods under either linear or equipercentile equating like Levine equating or chained equipercentile equating and neither of the research studies used chemistry as a school subject.

None out of all the literatures reviewed analysed Levine linear equating and chained equipercentile equating of SSCE Chemistry multiple-choice papers using non-equivalent anchor test design, equating methods which are often used when group of testees are dissimilar. Thus, the gaps that are left by previous researcher were filled by this study. These public examination bodies have been set up to conduct senior school certificate examination for candidates which can be used for admission purpose into different higher institutions. It is therefore imperative to investigate whether their scores can be equated and used interchangeably. Also, the study investigated the invariance of equated scores across equating methods.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter explains the methodology that was used in carrying out this study under the following sub-headings:

a. Research Design

b. Population, Sample and Sampling Techniques

c. Instrumentation

d. Procedure for Data Collection

e. Method of Data Analysis

Research Design

The Non-Equivalent Groups Anchor Test Design (NEAT) also known as Common-items Non-Equivalent Group (CINEG) design was used in this study. According to Sinharay and Holland (2006), the NEAT design deals with two non-equivalent groups of examinees and an anchor test. In a Non-Equivalent Group Anchor Test (NEAT) design, samples from two different populations take two test forms X and Y on two different occasions, the two populations are not

compulsory to be equivalent (Hou, 2007). The design is considered suitable for this study because there is variation in the ability level of examinees who participated in the study. NEAT design is said to be the most flexible tool available for equating tests (Kolen & Brennan, 2004; Livingston, 2004).

Anchor items which are also called common items were administered alongside with the test forms. The common items non-equivalent groups (CINEG) design according to Powers (2011) provides a way to adjust for differences in form difficulty by imbedding a subset of items from a previous form into a new form. Scores obtained from the subset of items, that is, the common items are used to make adjustments for differences in form difficulty, taking into account differences in group performance.

When using NEAT design, it is assumed that the groups of examinees are not equivalent and are not taking the same test forms, they must, therefore, be connected through anchor test items, and these are used for equating the test forms which also account for group differences in ability (Agah, 2013). Common items should be representative of the total test in content and statistical characteristics when using NEAT design. The proportional content representation of anchor items should almost be the same as the proportional content representation of the entire test form, even to the point of considering the set of anchor items to be a "mini-version" of the full test form (Kolen & Brennan, 2004). Each common (anchor) item should be of sufficient length. It is expected by psychometricians to see at least 15 to 20 anchor items for longer test forms (Ryan, 2011). Each common item should occupy the same position or location in the test forms.

Non-equivalent anchor group (NEAT) design is very useful for equating test scores from non equivalent groups of examinees, and the procedure of data gathering in non equivalent group

design is achieved through achievement test (Kolen & Brennan 2004). This study therefore, adapted the 2017 WAEC, NECO, and NABTEB Chemistry Multiple Choice Question papers in order to collect data. This design helped the researcher to equate the Senior School Certificate Examination (SSCE) Chemistry multiple-choice papers of the three different examination bodies, and also permitted the researcher to investigate the invariance of equated scores of Senior Secondary Certificate Chemistry multiple-choice papers across equating methods. This research helped to identify which of the equating methods, Levine linear or chained equipercentile method is better. Table 6 shows the plan of the study

Table 6: Nonequivalent Groups with Anchor Test (NEAT) Design

Population	**Sample**	**X**	**V**	**Y**	**Z**
A	1	@	@		
B	2		@	@	
C	3		@		@

X - WAEC Multiple Choice Items

V - Anchor Test Items

Y - NECO Multiple Choice Items

Z - NABTEB Multiple Choice Items

@ - denotes examinees in sample for a given row take test indicated in a given column

Population, Sample and Sampling Techniques

All public senior secondary schools in South-west, Nigeria constituted the study population. South-west Nigeria consists of six states: Oyo, Ondo, Ogun, Osun, Lagos and Ekiti States. Simple random sampling technique was used to pick three states (Ogun, Ondo and Ekiti) out of the six states, each state was selected independently of the other states. This was done so

as to afford each state equal chance of being selected for the study. Each of the state has three senatorial districts. Records from the Ministries of Education revealed a total of 6,509 senior secondary schools in all the states in South-West Nigeria with a total enrolment figure of 1,961,505 students and a total of 322,484 Senior Secondary III (SS III) students. According to Yamane formula in Israel (2003), sample size can be obtained by:

$$n = \frac{N}{1 + N(e^2)}$$

Where n is the required sample size, N is the total population and e is the margin of error (MoE), e = 0.05 is usually used based on research condition. In this study, e = 0.03 was used because of the large population involved.

$$N = \frac{322,484}{1 + 322,484\ (0.03)^2} = \frac{322,484}{1 + 290.2356}$$

$$= 1107$$

1,461 students were eventually selected as samples for the study because intact classes were used. This was achieved by using simple random sampling technique to select five schools in each senatorial district in each state. In total, 45 public senior secondary schools were selected. Purposive sampling technique was used in this study to select senior secondary three (SS III) chemistry students from all the 45 public senior secondary schools that were selected. These students are in the best position to respond to instrument that was used in this research because they were expected to have completed a significant part of senior school certificate chemistry syllabus and they were preparing for WAEC, NECO and NABTEB. A total number of 1,461

Chemistry students participated in the study. Table 7 showed the procedure for selection of sample in this study.

Table 7: Sample Size Table

South West States	Selected States	Senatorial districts	Number of schools	Sampled school	Sampled students (intact classes)
Lagos	Ondo	North	89	5	166
Ondo		Central	96	5	172
Oyo		South	99	5	187
Ekiti					
Osun	Ekiti	North	63	5	121
Ogun		Central	71	5	133
		South	68	5	127
	Ogun	East	97	5	176
		South	94	5	183
		West	103	5	196
Students' Population: 322,484					**Sample: 1,461**

Instrumentation

The 2017 WAEC, NECO and NABTEB Chemistry multiple-choice papers were adapted and used as instrument for data collection in this study. The papers comprised unique and anchor items for this study. Each test form had unique items and a set of anchor (common) items that

were positioned as a block at numbers 11 - 30 in each test form. This is referred to as appended internal anchor (Ryan, 2011). Test form A (WAEC), test form B (NECO) and test form C (NABTEB) contained 30, 40 and 30 unique multiple choice items respectively, and each test form also contained 20 multiple-choice common/ anchor items. The 20 multiple-choice common items were selected from the three test forms as there are similar items among the test forms.

The researcher determined the content validity of the instrument by calculating the percentage-difference coefficient of correlation between the test forms and syllabi contents. Coefficients of content validity obtained for WAEC, NECO and NABTEB were 0.78, 0.75 and 0.76 respectively. Also, the instrument was given to chemistry experts (experienced chemistry teachers) to give keys to the items.

To establish the reliability of the instruments (Test forms A, B and C), the researcher used measures of internal consistency. Measures of internal consistency was used because it is appropriate for a test containing only multiple-choice items (Gao, 2004). Internal consistency of a test shows whether test items that are supposed to measure the same construct produce consistent results (Tang, Cui & Babenko, 2014). If items of a test consistently measure the same construct, then the test can be said to be internally consistent. This was done by using the split-half method. The test items were divided into two halves by putting odd-numbered items in one group and even-numbered items in the second group for each testee. Scores that were obtained from the two groups were correlated using Pearson's Product Moment Correlation Co-efficient (PPMC). Spearman Brown correction formula was applied and coefficients of reliability for form A was 0.81, that of form B was 0.77 and form C had a coefficient of reliability of 0.78. This indicates reliability of the test forms. Test forms A, B and C are in appendices A, B and C.

Procedure for Data Collection

The instruments that were used to collect data for this study were the adapted 2017 WAEC, NECO and NABTEB Chemistry multiple-choice papers. Before administering the instruments, the researcher asked for a letter of introduction from the Head of Department, Social Sciences Education, so as to be able to seek for permission from the authorities of the schools that were involved in the research. And also to make proper arrangement as per the dates and times the researcher can administer the instruments (tests). The researcher and five trained research assistants administered the tests on the scheduled date and time. A period of 1hour was given to each testee that was administered test forms A and C while testees who attempted test form B had 1hour 20minutes each to answer the questions. The instruments were retrieved from the testees on conclusion of the process. Table 7 shows the schedule for data collection.

Table 8: Schedule for data collection

STATES	WEEKS	ACTIVITIES
Ogun	1	Giving out of introductory letters to schools
	2	Same as week 1
	3	Administration of instruments
	4	Same as week 3
	5	Same as week 4
Ekiti	1	Giving out of introductory letters to schools
	2	Same as week 1
	3	Administration of instruments
	4	Same as week 3
	5	Same as week 4
Ondo	1	Giving out of introductory letters to schools
	2	Same as week 1
	3	Same as week 2/ administration of instruments
	4	Administration of instruments
	5	Same as week 4

Ethical Consideration

Consent of the potential participants of the research work was sought and details of the purpose of the study were fully explained to them. After that, they were given room to willingly take part in the study. The participants were assured of anonymity as code numbers were given to them for identification instead of using their real names, hence their identity was not known. Participants were informed of their right to freely withdraw from the study at any time if they wish to and their data will not be used in the study. In addition, respondents will be assured of utmost confidentiality as their answer booklets cannot be traceable to them.

Data Analysis Techniques

The data collected from this study were analyzed with the use of descriptive statistics of mean, standard deviation, standard score deviate, percentile rank and coefficient of variation to answer the five generated research questions. Mean was used to answer research questions one and two. Research question three was answered using standard score deviates of T-score, this is used to transform raw scores into a common standard using mean and standard deviation. A computer programme known as common item programme of equating (CIPE) was used. Test form B was equated to test form A, test form C was equated to test form B and test form A was equated to test form C.

Percentile rank was used to answer research question four. Percentile rank is defined as the proportion of cases that fall below a given point on the measurement scale. It is the position on a scale of 100 to which an individual score lies which describes an individual's position in relation to a known group. Statistical Package for the Social Sciences (SPSS) was used to obtain

percentile rank for each score. Coefficient of variation statistic was used to answer research question five. Coefficient of variation statistic can be described as the ratio of the standard deviation to the mean. It was calculated by summing up testees' equated scores in the two equating methods. Computation of their means and standard deviations was done. It is translated in percentages.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

This chapter covered the data involved in this research work and the results of the analysis of the data. The data gathered from the responses were given to 1,461 SSS III students of Chemistry in South-West, Nigeria. This chapter dealt with presentation of results according to research questions raised, however, descriptive statistics of the respondents will first be presented.

Descriptive Statistics

Table 9: Sample Size of Respondents According to Gender and Test Forms Answered

	Gender		
Test form	Male	Female	Percentage
A	267	228	33.88
B	255	230	33.20
C	248	233	32.92
Total	770	691	100

Results in table 9 showed that out of the 1,461 students that were sampled, 770 (52.7%) respondents were male while 691 (47.3%) respondents were female. 495 (33.88%) respondents answered test form A, 485 (33.20%) respondents answered test form B and 481 (32.90%) respondents answered test form C.

Five research questions were generated in this study, research questions one and two were answered using mean. Standard score deviate of T-score was used to answer research question three, percentile rank was used to answer research question 4 while coefficient of variation statistic was used to answer research questions five.

Research question one

What is the profile of students' performance on the common items of SSCE Chemistry multiple-choice papers?

Mean and standard deviation of students' common items scores from test forms A, B and C were computed.

Table 10: Profile of Respondents' Performance in Common Items

Test form	No of students	Mean	Maximum	Minimum	Standard deviation	Skewness	Kurtosis
A	495	5.46	16	1	2.74	.514	.537
B	485	6.04	15	1	2.71	.917	1.487
C	481	7.36	17	1	3.58	.496	-.337

Results in Table 10 showed that the mean performance of respondents in common items in test form A was 5.46, while respondents' common items scores in test forms B and C had the mean performance of 6.04 and 7.36 respectively. Respondents' highest scores in the three test forms were 16, 15 and 17 respectively and their lowest score was 1. Also, test form A had standard deviation of 2.74, test form B with standard deviation of 2.71, and test form C had standard deviation of 3.58. The three test forms (A, B and C) had skewness of .514, .917 and .496 respectively. Their positive skewness values indicated that all the scores from common items of the test forms are clustered to the left at the low values, the distribution is moderately skewed. Difference in their mean performance suggested that there was difference in the examinees' proficiency level, as a result, Levine linear equating and chained equipercentile equating methods were found suitable to equate examinees' scores in this study. This was supported by the assertion of von Davier and Kong (2003) and Kolen and Brennan (2004), that Levine linear equating was preferred to be used if the standardized mean difference of the anchor scores in two samples is between 0.80 and 1.25.

Research Question two

What is the profile of students' performance on the unique items of SSCE Chemistry multiple-choice papers?

Mean scores of testees were used to determine the performance of respondents in unique items of the test forms.

Table 11: Profile of Respondents' Performance in Unique Items

Test form	No of students	Mean	Maximum	Minimum	Standard deviation	Skewness	Kurtosis
A	495	17.32	40	6	5.25	1.008	1.267
B	485	18.03	48	6	7.01	1.346	2.356
C	481	20.48	45	4	7.35	-1.072	0.983

The result in Table 11 showed the performance of testees on the unique items. Test form C with a mean performance of 20.48, standard deviation of 7.35, skewness of -1.072 and kurtosis of .983 indicated best performance. The skewness value of -1.072 showed that test form C scores clustered to the right at the high values and the kurtosis value of 0.983 indicated that the distribution of scores is relatively flat. Test form B with better performance had mean score of 18.03, standard deviation of 7.01, skewness of 1.346 and kurtosis of 2.356 while testees who did test form A had the least performance. The mean score of test form A was 17.32 and standard deviation of 5.25. The skewness and kurtosis values are 1.008 and 1.267 respectively.

Research Question three

What are the results of Levine Linear Equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using standard score deviates?

Raw scores and their equivalent scores on the three test forms are shown in table 12 using Levine linear equating method under NEAT design. The scores obtained from the three test forms were standardized using standard score deviates following the equating of the scores. NECO scores were equated to WAEC scores which gave equated scores of WAEC, NABTEB scores were equated to NECO scores, and this resulted to equated scores of NECO. Likewise, WAEC scores were equated to scores obtained from NABTEB and this resulted to equated scores of NABTEB.

Table 12: Levine Linear Equating of WAEC, NECO and NABTEB SSCE Chemistry Multiple-choice Papers

Raw Scores	Equated WAEC Scores	Equated NECO Scores	Equated NABTEB Scores
0	0.98	-3.25	-2.69
1	1.11	-2.07	-1.59
2	1.98	-0.88	-0.48
3	2.45	1.31	0.63
4	3.79	2.49	1.74
5	4.95	3.28	2.85
6	5.7	4.86	3.96
7	7.45	6.98	5.07
8	8.2	7.86	6.17
9	9.94	8.01	7.28
10	10.69	9.22	8.39
11	11.14	9.8	9.5
12	11.96	10.98	10.61
13	12.32	12.17	11.72
14	13.01	13.35	12.83
15	14.23	14.54	13.93
16	15.01	15.33	15.04
17	16.89	16.91	16.15
18	17.99	18.1	18.45
19	19.42	19.29	19.37
20	20.17	20.47	21.97
21	20.92	21.66	22.56
22	21.67	22.84	23.78
23	22.42	23.22	26.65
24	23.17	24.13	25.46
25	23.91	24.87	26.98
26	24.66	25.21	27.59
27	25.41	25.72	27.79
28	26.16	26.77	28.35
29	26.91	27.2	29.46
30	27.65	28.11	30.56
31	28.4	29.51	31.67
32	29.15	30.22	32.78
33	29.9	31.26	33.89
34	30.65	32.98	35
35	31.4	33.1	36.11
36	32.14	33.94	37.22
37	32.89	34.2	38.32
38	33.64	35.35	39.43
39	34.39	35.93	40.54
40	35.14	37.05	41.65
41	-	38.44	42.45
42	-	39.89	42.99
43	-	40.76	43.55
44	-	41.22	33.42
45	-	42.31	44.79
46	-	43.98	-
47	-	45.08	-
48	-	47.76	-

A raw score of 20 was found equivalent to equated scores of 20.17, 20.47 and 22.97 in WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers respectively and a raw score of 25 was equivalent to equated scores of 23.91, 24.87 and 26.98 in WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers respectively. Also, a raw score of 30 was found equivalent to equated scores of 27.65, 28.11 and 30.65 in WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers respectively and a raw score of 36 was equivalent to equated scores of 32.14, 33.94 and 37.22 in WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers respectively. Using Levine linear equating under NEAT design, it was found out that WAEC and NECO multiple choice items were more equivalent when compared to NABTEB multiple choice items. Figure 3 shows a line graph of Levine linear equating of equated scores from test forms A, B and C (WAEC, NECO and NABTEB) SSCE Chemistry multiple choice papers.

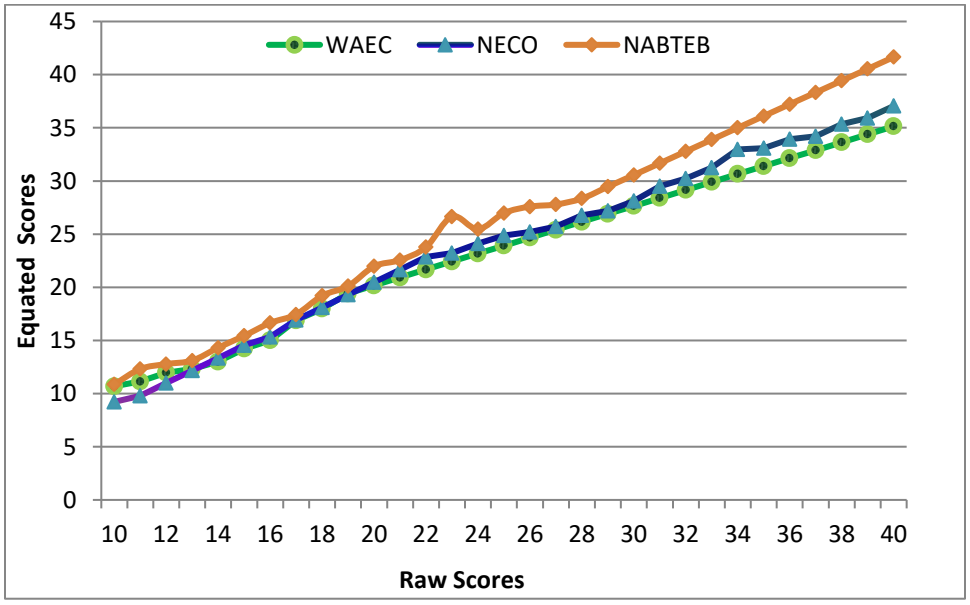

Fig. 2: Levine linear equating of equated scores from test forms A, B and C (WAEC, NECO and NABTEB) SSCE Chemistry multiple choice papers

Graph presented in figure 2 showed the scores obtained from the three test forms. Test form C had higher values of equated scores while test forms A and B had very close values. It can be said that the two test forms (A and B) produced almost equivalent scores.

Research Question four

What are the results of chained equipercentile equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers using percentile ranking?

Percentile ranks were used to determine the equivalence of scores obtained from WAEC, NECO and NABTEB such that scores with the same percentile ranks were considered equivalent.

Table 13: Summary Showing Chained Equipercentile Equating of WAEC, NECO and NABTEB SSCE Chemistry Multiple-choice Papers

WAEC (FORM A)	NECO (FORM B)	NABTEB (F0RM C)
9	8	10
10	9	11
11	10	12
14	14	16
17	17	19
18	18	21
20	20	23

Results in Table 13 revealed that scores of 9, 10, 11, 14, 17, 18 and 20 in WAEC (test form A) was equivalent to scores of 8, 9, 10, 14, 17, 18 and 20 in NECO (test form B) and same with scores of 10, 11, 12, 16, 19, 21 and 23 in NABTEB (test form C). That is, scores of 9 in WAEC was equivalent to 8 and 10 in NECO and NABTEB respectively and scores of 10 in WAEC was equivalent to 9 and 11 in NECO and NABTEB respectively. Also, scores of 14 in WAEC was equivalent to 14 and 16 in NECO and NABTEB respectively. Likewise, scores of 17 in WAEC was equivalent to 17 and 19 in NECO and NABTEB respectively, scores of 18 in WAEC was equivalent to 18 and 21 in NECO and NABTEB respectively and scores of 20 in

WAEC was equivalent to 20 and 23 in NECO and NABTEB respectively. Based on chained equipercentile equating using the NEAT design, it could be said that WAEC, NECO and NABTEB multiple choice items tended to be equal on the other hand WAEC and NECO were more equivalent. Figure 4 is a graphical representation of a line graph showing percentile rank of SSCE Chemistry multiple choice papers scores.

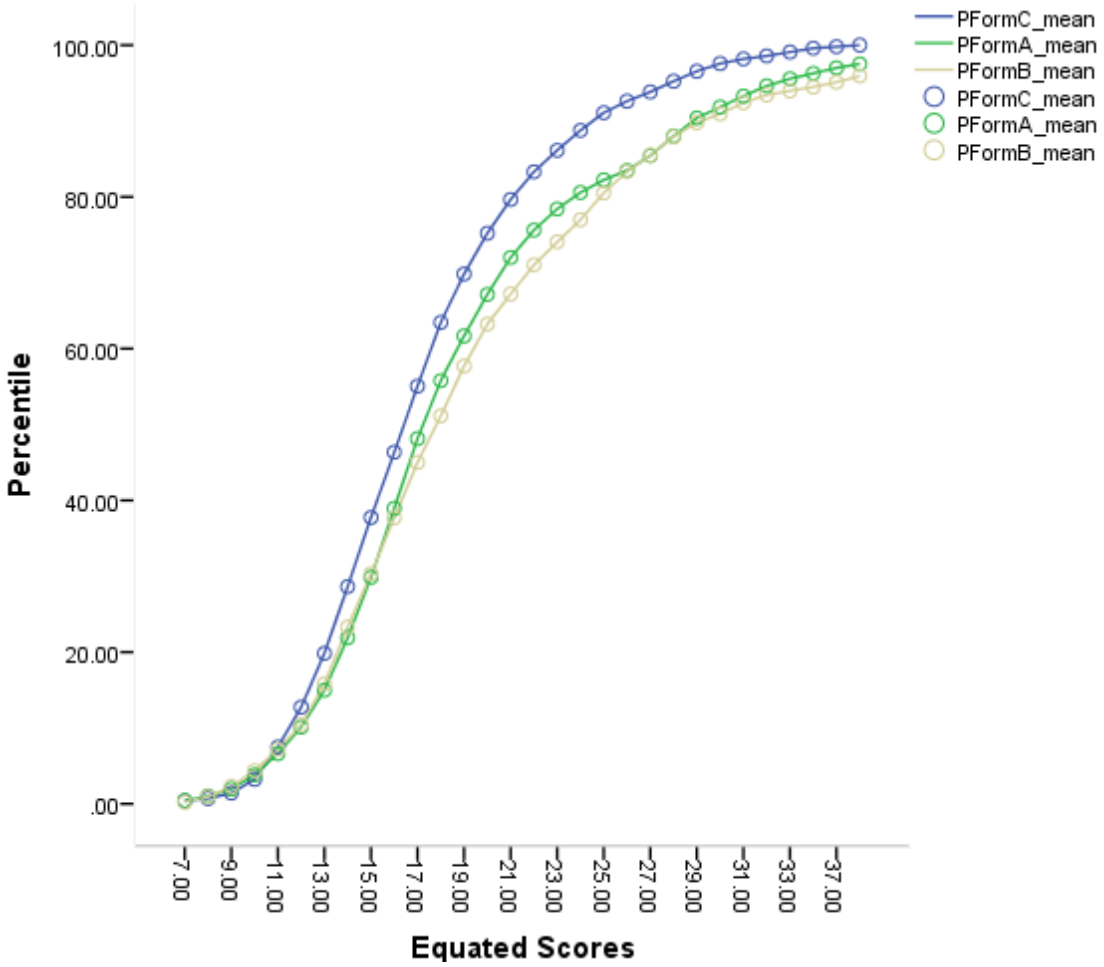

Fig. 3: Percentile rank of scores on WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers

The graph in figure 3 showed that scores from all the test forms are comparable, the distribution of scores had similar shape. Though scores obtained from test forms A and B (WAEC and NECO) were more equivalent than those from test form C (NABTEB).

Research Question five

How invariant are the equated scores of WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers across equating methods?

To examine how invariant the equated scores of SSCE Chemistry multiple choice papers were, testees' equated scores were independently summed and means and standard deviations were computed so that coefficient of variation of the equating methods were calculated.

Table 14: Summary Showing the Invariance in the Score Obtained on the Forms from the Two Equating Methods.

Equating method	Test Forms	Mean	Standard Deviation	Sum of Means	Sum of Standard Deviations	Coefficient of Variation
Levine Linear Method	WAEC	17.32	5.25	55.82	19.62	35.1
	NECO	18.03	7.02			
	NABTEB	20.47	7.35			
Chained Equipercentile Method	WAEC	21.2581	9.54	72.53	33.0	45.5
	NECO	25.3846	12.05			
	NABTEB	25.8890	11.41			

Results in Table 14 revealed that using Levine linear equating method 55.82 was the mean and 19.62 was the standard deviation of the equated scores and 35.1% was recorded as the coefficient of variation. 72.53 was the mean and 33.0 was recorded as the standard deviation of equated scores when chained equipercentile equating method was used and a higher coefficient of variation of 45.5% was obtained. These results brought about the lower coefficient of variation for Levine linear equating method across the three test forms when compared to variation recorded using the chained equipercentile method.

Since the purpose of equating was to reduce bias due to variance in the difficulty level of items there by leading to measurement error. It could therefore be said that the lower the variance in equating scores the better the test. Therefore, estimating test equating of the three test forms, Levine linear equating method was the best having reduced the variance in test forms.

Summary of Findings

The key results of this study are:

1. The mean performance of testees in common items in the three test forms A, B and C showed that there was difference in the examinees' proficiency level, thus, Levine linear equating and chained equipercentile equating methods were found suitable to equate examinees' scores in this study.
2. Testees performed differently in the unique items across the three test forms. Performance of testees in test form C indicated a high performance when compared to test forms A and B.
3. The result of Levine linear equating method after equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers showed consistency in the pattern of equated scores from raw score of 17.
4. The result of Chained equipercentile equating after equating of WAEC, NECO and NABTEB SSCE Chemistry multiple-choice papers showed consistency in the pattern of equated scores from a score of 6.
5. It was found out that Levine linear equating method had a lower coefficient of variation across the three test forms when compared to chained equipercentile method with higher coefficient of variation. Levine linear equating method was therefore, found to be more preferred than chained equipercentile equating method because it had a lower coefficient of variation.

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

This chapter is concerned with discussion and conclusion based on data analyzed and results presented in the preceding chapter. Recommendations and suggestions were made based on the results obtained from this study. This study aimed at equating the multiple-choice Chemistry items of WAEC, NECO and NABTEB Senior School Certificate Examination (SSCE) using Levine linear equating and Chained equipercentile equating methods.

Discussion of findings

The results of the study revealed that examinees differed in their proficiency level because the mean performance on their common items scores on the three test forms A, B and C were 5.46, 6.04 and 7.36 respectively. This result might be attributed to differences in the proficiency level of testees who sat for the three test forms, this implied that the use of Levine linear equating, which is a linear equating method used when testees' abilities differed and chained equipercentile equating methods, an equipercentile equating method are suitable for this study. This is in agreement with the findings of von Davier and Kong (2003), Kolen and Brennan, (2004) and Wang, (2008). The researchers all found out that when the proficiency level of the examinees who sat for the different test forms are at variance, Levine linear and chained equipercentile equating methods were found appropriate to equate the test forms.

In unique items, this study showed that there were differences in the performance of examinees across the test forms. Unique items of test forms A, B and C had mean performance of 17.32, 18.03 and 20.48 respectively. The highest mean value for test form C might be as a result of most items of NABTEB chemistry multiple choice paper falling under the lowest level of learning outcomes in the cognitive domain, that is, knowledge and comprehension levels, thereby, making it have more relatively easy items. This result is in line with the findings of Kolen and Brennan (2004) that constructing multiple forms of tests that are parallel is almost impossible. Equating therefore, becomes necessary because it adjusts for differences in difficulty across test forms that are constructed as similar as possible in difficulty and content just like the senior school certificate examinations that were examined in this study.

Results from Levine linear equating of WAEC (test form A), NECO (test form B) and NABTEB (test form C) Senior School Certificate Chemistry multiple choice papers with the use of standard score deviates showed equivalence in the test scores obtained from both test forms A and B which was not so with test form C. This could be as a result of similarity in the distribution of questions constructed by WAEC and NECO as was specified by Okoye and Nwafor (2009) that the similarity in the distributions of questions set by the two testing agencies could be accounted for in terms of the fact that they both essentially draw their examiners from the same pool. A report was therefore given that a significant difference does not exist between the distributions of questions by WAEC and NECO for biology, chemistry and physics. This finding corroborates the assertion of Obinne, Nworgu, and Umobong (2013) that there is no significant difference in the DIF of Biology tests conducted by WAEC and NECO. This was also supported by the submission of Obinne (2011) that Biology test conducted by WAEC and NECO were equally reliable. The finding is contrary to that of Olutola (2011) who found out that

WAEC SSCE multiple-choice Biology items were more difficult than NECO SSCE multiple-choice Biology items. This could be attributed to the findings of Obioma and Salau (2007) who discovered that, out of all Nigerian O'Level examination bodies, students' performance in WAEC was the best way to analyse performance of students in higher institution.

Results from Chained equipercentile equating of WAEC (test form A), NECO (test form B) and NABTEB (test form C) Senior School Certificate Chemistry multiple choice papers with the use of percentile ranking showed that scores of 9 in WAEC was equivalent to scores of 8 and 10 in NECO and NABTEB respectively. Similarity in how effective the three examination bodies are and the degree of reliability of their SSCE Chemistry multiple choice items might be responsible for this result. This finding is in line with that of Bandele and Adewale (2013) whose study revealed that WAEC, NECO and NABTEB were comparable and equivalent when the validity and reliability coefficient of Mathematics achievement examination conducted by the three testing agencies were compared. The finding disagrees with Alfred's (2011), that there was a significant difference in the difficulty level of Economics multiple choice items conducted by WAEC, NECO and NABTEB.

Chained equipercentile equating also revealed that scores of 14 in WAEC was equivalent to 14 in NECO and 16 in NABTEB. This result might be due to long existence of WAEC and NECO as examination bodies that conduct SSCE and have acquired a lot of experiences over time. This finding validates the outcome of the study of Udofia and Udoh (2017) that WAEC and NECO senior secondary Mathematics examination are similar and comparable at .05 level of significance. This finding negates that of Salako, Adegoke and Ogundipe (2017) that WAEC and NECO successes in Mathematics and Physics were not correlated. Further results revealed by

chained equipercentile equating in this study showed that though multiple choice items of WAEC, NECO and NABTEB tended to be equal, WAEC and NECO were more equivalent.

From this study, it was found out that the equated scores of Senior School Certificate Chemistry multiple choice items from Levine linear equating with coefficient of variation of 35.1% was different from that of chained equipercentile equating with coefficient of variation of 45.1%. Levine linear equating with lower coefficient of variation was found better to use in this study which was contrary to the research result of Sinharray and Holland (2009), who found that chained equipercentile equating method is more satisfactory than other equating methods. It can therefore be said that equating scores with lower variance show a better test. The variance noticed in the equated scores of WAEC, NECO and NABTEB SSCE Chemistry multiple choice papers across equating methods was as a result of variance in test forms difficulty level in addition to the differences in the score distributions.

Ozdemir (2017) who equated Trends in International Mathematics and Science Study (TIMSS) mathematics subtest scores obtained from TIMSS 2011 to scores obtained from TIMSS 2007 test form with different equating methods, included Levine and chained equipercentile equating methods and discovered that Levine equating method with lower bias, outperformed chained equipercentile equating method and was better to use, this supports the finding of the present study. This finding is contrary to that of Livingston and Kim (2009) who compared Levine linear, Chained equipercentile, Chained linear, Chained mean and Identity equating methods. Findings from the study showed that Chained equipercentile method performed better when than other equating methods.

Conclusion

From the findings of this study, it could be concluded that Levine linear equating was found more efficient than Chained equipercentile method for equating of Chemistry SSCE scores. This is because the level of invariance of Levine linear equating method is smaller when compared with chained equipercentile equating. It could also be concluded that the examination standard of both WAEC and NECO are comparable and of similar standard when compared with NABTEB.

Implications of the Findings of the Study

The implication of this study based on the results discussed above is this:
The most efficient method for each subject at each level should first be determined by test experts before equating of test scores since different methods of equating test scores exist apart from the two used in this study. Inappropriate use of equating method might lead to inaccurate results.

Limitations of the Study

An observable short coming of this study was the use of Classical test theory, which is known to have a major weakness, owing to the limitation of generalization because test accuracy is often solely attached to the study population, there can therefore be no generalization of the results of this study. Regardless of this limitation, the researcher still adjudge the findings of this

study valid since the study was done empirically and relevant statistical methods were used to analyse the data. Therefore, findings from this study could be considered valid and reliable.

Recommendations

On the basis of the findings of this study, the following recommendations are proposed:

1. Examiners and recruiting firms are encouraged to use Levine linear equating method in equating Chemistry scores.
2. Uniformity in standards of all examination bodies should be encouraged and this can be attained by employing experts in measurement and evaluation who will serve as monitoring team for constructing and conducting of examination as well as certification.
3. Test developers are encouraged to adopt equating method with lower coefficient of variation (Levine linear equating) for equating test scores.

Suggestions for further studies

This study equated WAEC, NECO and NABTEB Chemistry multiple choice papers using Levine and chained equipercentile equating methods under non-equivalent anchor groups test design in South-west Nigeria. Prospective researchers can replicate the study involving other school subjects at the senior secondary level in another geo political zone across years. This study adopted a sample size of 1,461 testees, further studies can accommodate larger testees in south west Nigeria. Further studies can be carried out by equating scores of essay tests. It can be extended to equating of examination scores at tertiary education level. Different equating methods can be compared using other equating design and subsequent studies can use item

response theory equating approach instead of classical test theory that was used in this study. Also, other studies on equating can carry out a comparative analysis of CTT and IRT methods.