Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Original Article

Equating Two Alternate Forms of Basic Statistics Test Using the Single Group Random Design: An Application of the Traditional Equating Methods

Frank QUANSAH[1]

[1] Department of Educational Foundations, School of Education and Life-Long Learning, University of Education, Winneba, Ghana

Correspondence: fquansah@uew.edu.gh

Abstract

Most testing situations in higher education institutions utilise different tests which measure similar or the same psychological construct to safeguard the security of the various test and to improve the validity of scores obtained from the examinations. This requires strategies like developing parallel tests and equating the scores. This paper demonstrates the application of traditional equating methods to equate two alternate forms of Basic Statistics test using the single group random design. Two alternate test forms (Form X and Form Y) were developed by test experts with measurement and statistics background. The development of the test forms was closely guided by the test specification table and item specification blueprint. The tests were administered to 146 students who were sampled through the convenience sampling technique. Half of the sample (n=73) took Form X and the other half were administered Form Y. The findings showed that relatively, the equipercentile equating appeared to produce scores that were similar and also within the range of the Form X scores. Consequently, the equipercentile equating was found to be statistically accurate for equating.

KEYWORDS: Equating, students, equipercentile, linear equating, mean equating, item specification, test specification table

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

1.0 INTRODUCTION

The use of multiple test forms has become necessary in contemporary times due to the threatened validity of scores from the test (Diego, 2017). Quite recently, Quansah and Cobbinah (2021) recommended the use of parallel test forms to curb the issue of examination malpractices and promote fairness in testing where test items are shuffled. The need to develop multiple test forms are motivated by the issue of test security which threatens the validity of test results. Test security is an important reason why experts are required to develop multiple test forms. In many high stakes testing programmes (such as licensure examination, entrance exams for educational institutions, university examinations, West African Secondary School Examination, Graduate Record Examination, among others), test security is a concern and multiple test forms are needed for examinees who are sitting for the examination on more than one occasion, those taking the examination on different days, or those closely seated beside themselves (Petersen, Kolen, & Hoover, 1989), especially when samples of these test get to the public domain.

Test equating is one of the rigorous means by which parallel forms of a test are used to safeguard test security and still ensure that fairness still prevails. According to Kolen and Brennan (2004), test equating is a statistical method employed to adjust obtained scores from test forms such that these scores can be interchangeably utilised. From Crocker and Algina's (1986) view, test equating is a process that determines comparable scores from two diverse measurement tools; they further indicated that when the percentiles equivalent to the X and Y obtained scores from distinct tests that have equal dependability and measure the similar construct are equal. Angoff (1984) also described test equating as the process of changing the unit system from one test form to the unit system of another test form, highlighting that obtained scores from different forms are equated after the scores are transformed. Consequently, test equating emerged because two or more tests forms that measure the same content and construct can produce different scores for the same examinees.

Equating reflects a statistical process employed to transform obtained scores from one test form to the scale of another test form. Through this statistical process, some conditions need to be met to equate two test forms. Literature mentions several distinct views associated with these conditions in the literature. Hambleton and Swaminathan (1985), for example, outlined these conditions to include equity, symmetry properties, group invariance and the same specifications. With regards to the symmetry property, the inverse of the transforming scores on Form X to Form Y scores should be valid (Kolen & Brennan, 2004). For the condition of the same specifications, both test forms to be equated needs to be similar in terms of statistical properties and content. The property of equity requires that indifference towards whether Form X or Form Y is administered to the test takers must be demanded. This condition, however, holds only when test forms are equivalent (Lord, 1980). To satisfy the group invariance condition, equating test forms should be independent of the examinee group; it does not matter which group is chosen for calculating the equating function between the scores from Form X and Form Y (Öztürk & Anıl, 2012). There are several procedures related to transforming forms (Dorans, Moses, & Eignor, 2010); these methods have been classified according to a theory based on methods of classical test theory and item-response theory (IRT). Classical test theory-based equating methods include identity equating, mean equating, linear equating, and equipercentile equating. This paper focuses on the classical test theory-based equating.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

1.1 Mean Equating Method

The mean equating methods operates on the principle that Form X is differentiated from Form Y in difficulty by a constant amount over the score scale (Kolen & Brennan, 2004). Under this method, no difference exists in examinees' ability levels. The scores of the two forms are determined using Equations 1 and 2.

$$x-\mu x = y - \mu y \qquad (1)$$

$$my(x) = y = x - \mu x + \mu y \qquad (2)$$

In these equations, x is the score from Form X; μx is the mean of Form X; y is the score from Form Y; μy is the mean of Form Y, and my(x) is the score transformed from x on Form X to Form Y by using mean equating.

1.2 Linear Equating

According to Crocker and Algina (1986), the linear-equating method operates on the assumption that the distributions of scores on Form X and Form Y are similar, but their means and standard deviations are dissimilar. This method of equating is employed when the standard scores obtained from these forms are deemed equivalent. From Donlon's (1984) view, if groups of examinees who were administered different test forms have equal levels of ability, then linear equating can be performed. Angoff (1984, p. 564) defined linear equating as "scores being equivalent when the scores on different forms of the test have similar standard-score deviations".

1.3 Equipercentile Equating

The equipercentile approach to equating is utilised when the score distributions of the forms are dissimilar. In this form of equating, Form X may have a high difficulty level compared to Form Y for high and low scores, nevertheless, it may be less difficult for middle scores (Kolen & Brennan, 2004). Thus, corresponding percentile ranks in both Form X and Form Y are used (Kolen, 1988). If the score distribution on Form X which transformed to scores on Form Y is equivalent to the score distribution on Form Y, the equating function between the two forms is known as the equipercentile equating function (Kolen & Brennan, 2004). The equipercentile equating method starts with the computation of percentile ranks for each test form. Hence, scores that relate to the same percentile rank are equal (Kolen, 1988; Livingston, 2004). These equating processes operate on the principle that the test scores are continuous variables. When test scores are discrete variables, the equipercentile equating function cannot be applied; but they are discrete in the real sense. Consequently, the discrete variables are operationalised as continuous variables by transforming scores into percentiles or percentile groups to overcome this challenge (Kolen & Brennan, 2004). In instances where an examinee receives no score when using the equipercentile equating, the middle point of the range of scores that correspond to the same percentile group is chosen as being equivalent.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

1.4 The Purpose

The purpose of equating in this writeup is to compare the results of the three traditional methods of equating as applied to the two forms indicated to equate two forms of achievement tests so that they can be used interchangeably. These three traditional equating methods are mean, linear, and equipercentile equating. Given this, two forms of achievement tests in Educational Statistics were used. This write up was guided by the seven guidelines or steps for conducting equating as outlined by Kolen and Brennan (2014). These steps include the following: (1) Decide on the purpose for equating, (2) Construct alternate forms, (3) Choose a design for data collection, (4) Implement the data collection design, (5) Choose one or more operational definitions of equating, (6) Choose one or more statistical estimation methods, and (4) Evaluate the results of equating. This paper is relevant in terms of demonstrating how two alternate test forms can be equated in a more practical sense. Further, this paper can also be used as instructional materials for students when teaching test equating.

2.0 METHODS AND MATERIALS

2.1 Design

The single random group design was adopted for this study- a type of design where two forms are administered to one group such that every test taker answers only form (Kolen & Brennan, 2004). Operating within the framework of the design, the administration of the alternate forms of the test was done using the spiral approach based on seating arrangement. For example, the first candidate was given the Form X, the second candidate administered the Form Y, the third examinee responded to the Form X, the fourth examinee was given Form Y, and so on. According to Crocker and Algina (1986), the determination of an equating design should be largely based on practical situations like test development challenges, test administration complexities, and meeting statistical assumptions. Meanwhile, the random group design was considered appropriate because it produces a relatively small error, compared with other data collection designs.

2.2 Participants

The study comprised 146 pre-service students in one of the colleges of education in Ghana. The participants were third-year students who were registered for the Educational Statistics course. The sample was obtained through convenience sampling. The students included 65.1% (n=95) females and 34.9% (n=79). The majority of the students were Christians (74.7%, n=109), 21.2% were Muslims (n=31), and 6 others did not indicate their religious affiliation. The mean age was 22.8.

2.3 Measures

Two alternate achievement tests were developed by Measurement and Evaluation experts based on the course outline and course/learning objectives of the Educational Statistics course. Because the tests were achievement tests (i.e., measuring the amount of learning students have acquired after instructional periods), the focus was on determining students' competency on the Educational Statistics course content (Quansah & Nugba, 2021). As at the time of the study,

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

seven topics were covered by the tutor which includes discrete/categorical and continuous data, descriptive/qualitative and quantitative data, parameters and statistics, frequency distribution, graphical organization of data, measures of central tendency, and measures of dispersion and variability.

The development of the test items for both Form X and Form Y tests went through a series of stages. First, the test specification table (see Table 1) and item specification blueprints (see appendix) were developed and utilised for the item development. The development of the items following all the test development protocols to ensure content and construct validity. For example, the item difficulty levels were varied, providing clear instructions, ensuring that distractors are effective, among others (Nitko, 2001). After the items were developed, two experts with measurement and statistics background reviewed the items to cross-validate the items for appropriateness. The majority of the items (11) measured knowledge, whereas the least number of items (1) each on application and analysis

Table 1: Test Specification Table

Content base category	**Level of cognitive operation**				
	KNG	**COM**	**APP**	**ANAL**	**Total**
Discrete/categorical and continuous data	1	1	-	-	**2**
Descriptive/qualitative and quantitative data	1	1	-	-	**2**
Parameters and statistics	-	2		-	**2**
Frequency distribution	4	-	-	-	**4**
Graphical organization of data	2	-	-	-	**2**
Measures of Central Tendency	2	1	1	-	**4**
Measures of Dispersion and Variability	1	1	1	1	**4**
Total	**11**	**6**	**2**	**1**	**20**

KNG- Knowledge; COMP- Comprehension; APP- Application; ANAL- Analysis

Before writing the actual items, three principles were adhered to: first, assessment tasks should focus on the important learning targets, in addition, tasks should elicit from students only the knowledge and performance relevant to the learning targets being assessed, and tasks that appear on assessment should neither prevent nor inhibit a student's ability to demonstrate that he has gained mastery on learning targets (Nitko, 2001). Based on these principles, the test was carefully planned. The learning targets were to be measured using objective-type items (multiple choice). Because three forms of tests were to be developed, item specifications were developed. These included the topic, objective, description of the item, and a sample of the item (see Appendix A). The essence of the item specification was to ensure that the two forms

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

of tests were as similar in content and statistical specification as much as possible. Based on this the forms were constructed and labelled Form X and Form Y.

2.4 Procedure

The data collection commenced after approvals have been obtained from the college principal after ethical approval was obtained. The researcher visited the school to farmiliarised themselves with the students and staff and also to brief them on the purpose of the research. Ethical standards were followed, which include informed consent (through signing a consent form), volition, protection of vulnerable participants, anonymity, and confidentiality. The students who were willing and available to participate in the study were assembled in one of the lecture rooms and the administration was done. Based on the dictates of the random group design, all the 146 students who took the test were seated while the two forms were administered. The forms were administered in a spiral manner based on the seating arrangements of the test takers. This was done in serpentine order. In all, 73 of the test takers took Form X and the other 73 also took Form Y. Both test forms were taken at the same sitting.

2.5 Data Analysis

The data analysis process began with the coding of students' responses to each test item from both Form X and Form Y using SPSS (Statistical Package for the Social Sciences). Each item was scored dichotomously (1 = correct, 0 = incorrect), and raw scores were computed for each participant. The dataset was thoroughly screened for entry errors, missing values, and inconsistencies. Any anomalies identified were addressed through cross-checking with original answer scripts and correction logs to ensure accuracy and completeness. As a preliminary step, descriptive analyses were conducted separately for each test form. This included (1) the frequency distributions of total scores to inspect score dispersion and detect irregularities, (2) measures of central tendency and variability (mean and standard deviation) to compare the relative difficulty levels of the two forms and (3) shape indices such as skewness and kurtosis were examined to assess the normality of the score distributions—a key consideration for selecting appropriate equating methods. These descriptive statistics provided a foundational understanding of how the forms performed and whether further adjustments via equating were warranted.

Three traditional equating methods, namely mean, linear, and equipercentile, were applied to adjust the raw scores from Form X and Form Y to a common scale, based on the assumption that the two forms were content-equivalent. This method was implemented in SPSS. It involved calculating the mean difference between the two forms and adjusting the scores accordingly. The assumption underpinning this method is that both test forms have equivalent score variability (i.e., standard deviations are approximately equal), and any difference is attributed solely to central tendency.

Linear equating adjusts for differences in both the mean and standard deviation between forms. It assumes a linear transformation relationship between the two score distributions. This procedure was carried out using the RAGE-RGEQUATE software, which enabled the computation of transformation constants and applied the linear model to produce equated

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

scores for each raw score on the alternate form. Equipercentile equating was also performed using RAGE-RGEQUATE. This method equates scores by matching corresponding percentiles across the two score distributions. It does not assume equivalent shape or spread but instead aligns score ranks. To account for sampling variability and to improve estimation accuracy, log-linear smoothing was applied to the score distributions prior to equating. The software generated an equating function and score concordance table for converting scores from one form to the other. Post-equating, the equated scores from all three methods were compared against the observed scores. The consistency, closeness, and interpretability of equated scores were evaluated to determine the most appropriate equating approach for the given dataset.

3.0 RESULTS

3.1 Descriptive Information

This section presents the results obtained. First descriptive information such as frequency counts, mean, standard deviation, skewness, and kurtosis are presented in Table 2.

Table 2: Frequency Distribution of Scores based on Forms

SCORES	FORM Y	FORM X
0	0	0
1	0	0
2	0	0
3	0	0
4	0	1
5	1	0
6	3	3
7	4	6
8	9	8
9	7	18
10	19	14
11	14	13
12	10	6
13	3	3
14	2	0
15	0	1
16	1	0
17	0	0
18	0	0
19	0	0
20	0	0

From Table 2, the lowest scores for Forms Y and X are 5 and 4, respectively. The highest scores from Forms Y and X are 16 and 15, respectively. The distribution has been plotted in Figure 1.

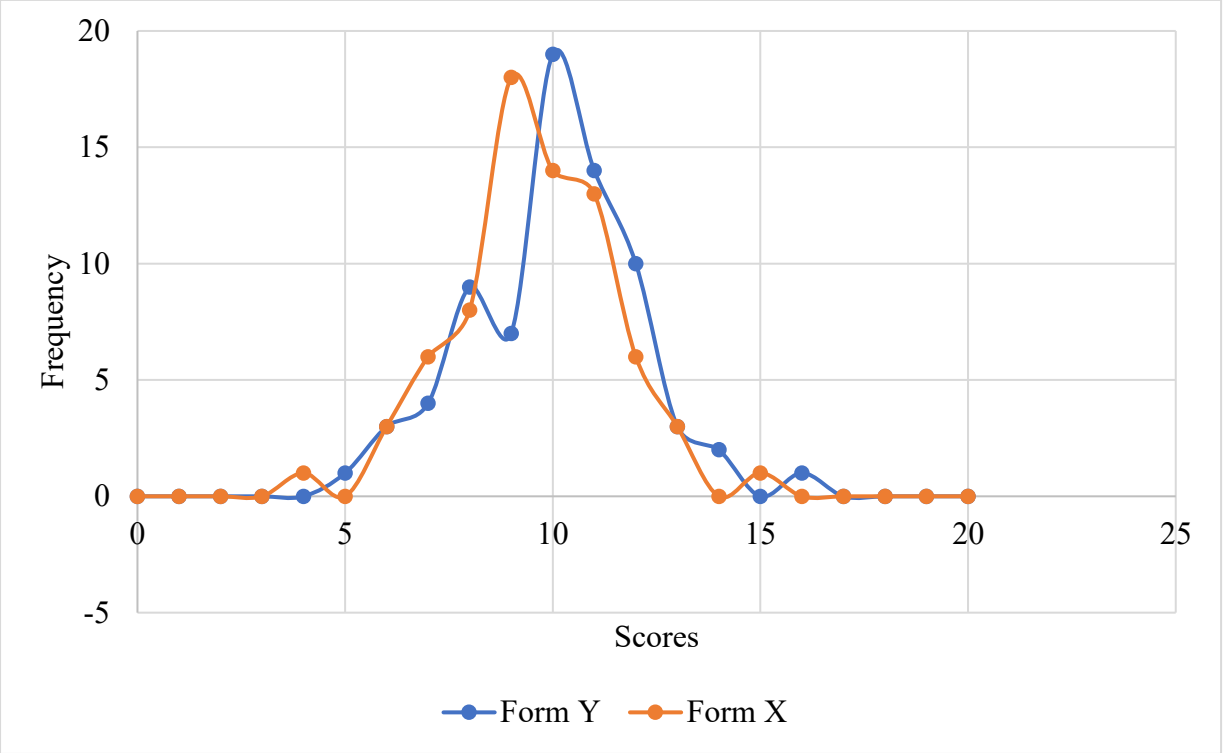

Figure 1: Raw score distribution for Forms X and Y

Table 3: Descriptive Statistics

Forms	μ	σ	sk	ku
Form Y	10.0411	2.04422	-.037	.410
Form X	9.5753	1.91426	-.088	.594

From Table 3, it appears the performance was better for Form Y (M = 10.0411) than Form X (M = 9.5753). Relatively, the spread of Form Y was larger than that of Form X based on the standard deviations. Scores on both forms were negatively skewed. The kurtosis for Form X is higher than Form Y. Generally, it can be said that Form X appears to be difficult than Form Y. Therefore, the scores on Form X were equated to that of Form Y using mean, linear, and equipercentile equating methods. This is to adjust for the differences in the difficulties in the forms so that they can be used interchangeably.

3.2 Equating Results

Mean equating was performed using SPSS, whereas RAGE-RGEQUATE software was used for the linear and equipercentile equating. The mean equated form of Form X to Y equivalent was computed using the formula as defined in equation 2: $my(x) = y = x - \mu x + \mu y$. This formula was computed in SPSS and the output generated is presented in Table 4.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Table 4: Form Y Equivalent Using Mean Equating

Form X Score	Mean equating
0	.466
1	1.466
2	2.466
3	3.466
4	4.466
5	5.466
6	6.466
7	7.466
8	8.466
9	9.466
10	10.466
11	11.466
12	12.466
13	13.466
14	14.466
15	15.466
16	16.466
17	17.466
18	18.466
19	19.466
20	20.466

As shown in Table 5, the Form Y score equivalent of 0 using mean equating is .466, a score of 1 is also equivalent to 1.466, and that of 20 is 20.466. In the case of linear and equipercentile equating, RAGE-RGEQUATE software was used.

To use RAGE-RGEQUATE software, scale scores need to be created. These scale scores are scale score equivalents of Form Y. These scores were constructed by following the six steps outlined in p.396 of the book by Kolen and Brennan (2014). The first step is to construct the relative frequency distribution of scores, g(y). The second step is to smooth the relative frequency distribution. This smoothed relative frequency distribution is referred to as adjusted relative frequency. The third step is to find the percentile ranks of the smoothed distribution. To find the percentile rank, cumulative adjusted relative frequencies are then computed. You then multiply each of the cumulative adjusted frequencies by 0.5 to get a modified adjusted cumulative relative frequency. These modified adjusted cumulative relative frequencies are then added to cumulative adjusted relative frequencies to get the percentile ranks. The fourth step is to transform the percentile ranks into z-scores, with a mean of 0, and a standard deviation of 1. For the fifth step, because of the negatives associated with the z-scores, the z-scores are either transformed to stanines, T-scores or transformed to normal curve equivalents (Kolen & Brennan, 2014, p. 396). For this article, the z-scores were transformed to T-scores. The last step requires that the resulting scale scores be rounded to the nearest whole number. Table 5 presents the relative frequencies, percentile ranks, z-scores, and T-scores.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Table 5: Relative frequencies, percentile ranks, z-scores, and T-scores

Scores	Relative frequency	Percentile ranks	z-score	T-score
0	.0000018	.00000090	-4.77467243	2
1	.0000151	.00000935	-4.27987299	7
2	.0001038	.00006880	-3.81244294	12
3	.0005777	.00040955	-3.34625849	17
4	.0026006	.00199870	-2.87836684	21
5	.0094255	.00801175	-2.40837984	26
6	.0273875	.02641825	-1.93625481	31
7	.0635251	.07187455	-1.46197121	35
8	.1171131	.16219365	-.98548214	40
9	.1708662	.30618330	-.50669819	45
10	.1964363	.48983455	-.02548376	50
11	.1771848	.67664510	.45833776	55
12	.1248515	.82766325	.94497136	59
13	.0684303	.92430415	1.43463302	64
14	.0290478	.97304320	1.92753010	69
15	.0095085	.99232135	2.42384166	74
16	.0023899	.99827055	2.92370714	79
17	.0004592	.99969510	3.42721769	84
18	.0000672	.99995830	3.93441378	89
19	.0000074	.99999560	4.44473566	94
20	.0000006	.99999960	4.93536744	99

Having obtained the T-scores, these scores were inputted together with the frequencies for Forms Y and X into the RAGE-RGEQUATE software in order to run the equipercentile and linear equating analysis. Table 6, therefore, summarises the raw-to-raw score conversion for Form Y equivalents using the three equating methods.

Table 6: Raw-to-raw Score Conversion

	Form Y equivalent using equating method		
Form X Score	Mean	Linear	Equipercentile
0	.466	-0.184	0.000
1	1.466	0.884	1.000
2	2.466	1.951	2.000
3	3.466	3.019	3.000
4	4.466	4.087	5.000
5	5.466	5.155	5.000
6	6.466	6.223	6.000
7	7.466	7.291	7.250

8	8.466	8.359	8.167
9	9.466	9.427	9.658
10	10.466	10.495	10.500
11	11.466	11.562	11.464
12	12.466	12.630	12.400
13	13.466	13.698	13.750
14	14.466	14.766	14.500
15	15.466	15.834	16.000
16	16.466	16.902	16.500
17	17.466	17.970	17.000
18	18.466	19.038	18.000
19	19.466	20.106	19.000
20	20.466	21.174	20.000

From Table 6, the Form Y equivalent of Form X score of 0 is -0.184 and 0 for linear and equipercentile equating. Also, the Form Y equivalent of Form X score of 4 is 4.087 and 5 for linear and equipercentile equating. Generally, the linear equating ranges from -0.184 to 21.174 for linear equating and 0 to 20 for equipercentile equating. It must, however, be noted that among the equipercentile scores, there is no score of 4 and also, a score of 5 is equivalent to Form X scores of 4 and 5. These are some of the irregularities in equipercentile equating. Therefore, the results on the equipercentile have to be smoothened to do away with some of these anomalies. It must be noted that equipercentile results in Table 6 are unsmoothed. Figure 2 shows the raw-to-raw score conversion.

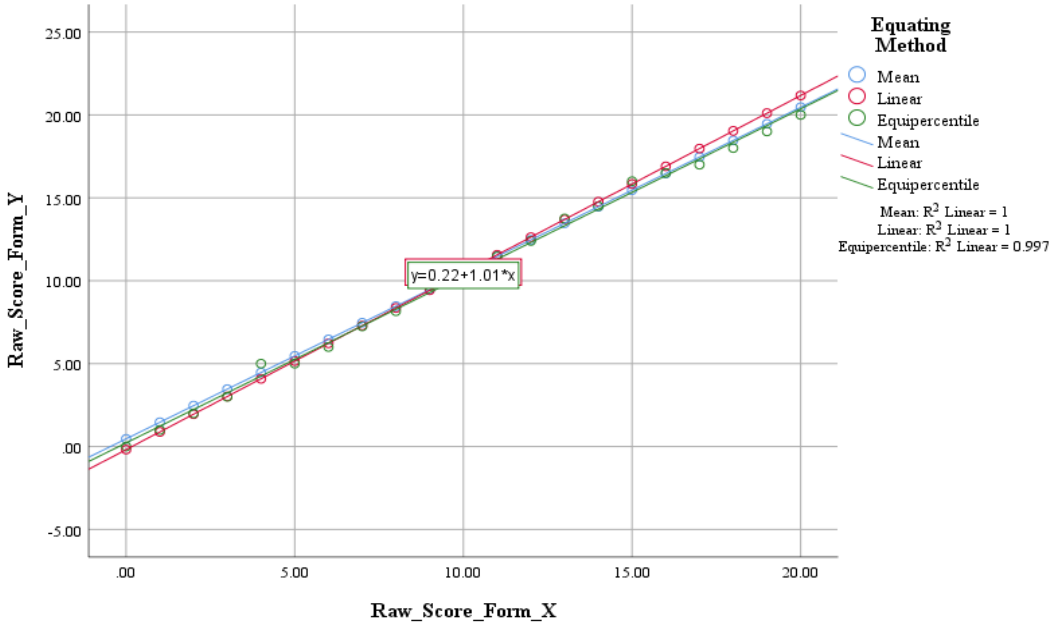

Figure 2: Raw-to-raw Score Conversion (output generated from SPSS)

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Table 7 presents the result on the four moments, namely, mean, standard deviation, skewness, and kurtosis of the Forms X and Y and Form X equated to Form Y using the mean, linear, and equipercentile equating.

Table 7: Moments for Equating Form X and Form Y

Test Form	μ	σ	sk	ku
Form Y	10.0411	2.04422	-.037	.410
Form X	9.5753	1.91426	-.088	.594
Form X equated to Form Y scale for various methods				
Mean	10.4660	6.20484	0	-1.20
Linear	10.4946	6.62617	0	-1.20
Equipercentile	10.2947	6.2607	-0.086	-1.246

As presented in Table 7, the third and fourth moments were the same for mean and linear equating. The first and second moments were never the same for all three equating methods. The moments for the equipercentile equating deviated significantly from those of mean and linear equating.

3.3 Smoothing strategies

Due to the anomalies and irregularities of the equipercentile equating, it equated scores were smoothed. The smoothing was done using presmoothing and postsmoothing methods, respectively. These were done by following the guidelines specified by Kolen and Brennan (2014), pp. 94-95. In terms of the presmoothing, the polynomial log-linear and strong true-score methods were used. The polynomial log-linear method has to do with the use of polynomial degree of C, overall chi-square, difference chi-square, and the Aikake information function (AIC). The presmoothing was done to smooth the score distributions. Table 8 presents the moments and fit statistics for presmoothing.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Table 8: Moments and fit statistics for presmoothing

Form Y:								
Method	**Mean**	**SE**	**sk**	**ku**	**Chi-Squared**	**df**	**Difference**	**AIC**
Sample	10.041100	2.044220	-0.037000	0.410000				
Beta4	10.041096	6.062399	-0.007131	1.792665	144.081	19		
Log-linear								
C=10	10.034280	2.044521	-0.090582	3.493899	5.318	10		27.318
C= 9	10.041102	2.030183	-0.036151	3.301092	5.358	11	0.040	25.358
C= 8	10.041095	2.030174	-0.036182	3.301033	5.399	12	0.040	23.399
C= 7	10.041096	2.030172	-0.036179	3.301031	5.574	13	0.176	21.574
C= 6	10.041103	2.030185	-0.036146	3.301104	5.840	14	0.265	19.840
C= 5	10.041095	2.030173	-0.036182	3.301036	9.172	15	3.332	21.172
C= 4	10.041096	2.030174	-0.036180	3.301056	9.597	16	0.425	19.597
C= 3	10.041099	2.030176	-0.036172	3.003768	9.801	17	0.203	17.801
C= 2	10.041096	2.030174	-0.000003	2.999874	9.817	18	0.016	15.817
C= 1	10.041097	6.055217	-0.008181	1.794638	143.763	19	133.946	147.763
Form X:								
Method	**Mean**	**SE**	**sk**	**ku**	**Chi-Squared**	**df**	**Difference**	**AIC**
Sample	9.575300	1.914260	-0.08800	0.594000				
Beta4	9.575342	5.974346	0.073169	1.821619	148.720	19		
C=10	9.575350	1.901113	-0.086273	3.472999	2.863	10		24.863
C= 9	9.575346	1.901107	-0.086293	3.472964	6.483	11	3.620	26.483
C= 8	9.575344	1.901103	-0.086303	3.472945	6.497	12	0.014	24.497
C= 7	9.575343	1.901102	-0.086305	3.472942	7.557	13	1.060	23.557
C= 6	9.575348	1.901113	-0.086275	3.473029	7.573	14	0.017	21.573
C= 5	9.575343	1.901105	-0.086295	3.473006	8.127	15	0.554	20.127
C= 4	9.575352	1.901129	-0.086220	3.473416	8.161	16	0.034	18.161
C= 3	9.575343	1.901102	-0.086308	3.022237	8.555	17	0.394	16.555
C= 2	9.575345	1.901122	0.000006	2.999963	8.646	18	0.090	14.646
C= 1	9.575342	6.046322	0.084601	1.804394	151.826	19	143.181	155.826

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

From Table 8, for Form Y, only the first moment for the Beta4 was approximately the same as the sample. Also, the χ^2 for the Beta4 (144.081) is greater than its degrees of freedom of 19, suggesting significant chi-square. This implies Beta4 does not fit. From among the log-linear C values, C=3 appears to be the smallest C with approximately the same as the first three moments of the sample. All the others largely differed. The χ^2 value of C=2 (9.817) is less than the table value of 28.869 at degrees of freedom of 18 at .05 level of significance, suggesting non-significant chi-square, so C=3 was selected as the criterion. In addition, the difference in chi-square, thus, $\chi^2_C - \chi^2_{C+1}$statistics was checked. At 1 degree of freedom, χ^2 value of C=2 (3.841) is greater than the difference of 0.016 suggesting a reasonable fit.

For Form X, only the first moment for the Beta4 was approximately the same as the sample. Also, the χ^2 for the Beta4 (148.72) is greater than its degrees of freedom of 19, suggesting significant chi-square. This implies Beta4 does not fit. Relatively, among the Cs, C=2 appears to be the smallest C with approximately the same as the first three moments of the sample. All the others largely differed. The χ^2 value of C=2 (8.646) is less than table value of 28.869 at degrees of freedom of 17 at .05 level of significance, suggesting non-significant chi-square, so C=2 was selected as the criterion. In addition, the difference in chi-square, thus, $\chi^2_C - \chi^2_{C+1}$statistics was checked. At 1 degree of freedom, χ^2 value of C=3 (3.841) is greater than the difference of 0.090 suggesting a reasonable fit.

Above all the aforementioned criteria, the AIC was examined using: $\chi^2_C - 2(C+1)$. Among the AIC values, C=2 had the least value on Form Y (15.817) and Form X (14.646). Based on this information, C=2 was considered as most fitting among all the log-linear C values. Table 9 presents the raw-to-raw score conversions for presmoothing.

Table 9: Raw-to-raw Score Conversion Presmoothing

		Form Y equivalent using equating method		
Form X Score	Standard error	Unsmoothed	Beta4	Log-linear C=2
0	1.9931	0.000	0.04632	-0.16057
1	1.9931	1.000	1.13629	0.88124
2	1.9931	2.000	2.22082	1.95969
3	1.9931	3.000	3.29966	3.04312
4	0.8621	5.000	4.37250	4.12568
5	1.4045	5.500	5.43903	5.20410
6	0.6732	6.000	6.49887	6.27601
7	0.8627	7.250	7.55158	7.33988
8	0.5047	8.167	8.59672	8.39504
9	0.2954	9.658	9.63374	9.44151

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

10	0.3129	10.500	10.66201	10.47975
11	0.3576	11.464	11.68079	11.51484
12	0.3429	12.400	12.68916	12.56212
13	1.0552	13.750	13.68604	13.61980
14	0.7022	14.500	14.67006	14.68781
15	0.8621	16.000	15.63943	15.76511
16	0.0001	16.500	16.59178	16.84959
17	0.0002	17.000	17.52373	17.93766
18	0.0001	18.000	18.42980	19.02025
19	0.0001	19.000	19.30043	20.03375
20	0.0001	20.000	20.11216	20.47348

Table 10 presents the moments for presmoothing

Table 10: Raw-to-raw Score Moments for Presmoothing

Test Form	μ	σ	sk	ku
Form Y	10.0411	2.04422	-.037	.410
Form X	9.5753	1.91426	-.088	.594
Form X equated to Form Y scale for various methods				
Unsmoothed	10.0460	1.9977	-0.0556	3.4213
Beta4	10.0412	6.0623	-0.0062	1.7931
Log-linear C=2	10.0346	1.9903	-0.0069	3.0529

As shown in Table 10, after presmoothing, only the first moments were approximately the same for the two smoothed methods (Beta4 and Log-linear C=2). Having obtained the equipercentile using the presmoothing, the equipercentile equivalents were smoothed using the Cubic Spline (S) using the following values: 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75, and 1.00. The raw score postsmoothing moments for the various Ss were compared. Details are shown in Table 11.

Table 11: Raw score moments for Postsmoothing

Test Form	**Mean**	**SE**	**sk**	**ku**
Form Y:	10.0411	2.0302 -	0.0362	3.3010
Form X:	9.5753	1.9011 -	0.0863	3.4729
Form X Equated to Form Y Scale				
Unsmth:	10.0460	1.9977	-0.0556	3.4213
S=0.01:	10.0434*	2.0127*	-0.0694	3.2981*
S=0.05:	10.0435	2.0020	-0.0639	3.3115

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

S=0.10:	10.0481	1.9876	-0.0461*	3.3867
S=0.20:	10.0581	1.9777	-0.0844	3.4676
S=0.30:	10.0581	1.9777	-0.0844	3.4676
S=0.40:	10.0581	1.9777	-0.0844	3.4676
S=0.50:	10.0581	1.9777	-0.0844	3.4676
S=0.75:	10.0581	1.9777	-0.0844	3.4676
S=1.00:	10.0581	1.9777	-0.0844	3.4676
Linear:	10.0411	2.0302	-0.0863	3.4729

From Table 11, among the Cubic Splines, S = 0.01 was the only one that had three of the moments equal to those of Form Y. From this, it can be said that S = 0.01 produced the most adequate equating. Finally, Table 12 and Figure 3 present the equated equivalents of the three equating methods.

Table 12: Raw-to-raw Score Conversion after Postsmoothing (Unrounded)

	Form Y equivalent using equating method		
Form X Score	Mean	Linear	Equipercentile (S=0.01)
0	.466	-0.184	0.105
1	1.466	0.884	1.314
2	2.466	1.951	2.524
3	3.466	3.019	3.733
4	4.466	4.087	4.942
5	5.466	5.155	5.447
6	6.466	6.223	6.075
7	7.466	7.291	7.127
8	8.466	8.359	8.224
9	9.466	9.427	9.581
10	10.466	10.495	10.548
11	11.466	11.562	11.494
12	12.466	12.630	12.498
13	13.466	13.698	13.651
14	14.466	14.766	14.666
15	15.466	15.834	15.956
16	16.466	16.902	16.790
17	17.466	17.970	17.614
18	18.466	19.038	18.439
19	19.466	20.106	19.263
20	20.466	21.174	20.088

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

From Table 12 the Form X score equivalents of 0 – 20 are .466 – 20.466, -0.184 – 21.174, and .105 – 20.088 for mean, linear, and equipercentile equating, respectively.

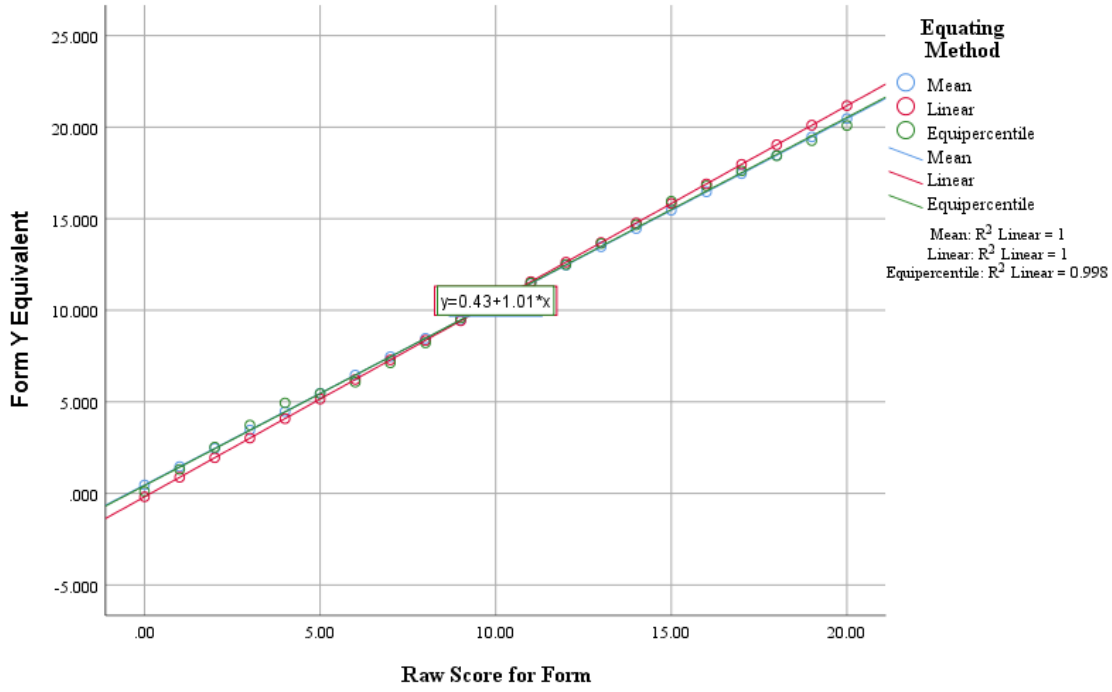

Figure 3: Raw-to-raw Score Conversion (output generated from SPSS)

Table 13 presents the raw-to-raw score conversion (rounded) after postsmoothing. From Table 13, the Form X score equivalents of 2 are 2, 2, and 3, for mean, linear, and equipercentile methods. Also, the Form X score equivalents of 4 and 5 are the same for equipercentile equating. Similarly, euipercentile of 11 was the same for Form X scores of 10 and 11. It was further noted that the Form X score equivalent of 20 is 21 using linear equating. This is greater than the actual scores on Form X.

Table 13: Raw-to-raw Score Conversion after Postsmoothing (Rounded)

	Form Y equivalent using equating method		
Form X Score	Mean	Linear	Equipercentile (S=0.01)
0	0	0	0
1	1	1	1
2	2	2	3
3	3	3	4
4	4	4	5
5	5	5	5
6	6	6	6
7	7	7	7

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

8	8	8	8
9	9	9	10
10	10	10	11
11	11	12	11
12	12	13	12
13	13	14	14
14	14	15	15
15	15	16	16
16	16	17	17
17	17	18	18
18	18	19	18
19	19	20	19
20	20	21	20

Table 14 presents the moments for Form X equated to Form Y for the various methods after postsmoothing of the equipercentile.

Table 14: Moments for Form X equated to Form Y after Postsmoothing

Test Form	μ	σ	sk	ku
Form Y	10.0411	2.04422	-.0362	3.3010
Form X	9.5753	1.91426	-.0863	3.4729
Form X equated to Form Y scale for various methods				
Mean	10.4660	6.20484	0	-1.20
Linear	10.4946	6.62617	0	-1.20
Equipercentile (S=0.01)	10.0434	2.0127	-0.0694	3.2981

From Table 14, the moments for the equipercentile, namely, mean, standard deviation, skewness, and kurtosis were approximately the same as that of Form Y.

4.0 Practical Implications

The most direct implication of this study's finding is its contribution to fair and equitable student evaluation. By demonstrating that different test forms, although constructed to be content-equivalent, can yield varying score distributions, the findings emphasize the need for score adjustments through equating. Therefore, institutions that administer multiple versions of the same exam, particularly in large classes or across campuses, should adopt equating methods to ensure fairness. In line with this understanding, test developers and lecturers can apply the study's

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

equating models (especially equipercentile) to adjust for variations that may arise due to form difficulty differences.

The findings also provide empirical support for the use of equipercentile equating over mean or linear methods in contexts with non-normal score distributions or unequal variances. This finding calls on educational bodies such as exam boards, universities, and curriculum authorities can use this evidence to guide policy formulation on test score equating. Hence, institutions could incorporate equipercentile equating into their standardized testing protocols, especially when stakes are high (e.g., entrance exams, promotions).

At the micro-level, university instructors can use this study as a framework to calibrate assessments administered at different times, such as supplementary exams or deferred assessments. In the same vein, lecturers administering alternative test versions can statistically align scores post-administration, ensuring consistency in grading. Faculty members can use tools such as RAGE-RGEQUATE to process and transform raw scores and include this as a routine part of test analysis.

The study serves as a practical teaching tool for postgraduate courses in educational measurement, statistics, and psychometrics. It illustrates the real-world application of classical equating models. The methodology and results can be used in training programs for lecturers, test developers, and graduate students. Courses on measurement theory, test construction, or item response modeling can adopt this study as a case study for classroom instruction or lab assignments.

The study findings add to the limited body of literature on psychometric practices in West Africa and can serve as a model for localized research, give the unique assessment context in Ghana. The findings, therefore, encourage indigenous psychometric research and local validation of international methodologies. Further, regional exam bodies (e.g., WAEC, NaCCA) can replicate this approach in large-scale assessments for basic, secondary, and tertiary levels.

4.1 Limitations

While this study provides valuable insights into the application of traditional equating methods within a single group random design, several limitations should be acknowledged. First, the study utilized a sample of 146 undergraduate students from a single public university in Ghana. Although adequate for traditional equating procedures, this sample may not capture the full variability in performance across different academic institutions, programs, or regions. Consequently, the generalizability of the findings is limited (Dzakadzie & Quansah, 2023). In addition, only classical equating approaches (i.e., mean, linear, and equipercentile) were employed. Although these methods are widely accepted, the study did not explore modern equating techniques such as Item Response Theory (IRT)-based or kernel equating, which could offer more sophisticated handling of item-level data and latent traits.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

4.2 Conclusion

The present study applied three traditional equating methods (i.e., mean, linear, and equipercentile) within a single group random design to investigate the comparability of two alternate forms of a Basic Statistics test. The results showed that while all three methods offered viable adjustments, the equipercentile equating method consistently demonstrated greater precision and alignment between score distributions. This outcome highlights that equating methods are not universally interchangeable; rather, their appropriateness is contingent upon the distributional characteristics of the test forms and the psychometric properties of the data. Specifically, in cases where score distributions deviate from normality or where form differences extend beyond mean and variance discrepancies, equipercentile equating presents a more robust alternative.

Beyond statistical accuracy, these findings have practical consequences for fair assessment practices, especially in contexts where multiple test forms are used for logistical or security reasons. Institutions, educators, and assessment developers are thus encouraged to adopt equating strategies that are empirically supported and contextually relevant. Future research should consider extending this investigation by incorporating Item Response Theory (IRT)-based equating techniques or exploring longitudinal impacts of equating on student progression. Additionally, simulation studies could help clarify the boundary conditions under which each method performs optimally.

DELCARATION

Data Availability: The data supporting the findings of this study were obtained through test administration and student responses to two alternate forms of a Basic Statistics examination. Access to the anonymized dataset can be made available upon reasonable request to the author, subject to institutional data-sharing policies and appropriate ethical approval.

Declaration of Conflicts of Interest: The author declares that there is no conflict of interest associated with this study.

Ethics Approval and Consent to Participate: Ethical clearance for the study was obtained from relevant academic gatekeepers. Informed consent was sought and obtained from all student participants prior to the administration of the test instruments, in compliance with ethical standards for educational research.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements: The author acknowledges the cooperation of the faculty members and students who participated in the study. Sincere appreciation is also extended to the university administration for granting permission to conduct the research.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

REFERENCES

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Diego, A. (2017). Friends with benefits: causes and effects of learners' cheating practices during examination. *IAFOR Journal of Education*, 5(2), 121–138. https://files.eric.ed.gov/fulltext/EJ1156266.pdf

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Dzakadzie, Y., & Quansah, F. (2023). Modelling unit non-response and validity of online teaching evaluation in higher education using generalizability theory approach. Frontiers in Psychology, 14, 1202896. https://doi.org/10.3389/fpsyg.2023.1202896

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications.* New York, NY: Springer.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29–37.

Kolen, M. J. (2005). RAGE-RGEQUATE manual (Console version)

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Method and practice* (2nd ed.). New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Method and practice* (3rd ed.). New York, NY: Springer-Verlag.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23–39.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Nitko, J. A. (2001). *Educational assessment of students*. New Jersey: Prentice Hall.

Nugba, R. M., & Quansah, F. (2021). Standardized Achievement Testing, Aptitude Testing, and Attitude Testing: How Similar or Different are these Concepts in Educational Assessment? *Asian Journal of Education and Social Studies*, 42-54. **DOI:** 10.9734/ajess/2021/v15i330383

Öztürk, N., & Anıl, D. (2012). A study on equating the scores of the academic staff and postgraduate education entrance exam. *Eğitim ve Bilim, 37*(165), 181–193.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics,* 8, 137-156.

Quansah, F., & Cobbinah, A. (2021). Equivalence of parallel tests in a Basic Statistics course in higher education using classical measurement theory. *Canadian Journal of Educational and Social Studies, 1*(2), 13-28.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

APPENDIX A
Item Specification

Item 1

Topic - Discrete/categorical and continuous data

Objective: To test students' knowledge on continuous variable

Description: Test items may include description of continuous variable assuming any value between two points on a number line.

Sample item:

1. A variable that assumes any value on a number line between two points is a …….. variable.
   a. continuous*
   b. discrete
   c. qualitative
   d. quantitative

Item 2

Topic – Parameter and Statistic

Objective: To test students' understanding on parameters

Description: Test items may include a scenario/situation that depict students' understanding of the term 'parameter'.

Sample item:

2. Mr. Osei, the Principal of Dafur Nursing Training College, describes the characteristics of newly admitted first year students in the school. He reports the mean age of all first year students as 21 years. Which one of the following concepts describes Mr. Osei's description?
   a. Parameter*
   b. Population
   c. Sample
   d. Statistic

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 3

Topic – Parameter and Statistic

Objective: To test students' understanding on statistic

Description: Test items may include a scenario/situation that depict students' understanding of the term 'statistic.

Sample item:

3. A Senior Nursing Officer is interested in describing the mass of infants attending weighing in a particular day. The modal mass of the first 30 infants who attended the weighing was reported as 7.2kg. Which one of the following concepts describes the Officer's description?
   a. Parameter
   b. Population
   c. Sample
   d. Statistic*

Item 4

Topic – Discrete/categorical and continuous data

Objective: To test students' understanding on qualitative and quantitative data

Description: Test items may include a scenario/situation that depict students' understanding qualitative/quantitative data which are also continuous or discrete in nature.

Sample item:

4. A tutor reports to the Principal of a college the number of days students have been in school. This information can be termed as ……..
   I. quantitative data **II. continuous data** **III. discrete data**
   a. I and II.
   b. I and III.*
   c. I only.
   d. II and III.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 5

Topic – Descriptive/qualitative and quantitative data

Objective: To test students' understanding on qualitative and quantitative data

Description: Test items may include a statement that depicts students' comprehension of qualitative/quantitative data which are also numerical.

Sample item:

5. Which one of the following data is only categorized but **NOT** numerical?
   a. Continuous
   b. Discrete
   c. Qualitative*
   d. Quantitative

Item 6

Topic – Descriptive/qualitative and quantitative data

Objective: To test students' understanding on qualitative and quantitative data

Description: Test items may include a statement that depicts students' comprehension of qualitative/quantitative data which are also continuous or discrete.

Sample item:

6. A tutor measures the temperature of a student in degree Celsius. This information can be termed as ……
   I. quantitative data **II. continuous data** **III. discrete data**
   a. I only.
   b. II only.
   c. I and II.*
   d. I and III.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Items 7 & 8

Topic – Frequency distribution and cumulative frequency tables

Objective: To test students' knowledge on relative frequency and mode for grouped data.

Description: Test items may include a set of grouped data from which students identify modal class, class mark, relative frequencies, and cumulative relative frequencies.

Sample items:

A teacher examined the performance of students in a quiz. The scores of the students are presented in the table below. Study the frequency distribution table and answer questions 7 – 8.

Classes	Frequency
46 – 51	13
40 – 45	10
34 – 39	9
28 – 33	8
22 – 27	15
16 – 21	4
Total	**59**

7. What is the relative frequency for the class 28 – 33 **(correct to 2 decimal places)**?
   a. 0.61
   b. 0.14*
   c. 0.22
   d. 0.54
8. What is the modal class?
   a. 16 – 21
   b. 22 – 27*
   c. 34 – 39
   d. 46 – 51

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 9

Topic – Graphical organization of data

Objective: To test students' knowledge on pictorial representation of data.

Description: Test items may scenarios/situations that depict the appropriate use and properties of pictorial representations.

Sample item:

9. Which one of the following is **NOT** a property of pictorial representations of data?
    a. They should be adequately labelled and titled
    b. They should be simple and clear in their meaning
    c. They should carry all the necessary information
    d. They should include only scores with high frequencies*

Item 10

Topic – Graphical organization of data

Objective: To test students' knowledge on pictorial representation of data.

Description: Test items may describe scenarios/situations that depict the appropriate use of pictorial representations, particularly histograms.

Sample item:

10. As a student in Statistics class, you are interested in comparing the examination anxiety levels of pupils from three different schools: School A, School B, and School C. Which of the following is **NOT** appropriate to use?
    a. Bar chart
    b. Cumulative frequency curve
    c. Frequency polygon
    d. Histogram*

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 11

Topic – Graphical organization of data

Objective: To test students' knowledge on pictorial representation of data.

Description: Test items may describe scenarios/situations that depict the appropriate use of pictorial representations, particularly cumulative frequency curve.

Sample item:

11. An ogive is also called ……..
    a. cumulative frequency curve.*
    b. frequency distribution.
    c. frequency polygon.
    d. histogram.

Item 12

Topic – Graphical organization of data

Objective: To test students' knowledge on pictorial representation of data.

Description: Test items may describe scenarios/situations that depict the appropriate use of pictorial representations, particularly frequency polygon.

Sample item:

12. Which of the following graphs join the mid-point with a straight line?
    a. Cumulative frequency curve
    b. Frequency distribution
    c. Frequency polygon*
    d. Histogram

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 13

Topic – Measures of Central Tendency

Objective: To test students' knowledge on the interpretations based on the centre of a data set.

Description: Test items may describe scenarios/situations that describe the centre of an entire data set.

Sample item:

13. Mrs. Mensah is interested in describing the performance of students in Psychology. Which of the following measures is **NOT** appropriate to use as a single score to describe the entire performance?
    a. Mean
    b. Median
    c. Mode
    d. Range*

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Items 14, 15, & 16

Topic – Measures of Central Tendency

Objective: To test students' knowledge and comprehension on estimation of measures central tendency from a given data set.

Description: Test items may describe scenarios/situations that describe the centre of an entire data set. It may require students to compute mean and apply the effect of mean in different scenarios.

Sample items:

The following are scores of students in a Psychology quiz:

13	5	6	9	11	7	6
9	10	8	21	10	8	4

Use the following information to answer questions 14-16.

14. What is the median of the distribution?
    a. 7.5
    b. 8.0
    c. 8.5*
    d. 9.1
15. How would you describe the performance of Mercy who had a score of 10 in the quiz?
    a. Above average*
    b. Average
    c. Below average
    d. More information is needed
16. It was later observed that Akwasi who obtained a score of 4 actually had 15. How would the new score affect the mean?
    a. Both mean would be the same
    b. The new mean would be less than the old mean

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

c. The old mean would be less than the new mean*

d. There would be no change in the mean

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Item 17

Topic – Measures of Dispersion

Objective: To test students' knowledge on the interpretations based on the variability of a data set.

Description: Test items may describe scenarios/situations that describe the variability of an entire data set.

Sample item:

17. The first quartile in the following distribution is

2 5 3 9 10 13 1

a. 1
b. 2*
c. 5
d. 10

Item 18

Topic – Measures of Dispersion

Objective: To test students' knowledge on the interpretations based on the variability of a data set.

Description: Test items may describe scenarios/situations that describe the variability of an entire data set, that is dispersion.

Sample item:

18. The measure of how widely scores are scattered is known as ……

a. central tendency.
b. dispersion.*
c. distribution.
d. mean.

Quansah. (2025), Vol. 6, Iss. 2, Pg. 01-33
https://doi.org/10.5281/zenodo.15778619

Items 19 & 20

Topic – Measures of Central Tendency

Objective: To test students' knowledge and comprehension on estimation of mean and how a change in score can affect the mean of a data set.

Description: Test items may describe scenarios/situations that require students to interpret the mean in terms of average. They may require students to compute mean and apply the effect of mean in different scenarios when a score changes.

Sample items:

The following data are the ages of teachers in a college:

21 24 19 17 29 35 31

Use this information to answer questions 19 – 20.

19. What description can you make about the ages of the teachers?
    a. Majority of the class are above average*
    b. Majority of the class are below average
    c. More information is needed
    d. The students are very old
20. Assuming a teacher with age 20 years was later added to the class, how would it affect the range of the entire data?
    a. More information is needed
    b. The range would decrease
    c. The range would increase
    d. There would be no change*

*Keyed response