

Test Equating Using Linear Equating Method

Umar Safianu Abukari

Saboba D/A Junior High School. P. O. Box 7. Saboba, Northern Region, Ghana

Robert Osman Iddrisu

Department of Educational Foundations

University for Development Studies, P.O. Box TL 1350 Tamale, Ghana

Simon Alhassan Iddrisu

Department of Educational Foundations

University for Development Studies, P.O. Box TL 1350 Tamale, Ghana

ABSTRACT

The study aimed to equate two test forms using the linear equating method and estimate the amount of equating errors associated with the method. The linear equating method was used to adjust for the difficulties of the test Form A and Form B on students' Mathematics test scores. A 20-item mathematics instrument was used for data collection. The bootstrap method was used in the determination of the amount of equating errors that come with this equating method. Data on participants' test scores on these two congeneric parallel test forms were used for this study. All the requirements for equating the two test forms were met with a reliability coefficient of .71. The scores of a new Form B of the test were then equated to its referenced Form A using a linear equating method. Findings from the study revealed that the linear equating method produced scores that were beyond the required possible observed scores, especially at the lower end of the scores. Again, the amount of error associated with the linear equating method seems to decrease consistently from the low ability scores to the high ability scores.

Keywords: Test forms, Congeneric Forms, Equating Designs, Linear equating, Equating errors.

INTRODUCTION

In the fields of education as well psychology, tests and scales are commonly employed to assess the learning abilities of individuals, categorize them into different proficiency levels, make selections, provide guidance, and offer clinical services [1]. Similarly, the utilization of high-stakes assessments occasionally necessitates the administration of several tests and scales at separate instances for security purposes. This prompts test developers to create distinct parallel test forms as a solution to this issue. However, the scores obtained from these many tests create a demand for psychometricians to accurately interpret the results obtained from multiple test administrations. The aforementioned statement has led to a surge in attention towards test-equating studies as well as test development in the field of testing [7].

Test equating involves adjusting for the difficulty levels of two parallel or alternate forms of a test, enabling them to be utilized in place of each other. It converts the scores obtained from one of the test forms to the scale of the new parallel test form [1]. Equating is the process of establishing equivalent ratings between two separate assessment tools. An equating process is conducted when there are different versions of a test and we want to set side by side the output of test takers who have been administered separate versions. In an anchor nonequivalent group design, when two distinct groups undergo the same test, it becomes necessary to make adjustments for the disparities in their performance. [2] argue that equating is performed when there are different versions of tests available, and the scores on these versions need to be compared. The subject of equivalence of students' observed test scores across several tests assessing the same skill is a matter of worry for test-takers, test developers, and measurement specialists, among others. How can the scores of two examinees be compared if they take different tests? This question is crucial when we need to make selection, pass/fail or certification decisions. This is because it should be inconsequential which test is utilized to make the decisions. To compare the test scores achieved on tests 'B' and 'A', it is necessary to conduct a procedure of equating the scores between the two tests. This technique establishes a correlation between the test scores on test B and test A, and converts the score on the new test form to the metric of the referenced test form. Therefore, an individual who achieves a score of b on test B will have a converted score on test A that can be directly compared to the score of an individual who took test A. This study involved the equating of two exam forms, namely Form B (New Form) and Form A (Reference Form). The parallel tests that were equated were of the congeneric type. This is because the means, variances, covariances, and validity coefficients were all estimated and it was seen that these statistics were not all equal.

Objectives of the Study

This study sought to:

1. use linear equating methods to equate two parallel test forms, a new form and a referenced form.
2. estimate the magnitude of equating errors associated with the linear equating approach.

Significance of the Study

The significance of this study lies in the fact that the findings will afford test users the knowledge and skills in converting test scores administered on different dates into forms that can be used interchangeably for decision-making purposes. The test a candidate took would no longer matter when a selection choice was to be made on them after these equating processes were employed to adjust tests of different difficulty levels.

LITERATURE REVIEW

The review covers the sections expanded on to bring clarity to the concept of equating and its associated processes. Aspects explained included conditions or requirements for equating, equating designs and procedures among others.

Conditions/Requirements for Equating

According to [4], equating relationships, especially within the classical theory framework must meet certain conditions. Among these conditions are:

1. The tests to be equated must measure the same traits. According to [3], this condition can be checked by allowing competent judges to inspect the content and wording of the

two sets of test questions through item specification. To ensure that this condition was met in this study, the two test forms were constructed from the same table of specifications. Each item specification of the parallel test was also similar. In addition, the wording, item specification and description were evaluated by experts before the administration of the parallel test.

2. The tests to be equated must not be of varying difficulty since such tests will not be equally reliable at different ability levels. To satisfy this condition the mean item difficulties of the two forms of test were estimated and concluded that both were seen to be moderate as form "A" had a mean difficulty of 0.458 while form "B" had a mean difficulty of 0.476. This indicates the similarities of item difficulties of the two forms of test.
3. Test scores on tests "B" and "A" cannot be equated unless the tests are parallel. The test scores for this study were initially examined to determine the degree of parallelism. In doing this, the content, difficulty, means, standard deviations, variances, and covariance of each of the forms were estimated and compared. These comparisons revealed that the two forms were seen to be a congeneric type of parallel test and hence equating was possible.
4. Only highly reliable tests can be equated. Equating of test forms is carried out when the reliability of those test forms is high. The researchers estimated the reliability of the test forms used in this study which were found to have a reliability coefficient of 0.71. These two forms had considerably higher reliability and were therefore suitable for equating.
5. Also, it is expected that the equating relationship be symmetrical. Either converting Form "A" scores to the scale of Form "B" or vice versa should provide the same equated scores; that is, the equating process should be uniform in both directions. According to published works, this symmetry assumption can be easily met by simply defining the term. The reason behind this is that the symmetric requirement is satisfied automatically by the conventional definitions of the linear equating functions [3]. Since linear regression calculations are based on the symmetry assumption, they cannot be used as an equalizing method [6].

Types of Equating Designs

[6] stated that there is a diverse range of designs available for data collection in the context of equating. Three typically utilized equating designs were found. However, it was recommended that the examinees involved in an equating study should be generally representative of the examinees who will take the test under ordinary test administration conditions. There are three types of designs: random group design, single group with counterbalancing design, and common-item non-equivalent group design.

For this study, the random group design was employed to gather data for the purpose of equating. This is due to the fact that the design offers several benefits in comparison to alternative designs, including ease of administration. For example, the single group counterbalancing design is typically preferred because each examinee only writes one Form of the test, hence reducing testing time. The spiralling process is a method employed in the random group design to assign forms to examinees in a random manner [6]. During this technique, the exam booklets were bundled by alternating between Form A and Form B. When distributing the booklets to the examinees, the first examinee received Form "A", the second examinee received Form "B", the third examinee received Form "A", and so on. An advantage of

utilizing the spiralling process is that it results in the formation of comparable, randomly equivalent groups that get Form A and Form B [6]. Furthermore, when considering the designs of the random groups, any disparities in performance detected between the groups on the two forms are regarded as clear evidence of the discrepancy in difficulty between the forms. Furthermore, the random group design ensures that each examinee is only required to take one version of the test, hence reducing the overall testing duration compared to a design where examinees are required to take many versions. An additional benefit is that multiple new forms can be equated simultaneously by incorporating them into the spiralling process. Group disparities are additionally addressed through the random allocation of forms to test takers [5]. The equivalent-groups design possesses significant limitations. It necessitates the simultaneous availability and administration of all the forms, which could provide challenges in some circumstances. Furthermore, a substantial number of individuals is necessary to obtain precise equating outcomes. The equivalent-groups design may necessitate a significantly larger number of test-takers, ranging from five to fifteen times more, in order to attain a similar level of accuracy as the counterbalanced approach. Nevertheless, we had a total of 30 individuals who took each version of the test. In using the spiralling procedure in the random assignment, the following practical issues were considered. First, the examinees were allowed to be seated randomly and not in any systematic way. In addition, two classrooms were used for the testing, and each of the classes had 30 students; this means that each class had an even number of students. Due to this, the last person in the first class had Form "B", so the next class had to begin with Form "A". The spirit for this was to prevent situations where examinees will take more of a particular form than the other.

METHODS

Linear Equating Method

Test scores on parallel test forms can be equated using identity, mean, linear and equipercentile methods of equating [6]. However, this study focused on a linear Equation method. According to [8], the linear equating approach considers a score on the new form (Form B) and a score on the reference form (Form A) to be similar within a group of test-takers if they are at the same distance, measured in standard deviations, from the group's mean. In linear equating, scores that have the same standard deviation and are equidistant from their means are considered similar. [6] explained that whereas mean equating assumes a consistent difference between two forms and utilises it for equating, linear equating allows for the variation in difficulty differences between the two test forms across the score scale. Linear equating permits Form B to demonstrate higher levels of difficulty for examinees with lower achievement, while showing lower levels of difficulty for examinees with higher achievement [6]. Linear equating permits for the variation of both the measurement units and the averages of the two test types. The linear conversion is determined by equating the standardised deviation scores (z-scores) on both versions of the test. If the standard deviations for the two forms are similar, it is clear that linear equating is equivalent to mean equating. If the standard deviations of the two forms are comparable, then both the mean and linear equating approaches get identical results.

Linear equating assumes that the sole discrepancy between the two distributions of test scores is in their means and standard deviations [6]. In order to build a correlation between scores on the new form and scores on the reference form among a group of test-takers, it is essential to convert each score on the new form to a score on the reference form that aligns with the same number of standard deviations above or below the group's mean. Linear equating assumes a

linear relationship between the scores on test B and test A. According to Livingston (2004), the linear equating formula is given as:

$$Y = \left[\frac{SD(Y)}{SD(X)} \right] X + \left[\text{Mean}(Y) - \left[\frac{SD(Y)}{SD(X)} \right] \text{Mean}(X) \right] = \text{adjusted } x \quad (1)$$

However, with the introduction of statistical software, test equating has become easier as compared to the manual calculations that were used in the past. As a result, in this equating study, the conversion of scores on the new form to the scale scores of the referenced form under the linear equating method was done by using RAGE-RGEQUATE computer programme.

RESULTS AND DISCUSSION

The analysis processes in this study involve a number of processes. The RAGE-RGEQUATE was used to compute smoothed frequencies which were then inputted into the SPSS to compute cumulative frequencies, adjusted 0.5, percentile ranks, Z-scores and T-scores. These T-scores were entered back into the RAGE-RGEQUATE again to compute the linear equating relationship. Therefore, in a complementarity form, the SPSS and RAGE-RGEQUATE were used to conduct a linear equating relationship. According to [6], the conversions for mean, linear, and equipercentile equating results are best estimated with the use of RAGE-RGEQUATE computer software. In addition, the Equating Error was used to estimate the amount of errors associated with this equating method in the study. The results obtained from the RAGE-RGEQUATE computer program are shown in Table 1.

Table 1: Raw-to-Raw Score Conversion Table

New form Score	Referenced Form similar score Using Linear Equating
0	1.303
1	2.226
2	3.149
3	4.071
4	4.984
5	5.917
6	6.840
7	7.763
8	8.685
9	9.608
10	10.531
11	11.454
12	12.376
13	13.299
14	14.222
15	15.145
16	16.067
17	16.990
18	17.913
19	18.836
20	19.759

Source: field data (2024)

It can be observed from Table 1 that, the linear equating method produced scores that were outside the range of possible raw scores, especially at the lower end of the scores. For instance, a score of 0 under Form B equated to a score of 1.303 under Form A. Then also, when Form B score is 1, its Form A equivalent is 2.226. Again, a score of 7 on Form B equated to a score of 7.763 on the reference form (Form A). These are the limitations associated with the linear equating. [6] highlight that the linear equating method may have a restriction where the equated scores can occasionally fall outside the range of feasible observed scores. [8] states that with linear equating, a score that is extremely high or extremely low on the new form can result in a score that falls beyond the range of possible scores on the reference form, namely Form A. This means that under linear equating, it is possible to obtain rescaled scores outside the permissible range. [4] advises that depending on the final reporting scale, these are often truncated to the lowest and highest permissible scores (arbitrarily decided). According to [6], there exist multiple approaches to address this issue. An option is to enable the top and bottom to "float." For instance, it is possible for the highest equated score to surpass the greatest raw score. A viable option is to limit the conversion by excluding the greatest and lowest values. In this study, we permitted the upper and lower ends to remain unrestricted because no participant seemed to have attained the extreme raw scores on Form B that could correspond to the unattainable numbers on Form A.

Estimating the Standard Errors of Equating In equating test forms to make them interchangeably sometimes introduces errors, therefore in order to estimate the amount of error, the bootstrap method was used to estimate the amount of equating errors related to the linear equating method used. In the bootstrap, samples are taken from the empirical distribution [6]. Thus, the bootstrap procedure uses empirical or actual distribution. This is appropriate for this study because the study used actual or empirical data collected using the random group design. For the purpose of assessing the data and determining the equating errors related to the linear equating method employed in this study, the Equating Error: A Programme for Computing Equating Error was utilised. The Bootstrap (Windows Console Version) was employed to quantify the extent of equating errors associated with the utilisation of the linear equating method. The bootstrap standard error of equating for the raw-to-raw score conversion using linear equating methods is displayed in Table 2. These errors were estimated at each score point. as indicated in Table 2.

Table 2: Bootstrap Standard Error of Equating for the Raw-To-Raw Score Conversion using Linear Equating Methods

Form B score	Linear equating error
0	6.9211
1	6.2339
2	5.6064
3	5.0677
4	4.6115
5	4.2070
6	3.8313
7	3.4791
8	3.1486
9	2.8459
10	2.5803

11	2.3646
12	2.2137
13	2.1425
14	2.1614
15	2.2743
16	2.4839
17	2.8172
18	3.3938
19	4.3023
20	5.3120

Source: field data (2024)

It can be observed from Table 2 that the amount of linear equating errors decreases consistently from the lower scores to the higher scores. This can be attributed to many reasons including the fact that the low-achievers introduce a lot of errors into the scores probably through guessing. Similarly, critical observation and analysis of the amount of equating errors associated with this equating method suggest that the total equating errors are estimated at 45.5522 with an average error of 2.1615. As shown table 2, majority of the errors from the various scores are above the average error with just few below average. This implies that majority of the errors resulted from almost all the score points in the equating. Generally, the average equating error is minimal (close to zero than the highest equating error of 6.9211), however it is recommended that error margins at individual score levels, especially at the lower end of the observed scores should be reduced. This can be done by removing or reducing the guessing parameter in the test.

CONCLUSION

The researchers concluded that test Form “B” scores were successfully equated to the scale scores of test Form “A” using the linear equating method. This means that this equating study has successfully adjusted for the differences in difficulty between the new test form and the referenced test form and for this reason, their scores can be used interchangeably. Therefore, after this successful equating in this study, for example, an examinee who got an equated test score of 13 on Form “B” can on average be regarded, as at the equal attainment level as the examinee with an observed score of 13.299 on Form “A” using linear equating method (See Table 1). The average equating error estimated in this study was seen to be 2.1615. The error margin was high among low-ability students and as a result, the researchers recommended that the guessing parameter in the test should be removed or reduced to further reduce the error margins at the lower ends of the observed scores.

References

- [1] Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice*, 16(2) 647-668.
- [2] Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline pre-smoothing method and the direct pre smoothing method. *Journal of Educational Measurement*, 46(2), 135-158.
- [3] Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.

-
- [4] Hambleton, R. K., Swaminathan, H., & Rogers (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
 - [5] Heh, V. K. (2007). *Equating accuracy using small samples in the random groups design* (Doctoral dissertation). Patton College of Education at Ohio University, Athens, OH.
 - [6] Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling and linking*. New York, NY: Springer.
 - [7] Lamprianou, I. (2007). *An investigation into the test equating methods used during 2006, and the potential for strengthening their validity and reliability*
 - [8] Livingston, S. A. (2004). *Equating test scores*. Princeton, NJ: ETS.