A comparative analysis of pre-equating and post-equating in a large-scale assessment, high stakes examination

Dibu Ojerinde
Dibu65ojerinde@yahoo.com
Joint Admissions and
Matriculation Board (JAMB),
Abuja Nigeria

Omokunmi Popoola
kunmipopoola@yahoo.com
Joint Admissions and
Matriculation Board (JAMB),
Abuja Nigeria

Patrick Onyeneho
patrickonyeneho@yahoo.co.uk
Joint Admissions and
Matriculation Board (JAMB),
Abuja Nigeria

Aminat Egberongbe
amiegberongbe@yahoo.com
Joint Admissions and
Matriculation Board (JAMB),
Abuja Nigeria

DOI: http://dx.doi.org/10.18820/2519593X/pie.v34i4.6
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
2016 34(4): 79-98

Abstract

Statistical procedure used in adjusting test score difficulties on test forms is known as "equating". Equating makes it possible for various test forms to be used interchangeably. In terms of where the equating method fits in the assessment cycle, there are pre-equating and post-equating methods. The major benefits of pre-equating, when applied, are that it facilitates the operational processes of examination bodies in terms of rapid score reporting, quality control and flexibility in the assessment process. The purpose of this study is to ascertain if pre- and post-equating results are comparable. Data for this study, which adopted an equivalent group design method, was taken from the 2012 Unified Tertiary Matriculation Examination (UTME) pre-test and 2013 UTME post-test in Use of English (UOE) subject. A pre-equating model using the 3-parameter (3PL) Item Response Theory (IRT) model was used. IRT software was used for the item calibration. Pre- and post-equating were carried out using 100-items per test form in an UOE test. The results indicate that the raw-score and ability estimates between the pre-equated model and the post-equated model were comparable.

Keywords: *pre-test, post-test, equating, ability estimates, equivalent group design*

1. Introduction

Developments in the field of education, psychology and statistics communities have immensely assisted researchers in assessment through its contributions towards the rapidly growing statistical and psychometric methodologies used in test equating. In large-scale examinations such as, the Unified Tertiary Matriculation Examination (UTME) where candidates' scores are used for high-stakes decisions, testing programmes require new versions of tests to be continually produced. The essence and expectation is that tests produced should be equivalent in test score difficulty as well as in functionality over time. The UTME is a computer-based test (CBT) conducted by the Joint Admissions and Matriculation Board (JAMB) for the purposes of selecting qualified candidates for admissions into Nigerian tertiary institutions. The examination, which comprised of 23 subjects including the UOE, is conducted at different times within a specified period of 14 days for over 1.5 million candidates. Therefore, the UTME is compulsory for any candidate seeking admissions into any tertiary institution in Nigeria. It is therefore a high-stakes test since results obtained from this examination is used in making important decisions about the candidates. Since this examination is conducted at different times and different days using several test forms in 23 subject areas, equating of the test forms is necessary. Equating is therefore a statistical procedure used in adjusting scores of two or more tests such that the resulting new forms of the test can be comparable. In supporting this assertion, Livingston (2004) defined equating as a statistical procedure that adjusts test scores for difficulty of the items. Equating as a statistical process refers to the derivation of transformations which places scores of different forms of a test onto a scale such that after transformation, the scores on the resulting forms are comparable. This definition can be likened to the meaning of equating by Kolen and Brennan (2004) who are of the opinion that it is a process that is used in adjusting scores on two or more test forms such that the scores can be used interchangeably.

Equating is an important component of any testing programme that produces more than one form for a test. It places scores from different forms onto a single scale. Once scores are placed on a single scale, the scores are interchangeable (Kolen & Brannam, 2004; Holland & Dorans, 2006). This development permits standardisation of scores across test forms such that what is applied to one test form is also applied to the other forms enabling consistency and accuracy across test forms in classification decisions. It is for this reason that equating has become essentially important to testing programmes that use test scores for the measurement of students' growth as well as high-stakes decisions. In the UTME, pre-equating is used in establishing a conversion table prior to the operational testing. Kirkpatrick and Way (2008) affirmed that a series of advantages arise from the use of the pre-equating over the use of post-equating. Top on the list of benefits stated include assessment that is more flexible and a better quality-control check for the tests.

Generally, what equating does is adjust test score difference because of score difficulty. Normally, it is desirable to have the same group of test takers take the new test form as well as the reference form at the same time. The difference in average performance on the two forms indicates the difference in form difficulty. After this, scores on the new test form can then be statistically adjusted to make the average performances on both forms equivalent. Nonetheless, in practice, it is not possible to compel test takers to take two different tests at the same time; rather it is more convenient to have the two different groups of test takers take the two forms of the test at the same time or on two different occasions. However, because these two groups of test takers could have different average abilities, Xuan and Rochelle (2011) are of the opinion that the difference in average performance on the two forms could be an indication of the existence of both group ability differences and form difficulty differences.

Equating may be classified as pre-equating or post-equating depending on the period when the equating practice is being conducted. Pre-equating according to Tong, Wu and Xu (2008) is to conduct equating prior to the operational testing while post-equating involves conducting equating after the operational testing. In their paper, they stated that pre-equating and post-equating are used in K-12 large-scale assessment programmes. In many large-scale, high stakes examinations such as the UTME where immediate reporting of scores are required, pre-equating is often a preferred alternative to post-equating since the equating transformation must be produced in a rather short period of time. Every prospecting UTME candidate is expected to enrol four UTME subjects including the UOE. The subjects are selected based on the faculty and course requirements. Normalised scores are reported based on the four subjects for each candidate. The normalised scores are based on Z-score and T-score transformations of the raw score. No other form of equating is carried out since the equating has been done prior to test administration.

The UTME results is solely used by the Joint Admissions and Matriculation Board (JAMB) and the tertiary institutions in Nigeria as an entrance examination for selecting eligible candidates into the various programmes/courses offered by the institutions. The computer-based testing administered by the JAMB takes place at different times and dates and so, several forms of the same test are required in each session in order to forestall item over-exposure of the items in the item bank. This is a strategy for curbing incidences of examination security breach. Since immediate score reporting is needed, all forms of tests for all the subjects are pre-equated in order to make them equivalent. This is to ensure that no candidate is in any way placed at a disadvantage because of administering any form of the test forms.

When embarking on equating, care must be exercised in order to avoid equating errors. If equating errors exceed some tolerable limits as a result of applying pre-equating, this can likely lead to multidimensionality. The probable cause for pre-equating error is the presence of bias in the item parameter estimates caused by the violation of the assumption of item local independence (Kolen & Brennan, 2004). A guide against committing serious equating errors through ensuring that model assumptions are to a reasonable extent complied with adds value to the final equating results.

2. Statement of problem

In many large-scale high stakes assessment enterprises such as the UTME, stakeholders need assessment evidence as quickly as possible to enable them to make informed decisions relating to admissions or other policy issues. The nature of the UTME assessment makes it pertinent to release candidates' results as quickly as possible in compliance to requests requiring meeting some deadlines in reporting scores. To facilitate this, test items are often calibrated prior to the operational administration with the raw score to scale score conversion tables prepared well ahead of the test administration to ease problems that impede quick reporting. The use of different forms of the same test for assessment often raises the issue of the comparability of test scores across forms. In order to use the scores from different forms of a test interchangeably, they must be put on a common scale. The problem is how to make the several test forms, which consists of different test items drawn from the same content areas of the syllabus, psychometrically equivalent so that whichever form is given to any candidate, s/he will not in any way be disadvantaged.

3. Purpose of study

Measurement equivalence is said to exist when candidates with the same scores on the latent trait have the same expected raw or true score at the item level. Raju, Laffitte and Byrne (2002: 517) inferred that without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully. The purpose of this study therefore, is to compare pre-equating and post-equating scores of candidates in the UTME high stakes examination in order to ascertain if the tests function the same way for students in a field test administration as well as in an operational test administration.

4. Literature review

While some researchers have varied views regarding the efficacy of pre-equating in a high stakes examination, other studies have suggested that pre-equating can achieve satisfactory results. For instance, a study by Livingston (2004) which adopted some sort of method similar to regression, demonstrated that pre-equating was highly accurate in three of the four New Jersey College Basic Skills Placement tests. Studies have also shown that there is a dearth of literature on post-equating. However, Kirkpatrick and Way (2008) were of the opinion that in post-equating, new operational data can be obtained for items selected from the calibrated item pool. They explained that item parameters are estimated for the operational data, and operational items are post-equated using the pool (old) and current (new) item parameters as well as a scale transformation procedure. If new field test items were administered with the operational items, this transformation can be applied to their calibration results as well.

Furthermore, in two of the most recent studies conducted by Domaleski (2006) and Tong *et al.* (2008), they supported the use of pre-equating by having similar pre- and post-equated scoring tables and similar accuracy of classifying students into different performance levels. Apart from different research findings about pre-equating, a literature review indicates that little research has been conducted on whether pre-equating agrees with the post-equating for a test let-based and computer-administered testing programme. What is more, given the controversial view towards the use of pre-equating and the appealing features that pre-equating can offer more research is clearly needed in this area. To this end, this study, which employed empirical data, aims at investigating whether the pre-equating results agree with the equating results based on operational data (post-equating). The study examined the degree to which the IRT pre-equating results agreed with those from IRT post-equating and the degree to which the two equating designs agree with each other.

Since pre-equating establishes a conversion table prior to the operational testing, a series of advantages often arise from the use of the pre-equating over that of post-equating (Kolen & Brennan, 2004) (Kirkpatrick & Way, 2008). These advantages include assessment that is more flexible, a better quality-control check for the tests and its ability to facilitate immediate score reporting of tests right after the test administration.

5. Equating designs and equating method

This research is based on the equivalent group equating design. The UTME test is a high stakes standardised test that is made up of 100 items. Twenty-three other subjects are also tested but candidates are only allowed to choose four subjects according to faculty and departmental requirements. The Use of English (UOE) subject is compulsory for all candidates and all the tests are administered via a computer-based testing mode using the linear-on-the-fly-testing (LOFT) method. In the UOE test, test forms C1, C2, C3 and C4 were created with each taking into cognisance the sub-sections of the syllabus and weights as stated in the UTME syllabus. In so doing, more than one parallel forms were created. Each of these trial-tested items was used in 2012 in creating tests administered in a subsequent operational examination.

The UTME test therefore contains many versions of the same test (test forms) created from the same rational content domain as stored in JAMB item banks. The test forms were built and made equivalent in terms of content and psychometric properties. For example, test form C1 in UOE from the trial-test was taken as a reference form while forms C2, C3 and C4, etc., were made equivalent and taken as the focal groups for the pre-equating. Data in these test forms were organised such that they have item distributions of mean = 0 in terms of item difficulties *b* and discrimination parameter *a* varying between 1 and 2. Test scores on different forms of the 2013 post operational exams were also equated using a common reference form – D1 and adjusting the test score difficulties of the other 3 test forms D2, D3 and D4 respectively. The 3-parameter IRT logistic model was used for the item analysis for the 8 UOE test forms comprising C1, C2, C3, C4, D1, D2, D3 and D4.

6. Data

Data for the study was extracted from the UTME master file after post-test administration as well as from the trial-test. The trial-test data is made up of responses of data from a representative sample of students from Senior Secondary Class III in the Use of English subject and indeed all other 22 UTME subjects. The students were administered the various test forms in a classroom setting at a period when they were psychologically ready for their senior secondary examination. The tests were administered to students in a scrambled form so that the groups of students taking each form were randomly equivalent. A pre-equating model, which employed the 3-parameter IRT logistic model, was used. The Xcalibre 4.0.0 software was used because of the necessity to have scoring tables prior to test administration. In this study, item parameter estimate and the raw score to theta (e.g., scoring table) relationship for pre-equating model were calibrated and developed on the field test data. To enable a comparison of the difference in equating results between pre- and post-equating, data based on the post-administration for the four different test forms in UOE of the field test of 2012 and 4 different post-administration data of test items in UOE in 2013 CBT were used. Each of the test forms consists of a sample of approximately 650 candidates' responses. In all, the data used is made up of 5,166 responses.

7. IRT pre-equating

Tong *et al.* (2008) defined pre-equating as conducting equating prior to the operational testing. The equating design used in pre-equating the UOE items was the IRT equivalent group equating procedure. In order to pre-equate the test forms in the 2012 UOE, the response data collected during the 2012 field test were first calibrated. Then, one of the test forms comprising of response data from the trial-test was calibrated using "*a* prior" information from previous operational data. Thereafter, the pre-test items were put on the same scale as the one calibrated using information from the operational items through the mean/sigma method. The item parameter estimates from the above step were then used to create the raw-to-scale conversion table for each form to the reference form using IRT pre-equating. The pre-equating process was carried out by applying the following procedures:

a. Estimates of item parameters were produced using the three-parameter IRT model on the 2012 trial-test data.

b. The item parameters were placed onto the reference scale by using the item equivalent group equating design.

c. Some items were selected from the item bank and used along with some pre-test items to build new test forms for parallelism

d. A raw score to theta relationship for these new test forms are developed using the trial-test pre-equated item parameters.

Despite the advantage of using pre-equating as a cushion where immediate score reporting is necessary and as a guide towards reducing incidences of examination security breach, this equating method can be vulnerable to equating errors and bias in a test.

8. IRT post-equating

In carrying out post-equating, the post administration item parameters and scoring table were produced using the operational data. During post equating, all the rules used in pre-equating were simulated during post-equating such as applying the mean/sigma equating method to place the item parameter estimates and scoring tables on the same scale. The following steps as suggested by Kolen *et al.* (2004) were applied during post-equating:

1. Calibrate all items on the operational test form by making the post-operational item difficulties centre at a mean value of zero and obtain raw score to theta scoring table.
2. Obtain mean test score difficulty using the post administration item parameters from the previous stage.
3. Obtain the scaling constant for post-equating by subtracting the mean item difficulty from stage 2 from the mean item difficulty from pre-equating.
4. Adjust all the post-administrational item parameters by adding the scaling constant obtained from stage 3.

9. Test calibration and analysis

A number of procedures can be performed to achieve item calibration and item linking such as carrying out separate calibration with linking, concurrent calibration or fixed parameter calibration. In this study, separate calibrations were carried out on all the test forms using the three-parameter IRT logistic model (3PL). The 3PL is an IRT model that specifies the probability of a correct response to a dichotomously scored multiple-choice item as a logistic distribution that introduces a guessing parameter in addition to the discrimination and difficulty parameters. Estimation of candidates' ability was done using the Maximum Likelihood Estimation (MLE) method. In statistics, MLE is a method of estimating the parameters of a statistical model's given observations by finding the parameter values that maximise the likelihood (or probability) of making the observations, given the parameters. Thereafter, the mean/standard deviation suggested by Livingston (2004) was used in placing the item parameters on the same scale.

10.Assessment criteria

In assessing the pre-equating and post-equating results, one major area of concern is the item parameter estimates. In order to compare the item parameters of two more test forms from post-equating, the two must be placed onto a common operational scale. Statistical methods such as correlation analysis can then be used in comparing the differences in the item parameter estimates obtained between the two. Correlation coefficients obtained are expected to be close to .90 and the average absolute differences between estimates are expected to be below 0.20. This same criteria may be applied when comparing pre- and post-equating results.

It is also important and interesting to observe how different the raw-score-to-theta scoring tables tend to be based on pre-post contrast. In the large-scale assessment context, decisions on classifications are also important. In this study, percentages of students in each of the performance levels are also contrasted between pre- and post-equating. Another reliability index examined is the classification accuracy. This is meant to establish what percentages of students were accurately classified. The classification method adopted by Gao, He and Ruan (2012) was applied to compute classification accuracy index for the pre- and post-equating results. To calculate the classification reliability index for a given ability score θ, the observed score θˆ is expected to be normally distributed with a mean of θ and a standard deviation of *SE*(θ) – the standard error of measurement associated with the given θ. The expected proportion of examinees with true scores in any particular level on high/low or pass/fail classification rates given by different equating methods was also reported. Each test has two cut scores, C and D cuts. Classification rates for the C and D cuts were reported for the UOE test in this study.

While there is no consensus on the best measures of equating effectiveness (Kolen *et al.*, 2004), three commonly employed measures used in equating studies include the Root Mean Square Error (RMSE), the Standard Error of Equating (SEE) and (3) BIAS of the equated raw scores (Pomplun, Omar & Custer, 2004). These measures represent total equating error, random equating error and systematic equating error, respectively. Notice that all three indices were weighted by the frequency of number-correct raw score at each particular level. Total equating error and systematic error were calculated with the formulas below:

$$\text{BIAS} = \frac{\sum_i f_i\,(x_i' - x_i)}{\sum_i f_i} \qquad (1)$$

$$\text{RMSE} = \frac{\sum_i f_i\,(x_i' - x_i)^2}{\sum_i f_i} \qquad (2)$$

where *fi* is the frequency of number-correct raw score level *i*, X_i' is the equated score at each of the number-correct raw score level and X_i is the equated score from IRT pre-equating at the number-correct raw score level *i*.

The standard error of equating is a measure of random equating error and can be estimated with the RMSE and BIAS. The standard error of equating at each possible raw score was estimated with:

$$\text{SEE (fi)} = \sqrt{\text{RMSE}(f_i)^2 - \text{BIAS }(f_i)^2} \qquad (3)$$

where fi is the frequency of number-correct raw score level *i*.

11. Results

Table 1 shows the item parameter estimates disparity between pre- and post-equating results for test forms C1 and D1 representing the base test form for pre-equating and one test form from the post-equating. Columns1 and 2 in table 1 shows the p-values of test forms C1 and D1. Overall, the p-values appear to be higher for the post-equated form than for the pre-equated one. The reason perhaps may be attributed to the prevailing situation during the conduction of the pre-test, as most students do not often take trial-tests as serious as other high stakes examinations. However, the item parameter values from the pre-equating were found not to be different from the post-equating item parameter estimates because of the mean/sigma equating, the average of the item parameter estimates were equated to be the same for pre- and post-equating.

Table 1: Comparisons between pre–equated and post administration item parameter estimates of use of English

Item	Pre-equated item mean (p-value)	Post-equated item mean (p-value)	Pre-equated item parameter (a)	Pre-equated item parameter (b)	Post-equated item parameter (a)	Post-equated item parameter (b)	Pre-post difference
1	0.9983	0.9984	4.6574	-2.974	6	-2.0877	-0.8863
2	0.8831	0.2709	0.8005	-1.526	6	2.0134	-3.5394
3	0.2638	0.6256	1.3985	2.6616	2.6891	1.4339	1.2277
4	0.0885	0.092	1.2125	4	1.8418	2.7751	1.2249
5	0.0634	0.087	1.1432	3.7925	1.6143	3.0446	0.7479
6	0.0568	0.6273	1.1169	4	0.8547	0.2142	3.7858
7	0.0751	0.1084	1.111	4	1.4384	2.8714	1.1286
8	0.7446	0.0969	0.4233	-0.9661	1.5148	3.0391	-4.0052
9	0.0568	0.1051	1.175	3.5508	1.4855	2.9939	0.5569
10	0.0952	0.1002	1.0557	3.7965	1.4307	2.8522	0.9443
11	0.0351	0.0854	1.1231	3.9127	1.4269	2.7354	1.1773
12	0.0701	0.0542	1.1081	3.8988	1.5332	3.057	0.8418
13	0.0534	0.1117	1.1645	3.595	1.4537	2.8764	0.7186
14	0.2354	0.197	0.907	2.9634	1.3474	2.6269	0.3365
15	0.0501	0.1018	1.1076	3.9601	1.446	2.9046	1.0555
16	0.0501	0.4811	1.0984	3.9751	2.4416	0.3931	3.582
17	0.4474	0.1264	0.743	1.0776	1.488	2.9561	-1.8785
18	0.0751	0.0575	1.1289	3.5074	1.5058	2.9411	0.5663
19	0.4073	0.3251	0.6921	2.2887	1.0861	1.5251	0.7636
20	0.0835	0.087	1.1354	3.5565	1.4684	2.8734	0.6831
21	0.0684	0.5386	1.2627	3.2729	1.5105	0.3335	2.9394
22	0.7563	0.2545	0.6847	-0.7316	1.5121	3.0295	-3.7611
23	0.1619	0.2397	1.2698	3.0003	1.1595	1.8202	1.1801
24	0.1703	0.1166	1.1809	3.265	1.4871	2.9837	0.2813
25	0.606	0.1888	0.6372	0.117	1.5196	3.044	-2.927
26	0.3122	0.1987	1.1833	2.9771	1.5164	3.0494	-0.0723
27	0.5476	0.2562	0.8543	0.3928	1.4251	2.8371	-2.4443
28	0.1336	0.289	1.1425	2.9324	1.1929	1.4307	1.5017

Item	Pre-equated item mean (p-value)	Post-equated item mean (p-value)	Pre-equated item parameter (a)	Pre-equated item parameter (b)	Post-equated item parameter (a)	Post-equated item parameter (b)	Pre-post difference
29	0.5042	0.2841	0.624	0.7679	1.3913	2.7849	-2.017
30	0.2404	0.1478	1.0675	3.0722	1.493	2.9872	0.085
31	0.1386	0.3465	1.1119	2.8014	1.0235	1.2874	1.514
32	0.3706	0.1757	0.9611	2.5122	1.5055	2.9795	-0.4673
33	0.4558	0.1724	1.0753	0.8105	1.4874	2.9871	-2.1766
34	0.2237	0.1429	1.0618	2.5839	1.4775	2.9762	-0.3923
35	0.1135	0.2053	1.2798	2.8837	1.2783	2.5516	0.3321
36	0.0618	0.0788	1.1115	3.9954	1.4763	2.9489	1.0465
37	0.1018	0.1199	1.3219	2.8851	1.4842	2.9842	-0.0991
38	0.0534	0.4483	1.1268	3.7298	2.2712	0.5067	3.2231
39	0.172	0.4351	1.2424	2.9521	1.9371	0.5428	2.4093
40	0.0952	0.1232	1.2425	3.2605	1.4904	3.0011	0.2594
41	0.0501	0.4943	1.3752	3.0139	2.9113	0.3305	2.6834
42	0.828	0.3268	0.7107	-1.2719	1.5245	3.0592	-4.3311
43	0.8731	0.3415	1.0836	-1.2877	1.5193	3.05	-4.3377
44	0.8514	0.3333	0.9867	-1.2055	1.5167	3.0517	-4.2572
45	0.0935	0.4631	1.2949	3.1726	2.0128	0.5022	2.6704
46	0.0918	0.1051	1.2881	3.171	1.5082	3.0269	0.1441
47	0.1085	0.2085	1.2021	3.4037	1.4827	2.9819	0.4218
48	0.0701	0.1067	1.3129	3.1915	1.4777	2.9619	0.2296
49	0.7462	0.1297	1.0897	-0.5341	1.5085	3.032	-3.5661
50	0.6761	0.243	0.6311	-0.3727	1.4968	3.0168	-3.3895
51	0.1369	0.0952	1.0308	3.8366	1.5069	3.0255	0.8111
52	0.0568	0.1051	1.2974	3.1161	1.4843	2.979	0.1371
53	0.0985	0.2135	1.2716	3.1313	1.364	2.5987	0.5326
54	0.0584	0.0706	1.1358	3.7712	1.4739	2.8864	0.8848
55	0.1736	0.4959	1.2232	3.1266	1.0059	0.6887	2.4379
56	0.1152	0.1248	1.0961	3.972	1.5216	3.0533	0.9187
57	0.0518	0.1248	1.3067	3.1408	1.4027	2.793	0.3478

Item	Pre-equated item mean (p-value)	Post-equated item mean (p-value)	Pre-equated item parameter (a)	Pre-equated item parameter (b)	Post-equated item parameter (a)	Post-equated item parameter (b)	Pre-post difference
58	0.7813	0.2463	1.0157	-0.7803	1.5232	3.0553	-3.8356
59	0.0801	0.0837	1.3068	3.1834	1.5074	3.0035	0.1799
60	0.7179	0.197	1.1117	-0.4228	1.5052	3.0171	-3.4399
61	0.8197	0.2677	1.2243	-0.8248	1.5119	3.0391	-3.8639
62	0.0785	0.1182	1.311	3.1662	1.4906	2.9987	0.1675
63	0.1619	0.1527	1.1741	2.9733	1.5076	3.0316	-0.0583
64	0.0668	0.0788	1.1265	3.746	1.5092	3.0204	0.7256
65	0.1135	0.1264	1.2463	3.1242	1.5077	3.0346	0.0896
66	0.7295	0.2135	1.1906	-0.3924	1.4969	3.0122	-3.4046
67	0.6678	0.1363	1.2154	-0.1831	1.5131	3.0425	-3.2256
68	0.1219	0.1166	1.2231	3.3843	1.5007	3.0219	0.3624
69	0.0634	0.4269	1.3126	3.1709	2.1383	0.5417	2.6292
70	0.172	0.2003	1.1629	3.3349	1.4372	2.8649	0.47
71	0.0551	0.0772	1.18	3.7017	1.4915	3.0104	0.6913
72	0.621	0.2874	0.806	-0.015	1.4709	2.927	-2.942
73	0.8397	0.4089	0.8289	-1.229	1.5122	3.0493	-4.2783
74	0.7346	0.1757	1.1683	-0.4817	1.5076	3.0023	-3.484
75	0.0902	0.0542	1.1948	3.285	1.5017	3.0048	0.2802
76	0.1085	0.468	1.286	3.129	2.1233	0.5758	2.5532
77	0.1035	0.1856	1.2776	3.129	1.4768	2.9757	0.1533
78	0.202	0.1494	1.2011	3.0715	1.488	3.0004	0.0711
79	0.1386	0.0887	1.2572	3.152	1.5039	3.0131	0.1389
80	0.0668	0.3924	1.2062	3.3302	2.127	0.6414	2.6888
81	0.1135	0.2299	1.241	3.0987	1.1827	2.1068	0.9919
82	0.7646	0.2693	0.9913	-0.7055	1.5065	3.0265	-3.732
83	0.0735	0.3777	1.2802	3.0588	1.9903	0.7027	2.3561
84	0.0818	0.0788	1.1916	3.3853	1.483	2.9097	0.4756
85	0.1803	0.1839	1.1311	3.3307	1.4995	2.9886	0.3421
86	0.0935	0.1166	1.1268	3.5994	1.4732	2.9424	0.657

Item	Pre-equated item mean (p-value)	Post-equated item mean (p-value)	Pre-equated item parameter (a)	Pre-equated item parameter (b)	Post-equated item parameter (a)	Post-equated item parameter (b)	Pre-post difference
87	0.0868	0.1675	1.1469	3.6049	1.3443	2.7093	0.8956
88	0.1753	0.1297	1.0478	3.8787	1.4722	2.9693	0.9094
89	0.1336	0.1658	1.0891	3.5808	1.4459	2.891	0.6898
90	0.1068	0.1297	1.0719	3.8811	1.4595	2.8438	1.0373
91	0.222	0.1741	1.0369	3.8145	1.4599	2.9198	0.8947
92	0.0451	0.0427	1.221	3.4338	1.5177	3.0287	0.4051
93	0.8347	0.353	0.7053	-1.3243	1.5203	3.0551	-4.3794
94	0.1536	0.1248	1.2109	2.6945	1.5037	3.0246	-0.3301
95	0.0851	0.0558	1.115	3.7291	1.5172	2.9931	0.736
96	0.0835	0.0887	1.1955	3.419	1.494	2.9979	0.4211
97	0.7646	0.3103	0.849	-0.689	1.5052	3.0101	-3.6991
98	0.8047	0.3235	0.7359	-1.0372	1.5098	3.0369	-4.0741
99	0.0952	0.1494	1.2025	3.4407	1.3969	2.7885	0.6522
100	0.0863	0.1084	1.1259	3.6137	1.4341	2.8826	0.7311

The average absolute difference between the item parameter estimates were computed as .000342 for C1 and D1, .00491 for C2 and D2, .00572 for C3 and D3 and .00557 for C4 and D4. In addition, all were found to be less than the benchmark of .20. Table 2 also shows the correlations between pairs of pre-equating and post-equating item parameter estimates of C1D1, C2D2, DC3D3 and C4D4. The results revealed correlation coefficients of .995**, .954**, .995** and .996**.

Table 2: Correlation of pre-equating and post-equating item parameters

		C1_Pre	C2_Pre	C3_Pre	C4_Pre
D1_Post	Pearson Correlation	.995**	0.036	0.062	0.082
	Sig. (2-tailed)	0	0.722	0.54	0.42
	N	100	100	100	100
D2_Post	Pearson Correlation	0.026	.994**	0.071	-0.1
	Sig. (2-tailed)	0.795	0	0.484	0.323
	N	100	100	100	100
D3_Post	Pearson Correlation	0.06	0.076	.995**	0.191
	Sig. (2-tailed)	0.555	0.455	0	0.056
	N	100	100	100	100
D4_Post	Pearson Correlation	0.076	-0.111	.212*	.996**
	Sig. (2-tailed)	0.454	0.274	0.035	0
	N	100	100	100	100
**. Correlation is significant at the 0.01 level (2-tailed).					
*. Correlation is significant at the 0.05 level (2-tailed).					

Making decisions from the criteria earlier stated in assessment criteria (i.e., correlation being 0.90 and average absolute difference being less than 0.20), the item parameter estimates between the two equating models are the same. Figures 1, 2, 3, and 4 also show the scatter plot of the relationship between the pre-equating and post-equating test forms.

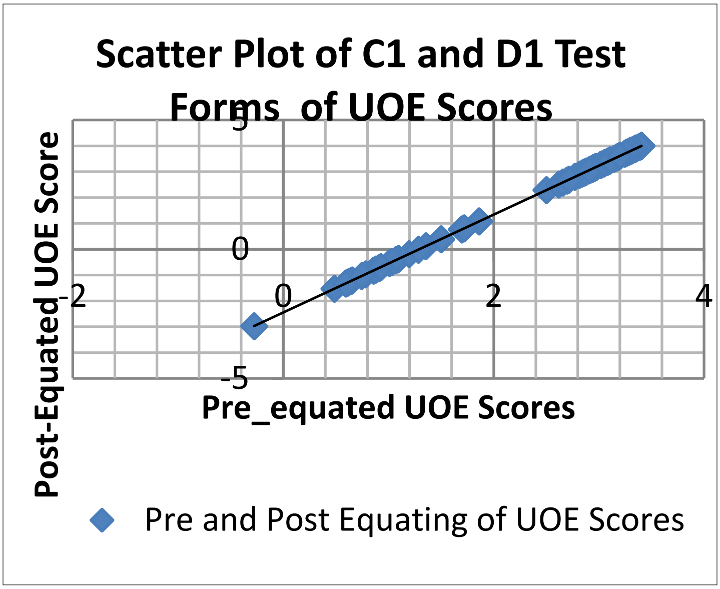

Fig 1: Scatter plot of relationship between pre-equating and post-equating of C1 and D1

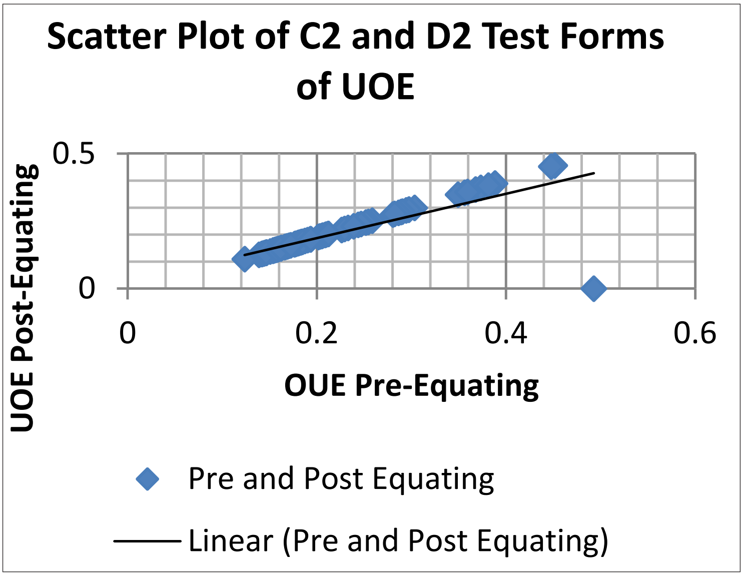

Fig. 2: Scatter plot of relationship between pre-equating and post-equating of C2 and D2 forms

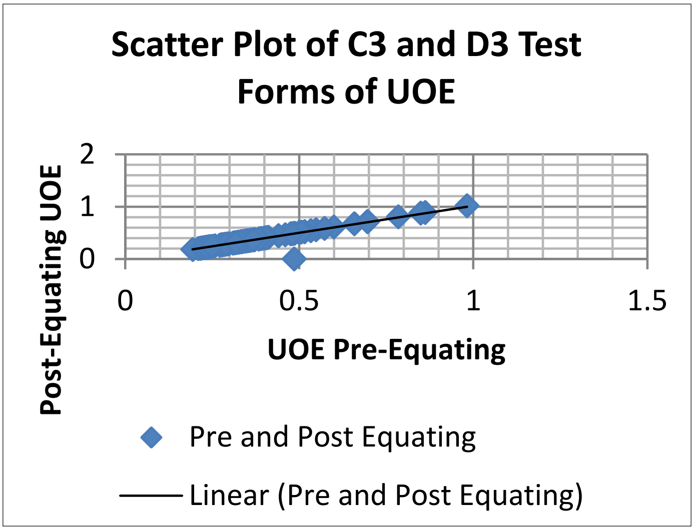

Fig 3: Scatter plot of relationship between pre-equating and post-equating of C3 and D3 test forms

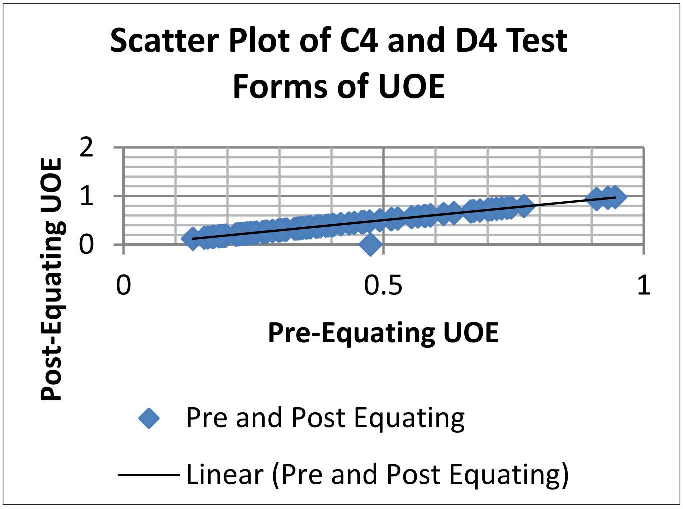

Fig. 4: Scatter plot of relationship between pre-equating and post-equating of C4 and D4 test forms

All the items constituting the two different forms were aligned to the linear straight line showing a highly close relationship. In the same way, figures 5, 6, 7 and 8 depict the raw score-to-theta-scoring tables based on the two equating models mentioned above. While the horizontal axis represents the ability estimates, the vertical axis represents raw scores. From the figures, it is certain that the raw score-to-theta scoring tables for pre-equating and post-equating models were overlapping each other.

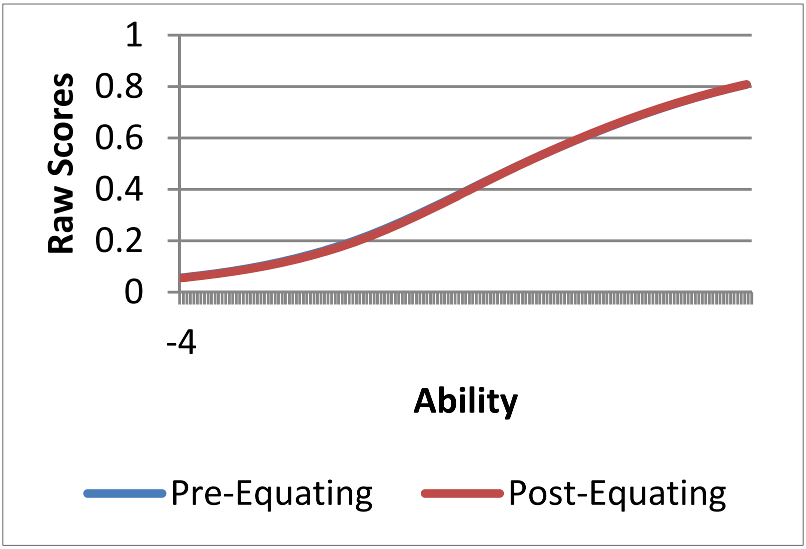

Figure 5: TCC of test forms C1 and D1

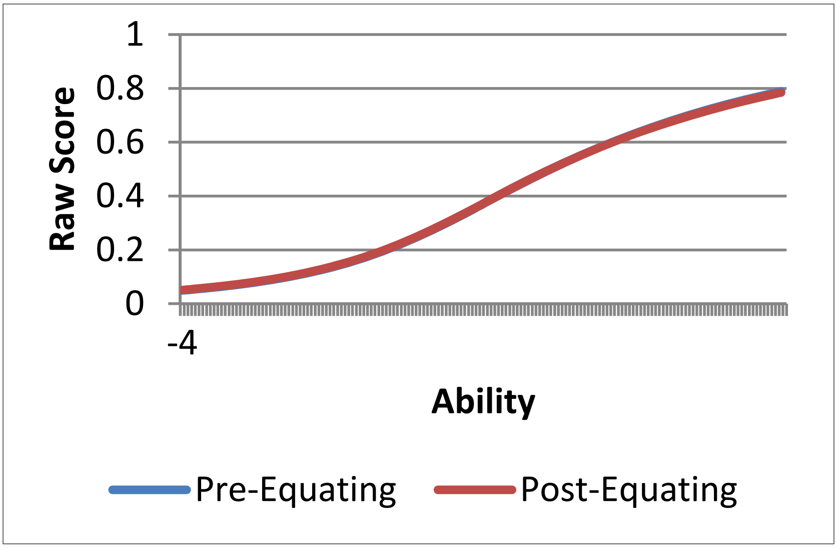

Figure 6: TCC of test forms C2 and D2

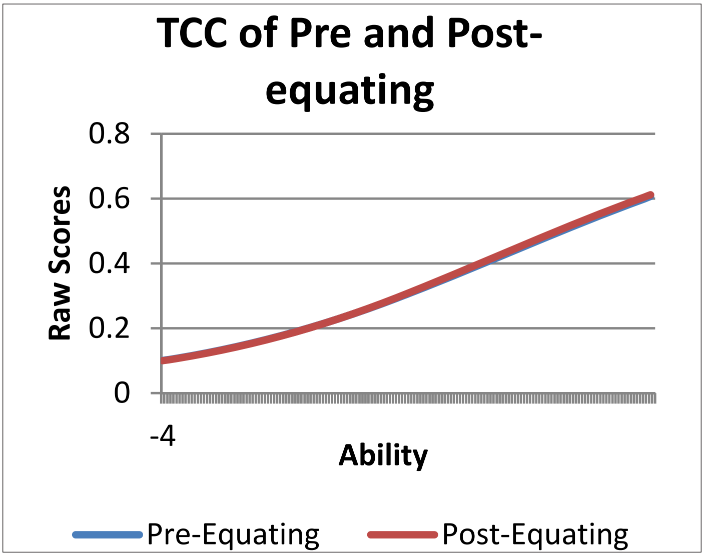

Figure 7: TCC of test forms C3 and D3

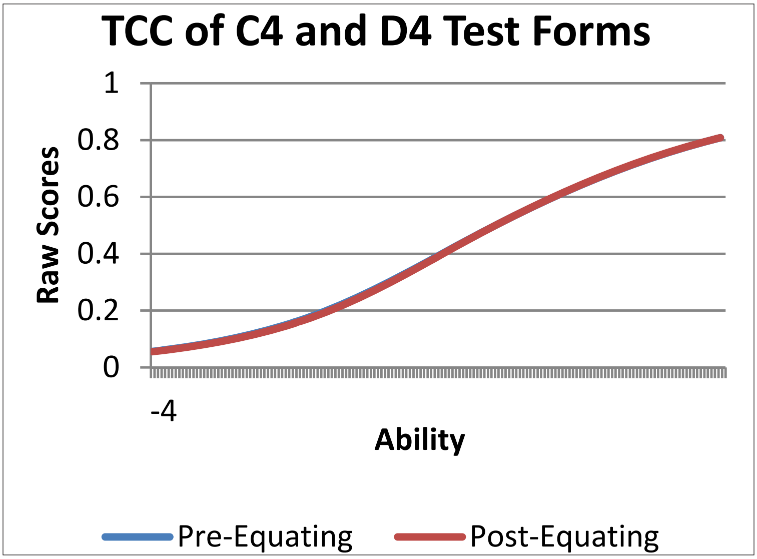

Figure 8: TCC of test forms C4 and D4

Table 3 shows that for the classification rate, the IRT post-equating tended to pass more examinees than the pre-equating methods in total. The table shows that the IRT pre-equating method tended to pass fewer examinees than the IRT post-equating method at the C cut and in total. However, the reverse is the case for the D cut, where the pre-equating method passed more candidates in test forms C1, C2 and C4.

Table 3: Classification frequency for aggregate pass rate, C-pass and D-pass rates for the UTME UOE

Test form	Equating method	No.	Total high (N)	Total high (%)	C-high (N)	% C-high	D-high (N)	% D high
C1	Pre		559	45.48	418	34	147	11.96
		1229						
D1	Post		670	54.51	565	45.97	105	8.54
C2	Pre		563	46.87	345	28.72	218	18.15
		1201						
D2	Post		638	53.12	452	37.63	186	15.48
C3	Pre		534	45.44	438	37.27	96	8.17
		1175						
D3	Post		641	54.55	543	46.21	98	8.34
C4	Pre		694	44.37	487	31.13	207	13.23
		1564						
D4	Post		870	55.62	671	42.9	199	12.72

The means and standard deviations of the equated scores from different equating methods are shown in table 4. From the table, it can be seen that the item parameters of the test forms from the pre-equating and post-equating consistently yielded almost the same values except for test forms C1, representing pre-equating and the corresponding D2 for post-equating which has slightly higher means and SDs.

Table 4: Means and standard deviations of the equated scores from different equating methods

Test Form	IRT-Pre equating		Test Forms	IRT Post equating	
	Mean	SD		Mean	SD
C1	65.045	14.757	D1	65.034	14.818
C2	64.991	14.921	D2	64.989	14.95
C3	64.82	14.503	D3	64.806	14.438
C4	64.87	14.784	D4	64.917	14.774

Finally, table 5 also presents the results of the three indices used to evaluate the equating results with IRT pre-equating results as the baseline. All three indices indicated that the IRT post-equating yielded closer results to the IRT pre-equating method by having the smaller RMSD, BIAS and SEE in all four of the test forms.

Table 5: Indices used in evaluate the equating results with IRT pre-equating as the baseline

Test Form	RMSE	BIAS	SEE
	IRT Post		
D1	0.01857	0.000345	0.018567
D2	0.07042	0.004959	0.070245
D3	0.07612	0.005794	0.075899
D4	0.07503	-0.00563	0.074818

12.Discussions on results

The perception on the higher p-values from the post-equating method can probably be explained. During field trials, the items constituting the UOE were administered in paper-and-pencil mode while the same items used in subsequent operational examination was done in a computer-based testing environment. The difference in the modes of examination could be a direct consequence for the perceived difference between the pre-equating method and post-operational method. The design of the UTME delivery system made it possible to include innovations such as the use of the four arrow keys on the keyboard as an alternative to the use of the mouse, review of items to reveal unanswered items prior to submission as well as inclusion of a timer among other things. These features added value to the test delivery system, distinguishing it from the paper-and-pencil mode of testing.

The seriousness or stake attached to the two examinations may also have contributed to the difference in the p-values observed. Since the trial-test does not often attract motivational gains, students often do not take the examination as serious as the UTME high-stakes examination. This could account for the difference in the overall performance of the candidates. Again, the level of preparedness of the students can constitute its own problem as well, which also affects performance.

Observing the performance of the candidates through direct examination of the p-values shows that for instance, test forms C1 and D1could offer more insight into differences in pre-equating and post-operational methods. Test form C1 represents the pre-equating while D1 stands for the post-equating method. Of the 100 items tested, 56 of them were found to be harder in the pre- than in the post-equating test form. Experience has shown that in the trial-testing situation, candidates are often less serious in taking examinations possibly because of a lack of motivation on the perceived consequences of the test. Wolf and Smith (1995) presented a research study, which showed that testing students in consequential condition compels them to out-perform other students in a non-consequential condition by an effect size of .26. They concluded that consequences influences motivation and motivation influences performance.

It is certain therefore that motivation is a likely contributor to performance differences found in this study between students that took the field test compared to students that took the UTME high stakes assessment. Indeed, it appears reasonable to say that students taking the field test according to Damaleski (2006) would not exert as much effort since no stakes were associated with this test event and, in fact, no student level results were ever reported. This lack of seriousness regarding trial-tests by students often accounts for the high rates of omitted and unreached items seen in many field tests and this possibly explains reasons why trial-test items were found to be harder due to the relatively large amount of missing or incomplete data.

The equality argument for fairness in assessment according to advocates assessing all students in a standardised manner using an identical assessment method, content and same administration, scoring and interpretation procedures. With this approach to assuring fairness, if different groups of test takers differ on some irrelevant knowledge or skills that can affect assessment performance, bias will exist. This situation is avoided by ensuring that pre-equating is carried out prior to real test administration. The analysis carried out in this study has shown that the pre-equating and post-equating methods have provided comparable results. This will mitigate the fears of stakeholders who are apprehensive of whether pre-equating is actually doing what it is supposed to do or providing validity evidence as to the equivalency of the test forms used in testing in the UTME UOE.

13.Conclusion/Recommendation

The result of this study has shown that all three major indices involving RMSE, BIAS and SEE which represent total error, systematic error and standard equating error indicated that the IRT post-equating yielded closer results to the IRT pre-equating method and are therefore comparable. However, carrying out equating using IRT is complex, both conceptually and procedurally.

Another score point for the post-equating method is that the method passed more candidates than the pre-equating especially in the total and c-cut. This shows that the field test items are predicting performance of candidates in the UTME operational examination. These results are pointers to the fact that item parameters obtained during the trial-test were remarkably equivalent to those obtained during the operational assessment of UTME in the UOE. All other 22 UTME subjects were also subjected to pre-equating prior to operational test administration and similar results were achieved. The extent to which those inferences are appropriate for different groups of test takers is an important aspect of fairness

The practice of using the pre-equating method to build score tables prior to an operational assessment should be sustained since the method yielded comparable results with the post-equating method. This occurs as long as the probable cause for pre-equating error such as the presence of bias in the item parameter estimates, which are caused by the violation of the assumption of item local independence, are removed (Kolen & Brennan, 2004). Pre-equating test forms prior to test administration in actual examination is a good way of assuring equity and fairness in assessment. When the tests given to the students are unbiased and function the same way for different groups of test takers, fairness is said to have been built into the test.

References

Domaleski, C.S. 2006. *Exploring the efficacy of pre-equating a large scale criterion-referenced assessment with respect to measurement equivalence*. Published PhD thesis. Ann Arbor, MI: ProQuest Information and Learning Company.

Gao, R., He, W. & Ruan, C. 2012. Does pre-equating work? An investigation into a pre-equated test let-based college placement exam using post administration data. *ETS Research Report Series*, 2012, i–18. https://doi.org/10.1002/j.2333-8504.2012.tb02294.x

Holland, P.W. & Dorans, N.J. 2006. Linking and equating. In R.L. Brennan. (Ed.). *Educational measurement,* 4th ed. Westport, CT: American Council on Education and Praeger Publishers. pp. 187-220.

Kolen, M.J. & Brennan, R.L. 2004. *Test equating, scaling and linking: Methods and practices,* 2nd ed. New York: Springer –Verlag. https://doi.org/10.1007/978-1-4757-4310-4

Kirkpatrick, R.K. 2005. The effects of item format in common item equating. Unpublished doctoral dissertation. Iowa: University of Iowa

Kirkpatrick, R. & Way, W.D. 2008. Field testing and equating design for state educational assessment. *A paper presented at the annual meeting of the American Educational Research Association*, New York.

Livingston, L. 2004. *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

Pomplun, M., Omar, H. & Custer, M. 2004. A comparison of Winsteps and Bilog-MG 144. for vertical scaling with the Rasch model. *Educational and Psychological Measurement*, 64, 600-616. https://doi.org/10.1177/0013164403261761

Raju, N.S., Laffitte, L.J. & Byrne, B.M. 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517-529. https://doi.org/10.1037/0021-9010.87.3.517

Tong, Y., Wu, S-S. & Xu, M. 2008. A comparison of pre-equating and post-equating using large-scale assessment data. *Paper presented at the American Educational and Research Association annual meeting*, New York City.

Wolf, L.F. & Smith, J.K. 1995. The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8(3), 227-242. https://doi.org/10.1207/s15324818ame0803_3

Xuan, T & Rochelle, M. 2011. Why do standardized testing programs report scaled scores? Why not just report the raw or percent-correct scores? *R&D Connections,* 16, 1-6.