JELS, Vol. 1, No. 3, Spring 2010, 113-128
IAUCTB

Test Score Equating and Fairness in Language Assessment

Purya Baghaei
Assistant Professor of Applied Linguistics, Islamic Azad University, Mashad Branch, Iran

Abstract

Equating test scores is an important issue in large scale testing. Almost all standardized tests have several forms which vary in difficulty. New forms are written and added every year. When the items in different forms of a test vary in difficulty, direct comparison of test-takers who have taken different forms and are at the same ability level is not possible; hence, the issue of test fairness arises. In such situations there is a need for equating test scores so that standards can be maintained from year to year. That is, there is the need to adjust the scores for the difficulty of the test forms and report a scaled score that is comparable across all forms of the test. In this study, two forms of a reading comprehension test were equated and the pass/fail decision consistency was investigated under two conditions of with and without equating. Concurrent common item equating with one parameter Items Response Model (IRT) was used to equate the two test forms. Results showed that the lack of equating leads to unfair pass/fail decisions. The implications for high-stakes large scale testing are thus discussed.

Keywords: Test score equating, Item Response Theory, common item equating, common person equating, concurrent equating, test fairness

Introduction

Testing companies normally administer their assessments more than once during a year and for security reasons, they cannot use the same test form over different administrations. This requires preparation of several test forms to be used in each administration of the test. The problem that arises as a result of having several test forms is the differing difficulty levels of the test forms and the incomparability of the abilities of examinees who take the different forms. Hambleton, Swaminathan, and Rogers (1991) state that if different groups of examinees take different tests, comparison among them is QRWSRVVLEOH ,Q RUGHU WR P DNH WKH H[DP LQHHV¶ SHUIRUP DQFH FRP SDUDEOH across different test forms, a procedure called equating is required. Equating

is a statistical process to adjust scores on different test forms to make them comparable (Kim & Hanson, 2002).

Equating test scores is an important issue in large scale standardized testing (Saida & Hattori, 2008; Shin, 2009). Standardized testing requires the stability of the pass-fail criteria over different runs of the assessment (Cook & Eignor, 1991). If a test in spring, for instance, tends to be somewhat easy, an examinee with a certain ability level might pass. If another examinee with the same level of ability takes the same assessment from the same company in fall, s/he may fail simply because the fall version of the test happened to be harder. Of course attempts are made to make all the test versions equally hard to avoid such complications but the problem is inevitable.

This problem exists in Iranian assessment contexts as well. Universities in Iran run English proficiency tests once a year to admit students to PhD programs. No attempt is made to make the passing standards comparable over different years. Therefore, in addition to all other factors, it seems that pass and fail to a certain extent depends on the year an examinee takes the test.

Another situation in which equating is essential is in longitudinal trend studies where following the trends of a particular ability or attitude over time LV RI FRQFHUQ 7R GHWHUPLQH VXGHQWV¶ SURJUHVV RYHU WLPH RU WHQGV LQ WKH abilities of different generations of students, educators and education policymakers have to compare them on a common scale of measurement (Saida & Hattori, 2008). For such comparisons that are usually made at international (such as Program for International Student Assessment, or PISA), national, or state-wide scales in North America and European countries, a single test cannot be used every year over long periods which are sometimes as long as two decades. Moreover, it is impossible to develop several forms of a test with exactly identical difficulty and reliability. Each year a separate test is used to measure the abilities of students. If the performance of the incoming students in a given year is better than the performance of the students in the preceding year, one can explain the difference in two ways. Either the incoming students are more proficient or the test they took was easier (Saida & Hattori, 2008). In order to exclude the second explanation and make sure that the observed trends are a valid UHIOHFWLRQ RI FKDQJHV LQ VWXGHQWV¶ DELOLWLHV RYHU WLPH WKH DVVHVVPHQW mechanism needs to equate scores from different test forms and adjust for variations in form difficulty.

Another context in which equating meets the surface is in program evaluation studies. Changes are interpreted in the scores as the result of

participation in the program in question. One of the threats to the internal validity of this inference is whether the observed changes in the scores are truly because of participation in the program or whether they can be DWWULEXWHG WR RWKHU YDULDEOHV VXFK DV SDUWLFLSDQWV¶ PDWXUDWLRQ DQG PRUWDOLW\ and changes in the measurement instrument rather than the treatment itself, among other things (Wolfe & Chiu, 1997). Wolfe and Chiu maintain that

> *Summated composite scores are not comparable across times when items are added, removed, or reworded; items are skipped by some subjects; or response options change from pre-test to post-WHVW DQG SURFHGXUHV WDWDUH FRPPR H T H WR DUH DUH H EHFD H RI SURFHGXUHV F D WH H FDQ J PHWR DUH often used to place measures from different administrations of a questionnaire onto a common scale.* (p. 7)

Consequently, in order to solve the problem of unequal difficulty of the tests, psychometricians have come up with certain empirical procedures called HTXDWLQJ DV H[SODLQHG DERYH 7HVWHTXDWLQJ LV D SURFHGXUH ³WR SURYLGH comparable scores on multiple forms of the same test, consequently avoiding some of the possible inequities that could occur if one examinee took a more GLIILFXOWRUP RI D WHVWWKDQWKDWWDNHQE\ DQRWKHUH[DPLQHH &RRN (LJORU 1991, p. 191). In the literature on equating, the form to which we adjust the QHZ IRUP LV UHIHUUHG WR DV WKH µHIHUHQFH IRUP¶ DQG WKH VFRUHV WKDWDUH FRP SDUDEOH DFURVV IRUP V DUH FDOOHG µVFDOHG VFRUHV¶

Equating Methods under Classical Test Theory

There are several methods for equating test scores, some of which are within classical test theory and some within IRT. The most common equating methods under the classical test theory are mean equating, linear equating, and equipercentile equating. In mean equating, the difference between the means of the two populations who take the two forms of a test is computed. This difference is then added to the scores of all the examinees who have taken the harder test form or subtracted from the score of those who have taken the easier version (Kolen & Brennan, 1995). Here the assumption is that the means of the two populations who have taken the two test forms are equal and any difference in their means is the result of differences in test GLIILFXOW 7KH P DIRU SURFHGXUH ZLWK P HDQ HTXDWLQJ LV WKDW³DGGLQJ WKH VDP H QXP EHUR SRLQW VR HYHU RQH¶V VFRUH LV QRW D JRRG ZD\ WR GHVFULWKH VFRUHV (Livingston, 2004, p. 13).

In linear equating, the scores on the new form are transformed to scores on the reference form that are the same number of standard deviations below or above the mean (Livingston, 2004). The mean of the adjusted scores on the new form are made equal to the mean of the scores on the reference form. The scores one, two, and three standard deviations below and above the mean in the adjusted form are also made equal to the scores one, two, and three standard deviations below and above the mean in the reference form. The procedure is also done for all the scores in between. In other ZRUGV ³VFRUHV RQ WZR WHVWV DUH HTXLYDOHQWL WKH\ FRUUHVSRQG WR HTXDO WDQGDUG VFRUH GHYLDWLRQV´ $QJRII S +RZHYHU QRQHDUHTXDWLQJ LV YHU VDP SDI GHSHQGHQW DQG WKXV ³: KHQ WKH WZR IRUP V RI WKH WHVWV GLIHULQ difficulty, the linear equating in a strong test-taker group can differ noticeably IURP WKH QQHDUHTXDWLQJ LQD ZHDN WHVWWDNHUJURXS´ /LYLQJVWRQ S

In equipercentile equating, scores on the new form are transformed to a score on the reference form that has the same percentile rank in the reference form group. In other words, two scores from the new form and the reference form are equivalent if they have the same percentile ranks in their corresponding groups. The problem with equipercentile equating is that adjustments do not exist for scores beyond the observed range of scores obtained by the test-takers. As Livingston (2004) pinpoints

> *One limitation of equipercentile equating is that the equating relationship cannot be determined for the parts of the score range above the highest score you observe and below the lowest score R RE HUH LWFD EH D SURE OP DW LJ FRUH GI HO R D LLIF 0V test, because some future test-taker may get a raw score higher than the highest score in the data used for the equating.* (p. 23)

Equating Methods under Item Response Theory

In equating, the conversion of raw scores to scaled scores, that is, scores which are comparable across forms, should be independent of the sample used to derive the scale scores. One of the most important properties of IRT is that if the data fit the IRT model, estimates of item difficulty and person ability parameters are independent of the sample of persons and the test form used for item analysis and person measurement. That is, person ability estimates will remain identical, within measurement error, regardless of whether an easy form or a hard form of a test has been given, after the two tests have been brought onto the same scale (Baghaei, 2009). This makes IRT a straightforward method for equating. In fact, under IRT, there is no

need for equating; only the measures from two different forms should be placed on the same scale.

IRT test calibration results in test-free person measurement and person-free item calibration. Two different calibration runs of the same test are only different in the origin of the scale of measurement (Wright & Stone, 1979). Setting a common origin for two analyses is in fact equating in IRT. In order to bring estimates from two separate IRT analyses onto the same scale, a transfromation is necessary.

Common Item Equating

In this method of equating, there are some items which are shared by both test forms. The difficulty estimates of these common items are used to adjust the person measures for the difficulty of the test forms. The difficulty of the common items estimated from separate analyses should first be compared before using them as anchor items. If the difficulty of the items from separate analyses differ greatly they cannot be used as anchors for equating (Skaggs & Wolf, 2010).

In common item equating, there are a number of items which are shared between the two test forms. Since the origin of an IRT scale is usually the mean of the items, the difference in the mean difficulty of the common items from the two analyses, which is called the *shift constant*, is used to transfrom the estimates from one test to another test. The shift constant is added to the ability estimates of all the examiness who have taken the harder form (Wright & Masters, 1982). In this way, the estimates from the two tests are brought onto the same scale. Wright and Masters state that five to 15 common items are required for common item equating to give stable results. Misfitting item pairs should also be removed to improve the quality of the link.

To investigate whether common items have behaved similarly in two different forms, we need to cross plot their difficulty estimates from the two analyses on the *x* and *y* axis. If the dots form a straight line, with a slope close to unity, then the items are useful for equating. If there are some dots that fall far from the line of best fit, their corresponding items should be dropped from the equating and should not be considered as anchor items. If after dropping these items, the dots still remain far from the line of best fit, Celsius-Fahrenheit equating is required (Linacre, 2007) in which equating the origin and unit of the scale are changed. The origin will be the intercept of the line of best fit and the unit will be one divided by slope of the slope of the line

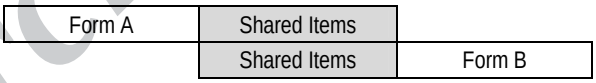
of best fit. The origin and unit are arbitrary in the IRT analysis. The origin of the scale is usually set at the mean of the item difficulty estimates and the unit is 1. This change of the mean and unit aligns the two tests.

Common Person Equating

The procedure for common person equating is exactly the same as common item equating. The only difference is that in common person equating, there are some persons who have taken both test forms. These persons should be spread across the ability scale. The ability estimates of these common persons are estimated on both test forms and then cross plotted. The ability estimates from the two analyses should form a straight diagonal line with a slope close to unity. If they do, the equating procedure explained for common person equating is repeated. That is, the shift constant will be the difference in the average ability of the common persons on the two forms. If not, Celsius-Fahrenheit equating as explained above is required (Linacre, 2007).

Concurrent Equating

Concurrent common item equating is also possible. That is, instead of computing the shift constant and adding it to the measures from the harder form, the analyst can easily set up one data file with all the items and persons from both forms, as is shown in Figure 1, and analyze them in one go to obtain measures for all persons and items on a common scale. This is the most common and straightforward method. In this equating design, there are considerable blocks of missing data, that is, items which are not administered. The IRT software can treat the unadministered items as missing data, exclude them from the calibration process, and estimate the parameters from the data matrix without problems (Smith & Kramer, 1992).



) L R F FRP P R L P T D L L KD D D KR K
common items)

Virtual Equating

In cases where there are no common items in both test forms and there are no common persons to link both forms of the test, virtual equating (Luppescu, 2005) is recommended. In virtual equating, pairs of items of similar content and difficulty in the two tests are selected. These are the pseudo-common items (Linacre, 2007). The procedure explained above for common item equating is repeated here to put all measures on a common scale.

The problem which is raised here is that since the items which link the two forms are not really the same and common to both forms, the quality of the link might be negatively affected. On the contrary, if research shows that virtual equating, that is, equating with no common person and items works as well as the other equating methods, this administratively easier method can always be employed.

Establishing a link by common items is always procedurally cumbersome and requires a huge amount of clerical work. There is also the additional problem of administering at least two test forms to one group of persons, which is something almost impossible under real testing circumstances.

It should be mentioned that equating methods only work for test forms which have minor differences in difficulty, and large differences in difficulty cannot be accounted for by these methods. The other important point is that the two forms should measure the same ability and have equal reliabilities. &RRN DQG (LJORU VWDWH WKDW³« D WHVW IRUP V PHDVXULQJ GLIIHUHQW DELOLWLHV FDQQRW EH HTXDWHG >« @ E REVHUYHG VFRUHV >« @RQ QHTXDO UHDEOH WHVW IRUP V FDQQRW EH IRUP DO HTXDWHG >« @DQG F REVHUYHG VFRUHV RQ WHVW IRUP V RI YDULQJ GLIILFXOW\ FDQQRW EH HTXDWHG´ S

Based on the importance of equating in making scores comparable on administrations of different forms of the test as mentioned in the introduction section and grounded on what was explained above regarding different methods of equating, the researcher posed the following research question:

Does the lack of test score equating lead to unfair pass/fail decisions?

In the following sections, it is demonstrated how common item equating was employed to answer the research question.

Method

Participants

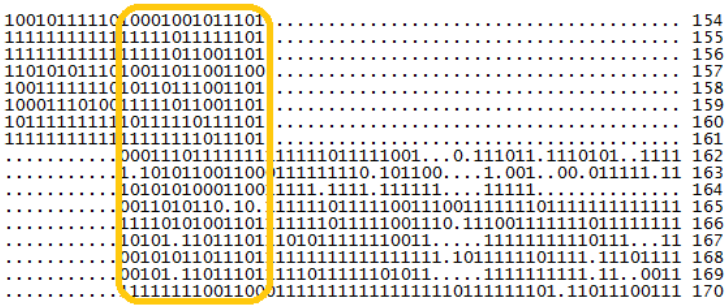
A sample of 264 undergraduate students of English Literature, English Translation, and Teaching English as a Foreign Language were recruited for this study. The participants were at different years of their studies in three English departments at Islamic Azad University, Ferdowsi University, and Emam Reza University in Mashad. Several instructors were asked to administer the tests during their normal class sessions for the purpose of this project. Test results had no consequences whatsoever for the participants and they were informed of this prior to testing. The sample comprised 169 females and 95 males, aged between 19 and 33 years with the average age of 22.2 (SD = 3.4).

Instrumentation

Two parallel English reading comprehension tests, Form A and Form B, were employed for this study. Items were randomly assigned to the two forms from a SRRO RI PRUH WKDQ UHDGLQJ FRPSUHKHQVLRQ LWHPV IURP %URZQV (Sharpe, 2004) and Longman TOEFL practice tests (Philips, 1996). The two forms contained 54 multiple-choice items and shared 14 common items. Items one to 40 in Test A were unique and items 41 to 54 were common or anchor items. Items one to 14 in Form B were common items, that is, they were exactly items 41 to 54 in Form A but items 15 to 54 were unique.

Procedure for Data Analysis

Forms A and B were randomly distributed among the 264 participants of the study. Form A was taken by 160 students and Form B was taken by 104 students at the universities mentioned above. Figure 2 below shows part of the data setup for common item equating. Students 1 to 160 took Form A and students 161 to 264 took Form B. As Figure 2 shows, the last 14 items in Form A, which are the first 14 items in Form B, were common or anchor items, that is, they were taken by both groups. In fact, these items linked the two datasets. The rest of the items in the two forms were unique.

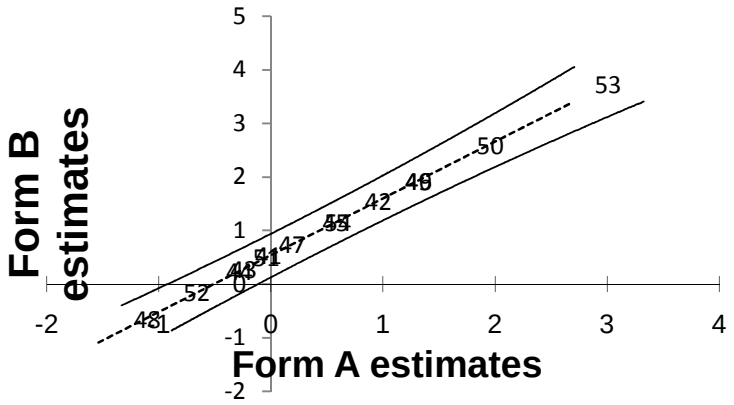


)L DD SIR FRPPR L P T DL

Concurrent common item equating design was used to place the items and persons from the two tests on the same scale so that the comparison of the abilities of the persons who had taken the two different test forms could become possible. In concurrent common item equating, after setting up the data in the fashion displayed in Figure 2, the entire dataset is calibrated in a single analysis. The anchor items take care of the difference in the difficulty of the two forms and bring the item and person estimates onto the same scale. Thus, the procedure allows the comparison of the difficulty estimates of the items in the two test forms and the ability estimates of the persons who have taken the two forms on a common scale. To analyze the data, one-parameter (1PL) IRT model or Rasch model as implemented in WINSTEPS (Linacre, 2009) version 3.69.1.10 was chosen.

Before running the equating analysis, the quality of the anchor items should be checked. As mentioned above, the difficulty estimates of the common items in two separate analyses should not be very different from each other; otherwise they cannot be used as common items. To make sure of this, a graphical check is usually carried out on the anchor item estimates in the two analyses. The difficulty estimates of the common items from separate calibrations of the two forms are cross-plotted on the *x* and y axes and quality control lines are drawn to check the closeness of the item parameter estimates (Baghaei, 2010; Wright & Stone, 1979). Items that fall outside the parallel quality control lines should be dropped from the analysis.

Figure 3 shows the quality control lines around the 14 common items in the two tests employed in this study. As Figure 3 shows, all the items function properly and cluster around the line of best fit and none of them is outside the parallel lines.



) L R S R RI FRPP R L P LIILF LP D ED R K R
forms

As was stated above, in concurrent equating there is no need to compute a shift constant. The test forms are linked by means of common items in a single data matrix as is shown in Figure 2 and the entire dataset is subjected to IRT analysis and the parameters of the two test forms are estimated in a single calibration. The derived parameter estimates are on a single scale.

Results

First, item separations, person separations, and reliabilities were computed for Form A and B. Whereas Table 1 shows the results of item separation and item reliability of the two forms, Table 2 depicts the results of person separation and reliability indices for the two forms.

DE P SD D LR D LDELL L LF RI) R P and B

	N	Item Separation	Reliability of Items
Form A	54	6.73	.98
Form B	54	4.58	.95
Combined Analysis	94	5.38	.97

Note: N is the number of items

As demonstrated in Tables 1 and 2, Form A has a person reliability of 0.92 and an item reliability of 0.98. The Root Mean Squared Error (RMSE) for the items is 0.22 and for the persons is 0.38. RMSE is the square root of the average of squared standard errors of measurement for all items and persons. The small values here show that the measurement has been precise. The data showed good fit to the Rasch model with only two items having infit mean square values outside the acceptable range of 0.7-1.3 (Bond & Fox, 2007).

Moreover, according to Tables 1 and 2, Form B has a person reliability of 0.81 and an item reliability of 0.95. The Root Mean Squared Error (RMSE) for the items is 0.37 and for the persons is 0.42. This form also fitted the Rasch model well. Three items had infit mean square values outside the 0.7-1.3 boundary. The common items showed good fit in both analyses.

DE 3 R SD D IR D IDELL L IF RI) R P D			
	N	Item Separation	Reliability of Items
Form A	160	3.31	.92
Form B	104	2.04	.81
Combined Analysis	264	2.82	.89

Note: N is the number of persons

Table 2 indicates that the combined analysis, when the two forms are linked by means of the 14 common items, yields a person reliability of 0.89 and Table 1 shows an item reliability of 0.97 in combined analysis. The RMSE for the items is 0.30 and for the persons is 0.39; five items out of the 94 items were misfits. The 14 common items all have good fit indices and cover a wide range of difficulty from -0.7 to 3.41 with a mean of 0.95 and a standard deviation of 1.06. The items which were used as anchor items all had acceptable fit indices and spanned over the difficulty continuum. Acceptable fit, high person and item reliability and separation indices, and small RMSEs in the combined analysis indicate that the equating procedure was successful.

Separate analyses of the two forms also indicated that the mean difficulty of the 14 common items in Form A was 0.54 logits and in Form B 1.11 logits. The mean of item difficulty estimates is usually centered on zero in most IRT software including WINSTEPS. This indicates that the average item on Form A (which had a difficulty of 0 logits in the Form A analysis) was 0.54 easier than the mean of the common items. The average item on Form

B (which had a difficulty of 0 logits in the Form B analysis) was 1.11 logits easier than the mean of the common items. Therefore, the average item on Form B was about 0.57 logits easier than the average item on Form A. 7KHUHIRUH) RUP %ZDV HDVLHUWKDQ) RUP $ &RPSDULQJ WKHWZRIRUPHUV¶5 DVFK ability estimates, which is the scaled score, and their raw scores with respect to the test form they have taken clearly shows this (Table 3).

DE D FR R FD FR FR LR DE

Person	Form	Raw score	Ability estimate
1	A	43	3.15
2	B	43	2.76
3	A	42	2.92
4	B	42	2.69
5	A	41	2.72
6	B	41	2.51
7	A	40	2.52
8	B	40	2.04
9	A	39	2.34
10	B	39	1.98
11	A	38	2.17
12	B	38	1.90
13	A	37	2.01
14	B	37	1.79
15	A	36	1.86
16	B	36	1.47
17	A	35	1.71
18	B	35	1.40
19	A	34	1.56
20	B	34	1.15
21	A	25	.38
22	B	25	.05
23	A	21	-.12
24	B	21	-.38
25	A	20	-.25
26	B	20	-.48
27	A	19	-.37
28	B	19	-.56
29	A	18	-.50
30	B	18	-.86

Table 3 shows part of the corresponding Rasch model ability estimates for different raw scores in the two forms. It should be mentioned that under the one-parameter Item Response Model (the Rasch Model), the persons with identical raw scores have identical ability estimates. However, this is not the case in the two and three parameter IRT models.

Table 3 shows that people who have identical raw scores but have taken different test forms do not have identical ability estimates. This is due to the differing difficulty of test forms, which is logical. Examines who have taken the harder form should have higher scaled scores than examines (with the same raw scores) who have taken the easier form. That is why it is essential to equate test scores. For example, as shown in Table 3, Person 1 took Form A and answered 43 items correctly; Person 2 took Form B and also answered 43 items correctly. Since Form A was harder than Form B, a raw score of 43 on Form A is worth 3.15 logits, while a raw score of 43 on Form B is worth 2.76 logits. A person who gets 43 items right on a hard test is more able than a person who gets 43 items right on an easier test. So, different difficulties in test forms are taken care of via equating.

Table 3 also shows that examinees with identical raw scores of 25 who took different forms of A and B had unequal abilities of 0.38 and 0.05 logits, respectively. Without equating examinees with identical raw scores who have taken different forms with varying difficulties are wrongly considered of equal ability. Note that Examinee 11, who took the harder Form A, with a raw score of 38, is more proficient than Examinee 8 with a raw score of 40 since Examinee 11 took the harder Form A.

Discussion and Conclusion

In order to better appreciate what unjust decisions which ignore test form difficulty might lead to, the two test forms analyzed above should be regarded hypothetically as different versions of the same high-stakes assessment system which is carried out twice a year to admit candidates to a very popular university program. Furthermore, the passing score for the test can be considered as 40. The pass mark in such examinations are constant and do not change across seasons or years. Examinee 11 who has taken Form A (the harder form) in spring, for instance, fails because s/he has scored 38. Examinee 8, who has taken the test in fall, when by chance the easier Form B was given, passes because s/he has scored 40. However, Table 3 shows that after equating, that is, after scores are adjusted for form difficulty, Examinee 11 with a raw score of 38 and a scaled score of 2.17 logits, turns out to be more proficient than Examinee 8, with a raw score of 40 and a scaled score of 2.04 logits. Therefore, scores from different test forms are not comparable unless they are equated and adjusted for form difficulty.

One of the main reasons for developing standardized tests is to measure FDQGLGDWHV¶DELOLWLHV REMHFWLYHO\ DQG IDLUO\ 6FRUHV IURP VWDQGDUGL]HG WHVWV DUH

RIWHQXVHG WR P DNH LP SRUWDQW GHFLVLRQV DERXW LQGLYLGXDOV¶ OLYHV : LWK D ZURQJ decision, a candidate may be excluded from the academic program of his/her interest or the practice of his/her favorite profession. Furthermore, important decisions about education policies and curricula are made on the basis of standardized tests.

Due to the importance of the results of standardized tests, every effort should be made to provide a fair measurement of the abilities of interest. The lack of equating and reporting raw scores across numerous forms which are used in multiple runs of an assessment over years and comparing FDQGLGDWHV¶ UDZ VFRUHV ZLWK D FXVWRPDU\ FRUH ZKLFK LV D UDZ VFRUH UHVXOW LQ QRQVHQVLFDO GHFLVLRQV DQG P HDVXUHP HQWDO DQ XQIDLU HYDOXDWLRQ RI H[DP LQHHV¶ VNLOO (Cook & Eignore, 1991).

The concept is directly related to test fairness and social aspects of validity. Messick (1989) defines validity as the appropriateness of inferences and decisions made on the basis of test scores. In other words, it is concerned with whether exclusion or inclusion of a particular candidate is appropriate and beneficial to the society and the institution which uses the test results. More recently, fairness and social responsibilities of test GHYHORSHUV KDV EHHQ GLVF XVVHG WR WKH YDOLGLW\ LVVXH ³9 DOLGLW\ WKHRUL P SOLHV considerations of social responsibility, both to the candidate (*protecting him or her against unfair exclusion*) DQG WR WKH UHFHLYLQJ LQVWLWXWLRQ< ´ 0 F1 DP DUD & 5 RHYHU S DXWKRUV¶ HP SKDVLV

In order to guarantee validity and fairness, testing bodies have to maintain the same standards from season to season and from year to year. As was mentioned earlier, standardized testing implies fair and stable measures and units which do not put some examinees at a disadvantage. The time of the year or the form of a test an examinee takes should not determine pass and fail results.

The direct implication of the argument and example delineated in this paper is for high-stakes assessments in Iran where in some instances, test score equating might not be implemented and thus unfair pass/fail decisions might be made. Thus, the researcher hopes to draw the attention of high-stakes test-developers and administrators in the country to the importance of test score equating as they are involved in a highly sensitive and critical decision-making. The ultimate goal is to have fair decisions about admissions, certificate issuance, and graduation. Equating is a process which needs to be implemented in all standardized high-stakes tests if testing authorities and educational policymakers want to avoid unjust decisions.

Received on March 2, 2010
Accepted on May 15, 2010

The Author

Purya Baghaei is Assistant Professor of Applied Linguistics at Islamic Azad University, Mashad Branch where he teaches language testing, research methods in TEFL, and language teaching methods. He holds a PhD in applied linguistics from Klagenfurt University, Austria and his major research interests are the application of Rasch models in language testing. He was a former data analyst and validation officer at the Klagenfurt Language Testing Center and has published numerous articles on the applications of Rasch models in standard setting and validation.

puryabaghaei@gmail.com

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thomdike (Ed.), *Educational measurement* (pp. 508-600). Washington DC: American Council on Education.

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträgeaus der aktuellen Forschung/The C-Test: Contributions from Current Research* (pp. 101-112). Frankfurt: Lang.

Baghaei, P. (2009). *Understanding the Rasch model*. Mashad: Mashad Islamic Azad University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: LEA.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practices*, *10,* 191-199.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage Publications.

Kim, J. S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*(3), 255-270.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating*. New York: Springer.

Linacre, J. M. (2007). *A HU J L H WR : ,16 (6 0,1,6 (5D F PR HO computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2009). *WINSTEPS* >&RPSXWHU6RIWDUH YHUVLRQ @&KLFDJR IL: winsteps.com.

Livingston, S. (2004). *Equating test scores (without IRT)*. Princeton: Educational Testing Service. Retrieved July 5, 2010, from www.ets.org/Media/Research/pdf/,9,1*6721 SGI

/ XSSHMFX 6 9LUWDCHTXDLQJ *Rasch Measurement Transactions, 19*(3), 10-25.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Boston, MA: Blackwell.

0 HVVLFN 6 9DOLG ,Q5 / / LQQ (G *Educational measurement* (3rd ed.) (pp.13-103). NY: American Council on Education & Macmillan.

Philips, D. (1996). *Longman preparation course for the TOEFL test*. (2nd ed.). New York: Addison-Wesley.

Saida, C., & Hattori, T. (2008). Post-hoc IRT equating of previously administered English tests for comparison of test scores. *Language Testing, 25*(2), 187-210.

Shin, S. H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation*, *14*(1). Retrieved September 17, 2010, from www.pareonline.net/getvn.asp?v=14&n=1

Skaggs, G., & Wolf, E. W. (2010). Equating designs and procedures used in Rasch scaling. *Journal of Applied Measurement, 11*(2), 182-195.

Sharpe, P. J. (2004). *How to prepare for the TOEFL* (11th ed.). Hauppauge, NY: %DUURQV (GXFDWLRQDO 6HULHV

Smith, R. M., & Kramer, G. A. (1992). A comparison of two methods of equating in the Rasch model. *Educational and Psychological Measurement, 52*, 835-846.

Wolfe, E. W., & Chiu, C. W. T. (1997). *Measuring change over time with a Rasch rating scale model*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL, March, 24-28. ERIC Documents.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA.