Practical Assessment, Research, and Evaluation

Volume 10 *Volume 10, 2005* Article 4

2005

Test Equating by Common Items and Common Subjects: Concepts and Applications

Chong Ho Yu

Sharon E. Osborn-Popp

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 10 Number 4, May 2005 **ISSN 1531-7714**

Test Equating by Common Items and Common Subjects: Concepts and Applications

Chong Ho Yu, Aries Technology
Sharon E. Osborn Popp, Arizona State University

Since the invention of z-scores (standardized scores), comparison among different tests has been widely conducted by test developers, instructors, educational researchers, and psychometricians. Equating, calibration, and moderation are terms used to describe broad levels of possible comparison among educational assessments (Dorans, 2004; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Linn, 1993; Mislevy, 1992). Equating is at one end of the linking continuum, involving the most stringent requirements of equivalence among the assessments and examinee populations to be linked, and compares tests that measure the same construct and have been designed to be equivalent. Less equivalent conditions involve calibration, which compares tests that measure the same construct but vary in design or difficulty, and moderation, which compares tests that measure different constructs. Psychometric approaches to linking assessments include linear equating, equipercentile equating, and item response theory (IRT). This article is a practical guide to conducting IRT test equating in two different scenarios:

1. **Common Item Equating:** In this scenario, subjects take tests that are composed of linking items that are common to all forms and non-linking items that are unique to each form. Two types of common item equating are illustrated here: alternate form equating (where common and unique items are analyzed simultaneously) and across sample equating (where different sets of unique items are analyzed separately based on previously calibrated anchor items).
2. **Common Subject Equating:** In this scenario, the same subjects take different tests. Tests are designed to the same specifications, but share no common items.

In all test equating studies, the central question remains the same: Could the scores yielded from different tests measuring the same construct be comparable to each other? Consider alternate test forms, which are widely used as a countermeasure against the cheating problem. According to IRT, alternate forms should be balanced in terms of equivalent test information functions (TIF). To be specific, an examinee who takes Form A should not be more or less advantaged than one who takes Form B or Form C. By using linking items across all forms, item calibration and parameter estimation can be grounded on a common base. Another practical example is when an exam is almost completely revamped in order to retire outdated materials, it is advisable to keep several items from the previous exam as anchor items. In this setup, it is unlikely that subjects who have obtained the license, certification, or qualification by passing the old test would take the new version of the exam. Nevertheless, even if different examinees take the old and the new exams, the parameters of those anchor items could be used as fixed parameters in the new test analysis, and thus the scores yielded from the new test are said to be comparable to that from the old test when equating using an IRT approach.

Test equating by common subjects may be employed in the situation where a drill-and-practice exam and the "real" exam are administered to the same group of subjects. The purpose of "drill-and-practice" is to give examines an orientation to the real exam, where no items in the former should resemble the latter, otherwise, the students could remember the items and seriously contaminate the real exam result. Even though no items in two tests are the same or look alike, the question of commensurability between exams could be directly answered by checking the strength of correlation between two sets of theta (ability estimates) yielded from the two tests.

According to Kolen and Brennan (2004), there are at least four aspects of "commonalities" between two tests that should be taken into consideration for test equating, scaling, and linking:

1. **Inferences**: To what extent are scores for the two tests or two alternate forms used to draw similar types of inferences?
2. **Constructs**: To what extent do the two tests measure the same construct? On some occasions, the tests may share a common construct yet also have unique constructs.
3. **Populations**: To what extent are the two tests designed to be used with the same populations? In some contexts two tests might measure the same construct but not be appropriate for the same population. In other contexts, two tests might measure the same construct yet the results can be generalized across different populations.
4. **Measurement characteristics or conditions**: To what extent do the two tests share common measurement characteristics or conditions, such as test length, test format, item type, administration procedures, etc.

The authors of this article have applied this categorization onto our examples of test equating. The degree of similarity for different types of test equating is summarized in Table 1:

Table 1 Degree of similarity for different types of test equating

	Inference	**Constructs**	**Populations**	**Measurement characteristics**
Common item equating (alternate form)	Same	Same	Same	Same
Common item equating (across sample)	Similar or dissimilar	Same	Same	Similar or dissimilar
Common subject equating	Same	Same	Same	Similar or dissimilar

For common item equating with alternate forms, all four aspects should have a high degree of resemblance. Alternate forms are considered the same test being expressed in different versions. Thus, even if different subjects take different forms, the constructs to be measured and the inferences to be yielded should remain the same. Also, since all subjects take the test in the same setting, there is no reason that the measurement characteristics should vary from form to form.

For common item equating across samples, it is important to note that even though the items are supposed to measure the same construct for the same population, their inferences might be different from each other, especially when the tests have different objectives. For example, items pertaining to TCP/IP in a Microsoft Windows Server exam and in a Microsoft Windows Network exam could

DOI: https://doi.org/10.7275/68dy-z131

lead to different inferences. In this form of test equating, the focus centers around item attributes rather than subject ability, and thus different interpretations of subject scores in different tests are allowed. Because common item equating across samples is implemented in different settings, it is expected that their measurement conditions could be similar or dissimilar.

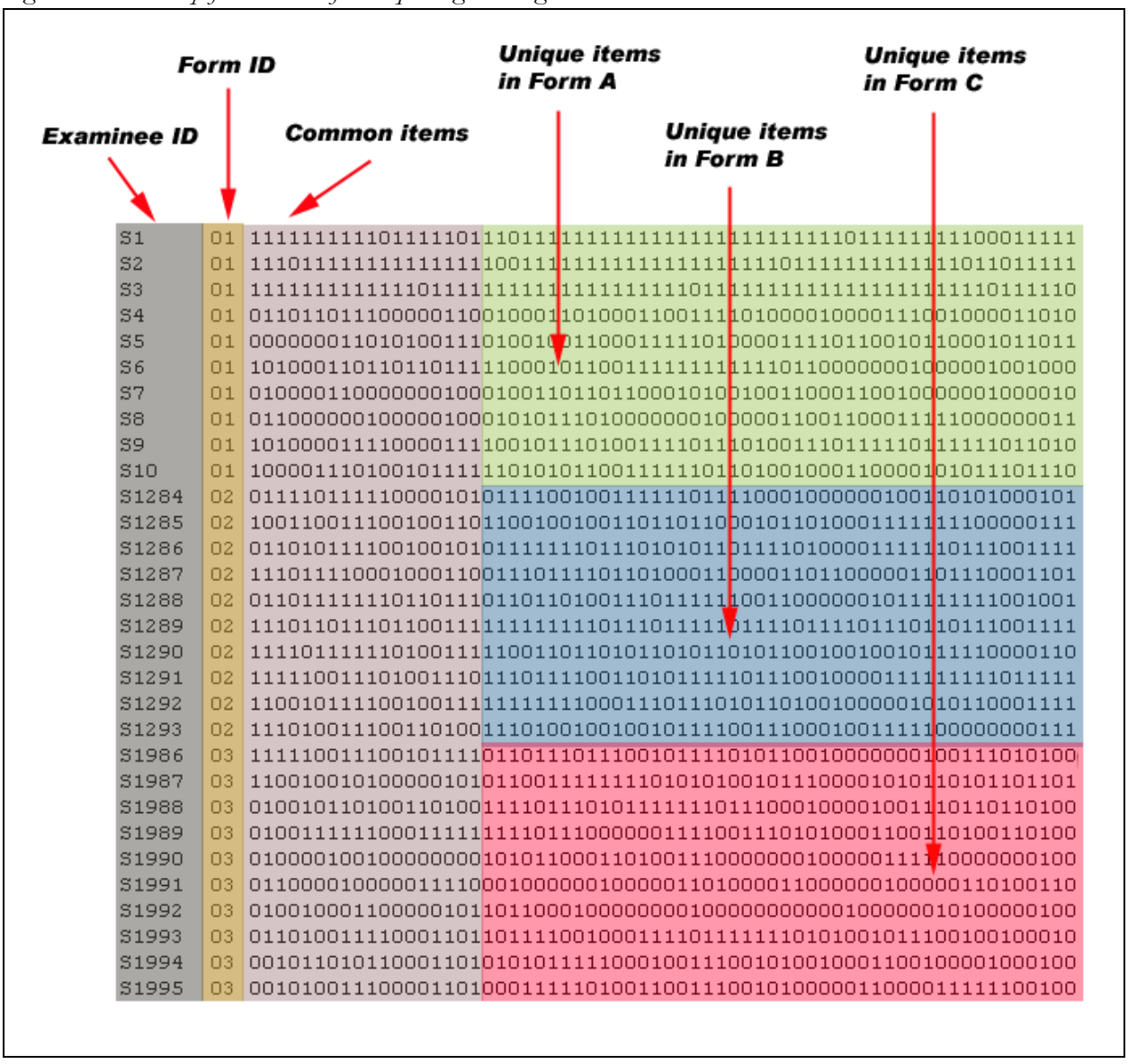
For common subject equating, the two tests are supposed to measure the same construct. But unlike common item equating across samples, the subject abilities in terms of logit are compared in this type of test equating; it is assumed that the two sets of scores are comparable and the same type of inferences can be made. However, since the two tests are administered in different settings, variations of measurement conditions are expected.

The procedures of how to conduct test equating in different scenarios are explained as follows:

COMMON ITEM EQUATING: EXAMPLE 1 USING BILOG-MG

The software package for alternate form equating is Bilog-MG (Zimowski, Muraki, Mislevy, & Bock , 2003), in which "MG" stands for "multiple group." To run an alternate form analysis, the researcher must set up the data by putting all common items at the beginning and the unique items at the end.

Figure 1. Data setup for alternate form equating in Bilog-MG



The following example is based upon the dataset named "Mathematical competence test" collected in a large metropolitan school district (see Figure 1), there are eighteen common items across all forms while each form contains forty-six non-linking items. It is important to point out that although in this data format the forty-six non-linking items occupy the same chunk of positions (Column 19-64), their Bilog item ID must be assigned differently in each form.

The code in Figure 2 is an example of how to assign Bilog ID to linking and non-linking items in an alternate-form setting. Each step will be explained.

Figure 2. Program setup for alternate form equating in Bilog-MG

```
>COMMENTS ;
    Form equating of Mathematical Competence Test;
>GLOBAL  NPArm=1, DFName='bilog.txt', SAVe, logistic;
>SAVE    PARM='bilog.PAR',
         SCOre='bilog.SCO',
         TSTat='bilog.TST';
>LENGTH  NITems= 156;
>INPUT   NTOT=156, NFMt=1,NIDCHAR=10,nalt=4,NFORM=3,SAMPLE=1995;
>ITEM    iname = (1(1)156);
>TEST    tname = bilog;
>Form1   LENGTH=64, INUM=(1(1)18,19(1)64);
>Form2   LENGTH=64, INUM=(1(1)18,65(1)110);
>Form3   LENGTH=64, INUM=(1(1)18,111(1)156);
  (10A1,I2, 64A1)
>CALIB   EMPIRICAL,TPRIOR,CRIT=0.01,PLOT=1.0;
>SCORE   METhod=2, INFo=1, noprint;
```

GLOBAL: "**NPARM**" is the abbreviation for "number of parameters." Bilog-MG can run one-, two-, or three-parameter modeling. The one-parameter model estimates item difficulty, the two-parameter-model estimates item difficulty and item discrimination, and the three-parameter model also attempts to estimate the effect of chance with the addition of a pseudo-guessing parameter. For further information on common IRT models, as well as their advantages and limitations, see Hambleton (1993) and Andrich (1988). On some occasions the researcher may want to compare two sets of parameters yielded from Bilog and Winsteps, which is a Rasch-model (one-parameter model) software application. In Bilog, users are allowed to choose between the logistic metric and the normal metric. If the logistic metric is used, certain rescaling schemes are needed in Winsteps in order to facilitate further comparison. With some minor adjustment, such as using a multiplier as D=1.7, the logistic and the normal metrics are virtually indistinguishable (McDonald, 1999). Please consult the Appendix at the end to learn the differences between Bilog-MG and Winsteps, as well as the research paper presented by Pomplun, Omar, and Custer (2002).

SAVE: Users can choose to save the item parameter file (filename.PAR) and the examinee theta score file (filename.SCO) in Bilog. These files are equivalent to the question file (filename.que) and person file (filename.per) in Winsteps. Since the score file will be output, in **SCORE** definition, which is the last line in the program, **NOPRINT** should be used to suppress printing all theta estimates on the screen. You can imagine that the screen will look extremely messy when there are 1,500 subjects. The parameter file is equivalent to the Phase Two output in Bilog. However, the former does not carry the Chi-Squared statistics, which are essential for examining the fitness of the items. In short, the parameter file should not be treated as a replacement of Bilog Phase Two output.

LENGTH and INPUT: Users should specify the total number of items, including the linking and

non-linking items. It is crucial to note that the total number is the total in all forms, not in one form. In this example, each form has 18 common items and 46 non-linking items. Users may mistakenly put down 64 (18+46) as the total. Actually, the total should be 156 (18+46+46+46). The length of examinee ID should be specified in **IDCHAR**, the number of alternatives in multiple-choice items in **NALT** and the number of forms in **FORM**. By default, Bilog-MG randomly subsets 1,000 subjects when the sample size exceeds 1,000. If you want to utilize all subjects, **SAMPLE** must be defined.

Form 1 to 3: As mentioned before, although the non-linking items are situated in the same positions in the raw data file, In the Bilog program they must be treated as different items and thus their ID must be re-assigned.

CALIB and **SCORE** are concerned with calibration and estimation algorithms for item parameters and examinee theta scores, which are beyond the scope of this practical guide. Please consult the Bilog manual for more information.

There is a major drawback in common item equating: the standard error of item parameters yielded from non-linking items are inevitably higher than that of linking items due to the difference in sample size. In this example, all linking items are answered by 1,995 subjects whereas non-linking items in the three forms are answered by about 650 subjects, respectively.

Figure 3 is a partial screen shot of Phase 2 output in Bilog. Item 1-3 are extracted from the linking pool whereas item 154-156 are from the non-linking counterpart. Note that the standard errors of the first item group are smaller than that in the second group.

Figure 3. Phase 2 of Bilog output

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	DISPERSN S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
Linking items							
ITEM0001	1.055	0.939	-1.124	1.065	0.000	45.0	9.0
	0.056*	0.007*	0.059*	0.009*	0.000*	(0.0000)	
ITEM0002	1.647	0.939	-1.754	1.065	0.000	60.0	9.0
	0.057*	0.007*	0.061*	0.009*	0.000*	(0.0000)	
ITEM0003	0.859	0.939	-0.915	1.065	0.000	30.7	9.0
	0.053*	0.007*	0.057*	0.009*	0.000*	(0.0004)	
Non-linking items							
ITEM0154	1.017	0.939	-1.083	1.065	0.000	8.3	9.0
	0.091*	0.007*	0.097*	0.009*	0.000*	(0.5057)	
ITEM0155	0.258	0.939	-0.275	1.065	0.000	25.3	9.0
	0.086*	0.007*	0.092*	0.009*	0.000*	(0.0028)	
ITEM0156	-0.318	0.939	0.339	1.065	0.000	5.2	9.0
	0.085*	0.007*	0.091*	0.009*	0.000*	(0.8220)	

There is no quick fix for this shortcoming. Researchers are encouraged to use as many subjects as possible for common item equating. For instance, 900 subjects seem to be a fairly adequate sample size for one-parameter IRT modeling. When those 900 subjects are spread across three forms, each non-linking item has only 300 subjects. It may or may not be adequate, depending upon the ratio between the number of subject and the number of items, as well as the factor structure of the test.

COMMON ITEM EQUATING: EXAMPLE 2 USING WINSTEPS

Across-sample item anchoring can be done in Winsteps (Rasch Measurement Software and Publications, 2002). Although no specific data format set up is required for this type of item anchoring, it is important to point out that common items should be placed in both tests to be equated, as illustrated in Figure 4. Needless to say, good items must be used for anchoring. If an unstable or imprecise parameter based upon a poorly-written item is used, the analysis will be worse than the one without test equating. Researchers are advised to take the following references into account while selecting anchor items from previous test: 1. Item parameter stability, 2. Item exposure, and 3. Item type. Each criterion will be discussed next:

1. **Item parameter stability**: In classical test theory, reliability in terms of stability is a temporal concept. This concept can be well applied into IRT. Simply put, if a parameter tends to be almost invariant across time, it should be a good candidate to be an anchor item.

Take the line plot in Figure 5 as an example. After Item 1, Item 2, and Item 3 were released in the first two months, they seem to be equally good because their difficulty parameters are close to zero, which means an examinee with average ability has approximately 50% chances to answer the item correctly. However, the plot indicates that several months later, the difficulty parameter of Item 2 fluctuates considerably while Item 3 tends to be easier and easier over time. Clearly, Item 1 is a better choice for item anchoring.

2. **Item exposure**: Item exposure may be related to decreasing parameter value. To be specific, even if an item is well written, the frequent use of the item may lead to the consequence that examinees are familiar with the pattern of similar items or that earlier examinees share the item content with later examinees. The following plot demonstrates this relationship. The y-axis is the item exposure in terms of the accumulated frequency of the item used in all exams while the x-axis is the item parameter of Item 3 from January to December (see Figure 6). If the researcher has not collected sufficient data to make temporal-based plots, it is recommended that an item with less exposure be used for item anchoring. For instance, if Item A and Item B have the same threshold parameter, but the former has been seen by 500 examinees while the latter has been seen by 1,000, Item A should be the first choice.

Figure 4. Linking items across samples

Figure 5. Plot for detecting item parameter stability

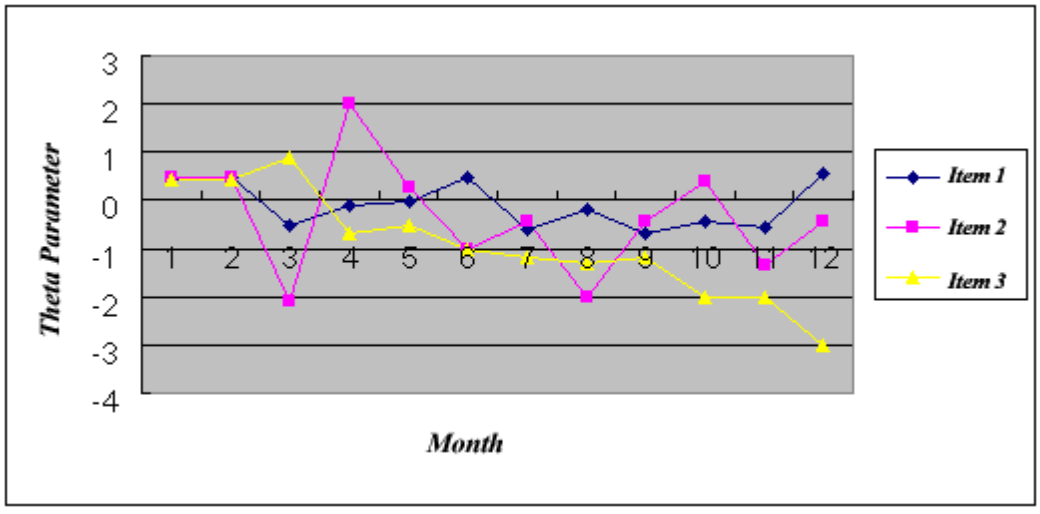
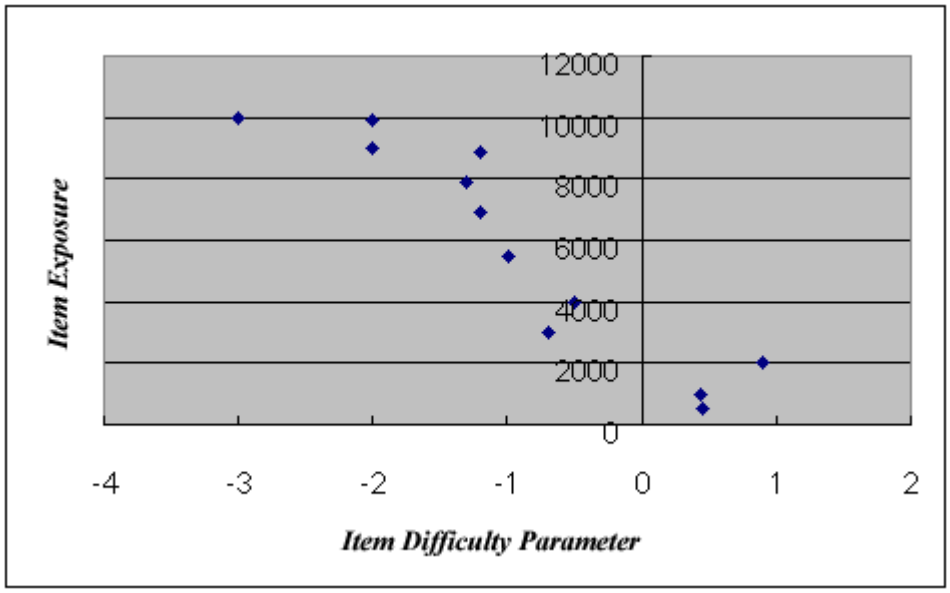


Figure 6. Scatterplot of item difficulty parameter and item exposure



3. **Item type**: Certain item types are more suitable than others as anchor items. For example, Multiple-Choice-Multiple-Answer (MCMA) items may be preferred to Multiple-Choice-Single-Answer (MCSA) items because in the latter, by default, the examinee has 1/5 or 1/4 chances to get the right answer by guessing, depending upon the number of options, while in the former the number of possible combination are too many (e.g. ABD, BCD, ACD, ABE, BCE etc.) for examinees to guess the right answers. In MCMA the correct answer is composed of a combination of several options. Thus, the difficulty parameter yielded from this type of question tends to be a true parameter. Keep the intended examinee pool in mind, however. Complex multiple-choice items like MCMA may not always be the choice (e.g., for elementary level assessments). On some occasions complex and wordy MCMA items may be a test of analytical ability rather than mastery of the subject matter. MCSA items that have well written distractors can also discourage guessing behavior, and will always be satisfactory choices for anchoring.

Some argue that case-study or simulation-based items are even better than MCMA items because examinees could memorize the item content and thus MCMA and MCSA items are vulnerable to item exposure. On the other hand, the text of case-study items are usually longer than that of MCMA and MCSA items, and thus passing item information from earlier examinees to later examinees is difficult, if not impossible. In addition, case-study items demand applications of knowledge whereas simulation-based items, which are common in computer-related tests, require procedural knowledge. For both item types, memorization of factual content is not helpful for examinees to boost performance and thus they are less subject to the detrimental effect resulted from item exposure.

Some kinds of items are inherently unsuitable to be anchor items. Testlet items, which are grouped together based on a common passage or other stimulus, may be problematic. First, since testlet items may be highly correlated, the assumption of local independence, which is essential to IRT, may be violated. Second, when the entire chunk of testlet items are used as anchors, it is expected that the quality of some items may be better than others, and as a result the researcher is forced to adopt certain unstable or imprecise item parameter as anchors. In addition to testlet items, partial-credit items should also be avoided in item anchoring. Partial-credit items involve a step function, in which the step of obtaining a higher score is more difficult than the step of obtaining a lower score. When an overall difficulty parameter across all steps of partial-credit item is used as an anchor, needless to say, there is a substantial loss of information.

In summary, MCMA, well-developed MCSA, case-study, and simulation-based items are recommended as anchor items while testlet and partial-credit items are not recommended for use in test equating. Maintaining test specifications should be the primary goal in choosing anchor items, so sustained attention to item content, item difficulty, and cognitive demand of items remains essential in any assessment development.

After anchor items are selected, several program parameters must be carefully specified as shown in the following (see Figure 7).

Figure 7. Program setup for across-sample item anchoring in Winsteps

```
&INST
TITLE='Winsteps program for test equating'
NI=64        ; Number of items
ITEM1=13     ; Position of where the first item begins
CODES=01     ; Valid answers
GROUPS=0     ; If there is no multiple group, use "0"
NAME1=1      ; The position of where the subject ID starts
NAMELEN=10   ; Length of Subject id
; Bilog sets the person mean to zero by default whereas Winsteps sets the
item mean to zero by default.
; We want to make Winsteps output as close to Bilog's as possible and thus
the person mean is set to zero.
UPMEAN=0
; Bilog uses probit when Winsteps use logit. Rescaling: Probits * 1.7 =
logits
USCALE = 0.59
; output file names
IFILE=math.que  ; que is the question file for item parameters
PFILE=math.per  ; per is the person file for person theta
; Prefix for person and item. They are arbitrary
PERSON=S
ITEM=I
TABLES=11111111111111111111  ; Return all tables
DATA=matha.dat    ; Name of the raw data file
; IA means Item Anchoring.
; The number in the first column is the position of the items, not the Item
ID.
; The numbers in the second column are the item difficulty parameter.
IAFILE=*
2 -.57
4 -1.15
7 -.34
9 1.67
12 -.36
14 .83
17 -.07
25 -1.40
29 -.12
31 -1.06
36 .57
39 1.69
44 .45
46 1.58
50 1.00
55 -2.06
58 -.27
63 -.41
*
; Enter Item ID in the following
&END
```

Figure 7. Program setup for across-sample item anchoring in Winsteps

```
Item1
Item2
...
Item62
END NAMES
```

Most comments inside the preceding code are self-explanatory, nonetheless, some require further explanation. Note that in Winsteps the user can either set the person mean or the item mean to zero. For test equating the person mean rather than the item mean should be set. The rationale is straight-forward: when the parameters of anchor items are fixed, all other unknown item parameters should be adjusted along these known parameters. In this case, there is no need to force the item calibration centering around the zero mean as another anchor. Further, since in Bilog the person mean is set to zero by default, adopting the same setting can facilitate comparison across Bilog and Winsteps, if necessary. In addition, the user can either prepare a separate item anchor (IA) file to specify the fixed parameters or simply enter the IA parameters into the control file. In this example the latter approach is adopted.

Figure 8 shows a partial screenshot of the Winsteps question file. The first column is the position of items, the second is the difficulty parameter, also known as measure in Winteps. The third is the flag that indicates which items are anchors. Items in position 24, 26 and 28 are anchors. Thus, all non-anchor items are flagged as "1" while these three items are flagged as "2", which mean the data for these items are ignored and fixed parameters supplied by the program are imported.

Figure 8. Winsteps question file output

MEASURE	STTS	COUNT	SCORE	ERROR	IN.MSQ	IN.ZSTD	OUT.MS	OUT.ZSTD	DISPL	CORR	WEIGHT	G	M
					Anchoring items								
-.57	2	636.0	479.0	.10	.88	-2.59	.91	-1.21	-.11	.43	1.00	1	R
-1.15	2	636.0	541.0	.11	.92	-1.24	.75	-2.40	-.25	.30	1.00	1	R
-.34	2	636.0	426.0	.09	1.02	.43	1.06	.89	.14	.40	1.00	1	R
					Non-anchoring items								
.19	1	636.0	378.0	.09	1.12	3.31	1.16	3.06	.00	.29	1.00	1	R
-1.74	1	636.0	564.0	.13	.99	-.08	1.15	.88	.00	.26	1.00	1	R
1.89	1	636.0	166.0	.10	1.24	4.57	1.32	3.64	.00	.15	1.00	1	R

TEST EQUATING BY SUBJECTS

Common Subject equating can be done in Winsteps, Bilog, or any IRT software. The same subjects are required to take both exams to be equated. Afterwards, separate IRT analyses are performed on the two exams to estimate subject ability in terms of theta scores. Next, the two sets of scores along with the 95% confidence band based upon the standard errors are plotted in a scattergram, as shown in Figure 9 (Bond & Fox, 2001). The 95% confidence band provides a means to evaluate the extent to which the two tests are measuring the same construct within a reasonable degree of measurement error.

Thirty subjects who took two computer competency exams in a large metropolitan area are used in this example. The steps for common subject equating is illustrated as follows:

1. Run Winsteps or Bilog to obtain the ability estimates and error estimates for examinees on each test. In Bilog, Phase three output or the Score Output (sco) contains the ability estimates while in Winsteps, usually the Person file (per) carries this information. It is important to emphasize that the user IDs on both exams must be consistent for merging. Figure 9 is an excerpt of the Bilog Phase 3 output, which has clear labels. However, the pipe symbol (|) in the table makes importing very difficult, and thus it is advisable to use the Score Output, in which control characters such as the pipe are omitted. Note that the Score Output produced by Bilog will not have header labels. The fields that will be used for equating are shaded in the figures. Figure 10 is an excerpt of Winsteps person file output.

Figure 9. Excerpt of Bilog Phase 3 output

GROUP WEIGHT	SUBJECT IDENTIFICATION SUBTEST	TRIED	RIGHT	PERCENT	ABILITY	S.E.	MARGINAL PROB
1 S1				|		|	
1.00	BILOG	64	57	0.8906 |	1.7876	0.4559 |	0.000000
1 S2				|		|	
1.00	BILOG	64	58	0.9062 |	1.9770	0.4334 |	0.000000
1 S3				|		|	
1.00	BILOG	64	60	0.9375 |	2.2702	0.3823 |	0.000000

Figure 10. Excerpt of Winsteps person file output

;ENTRY	MEASURE	STTS	COUNT	SCORE	ERROR	IN.MSQ	IN.ZSTD	OUT.MS	OUT.ZSTD	DISPL	CORR	WEIGHT	NAME
1	2.52	1	64.0	57.0	.42	1.14	.46	1.08	.14	.00	.35	1.00	3334321
2	2.71	1	64.0	58.0	.45	.96	-.14	1.75	1.00	.00	.43	1.00	3314221
3	3.18	1	64.0	60.0	.53	1.04	.09	.79	-.28	.00	.44	1.00	3334121

2. If the exams have both common and unique subjects, merge the subjects by user ID and retain only common subjects in a file.
3. Import this file into a spreadsheet or a statistics program.
4. Compute the mean of the ability estimates for each exam.
5. Compute the difference between the two means.
6. Adjust one of the exam ability estimates by the mean difference (add the difference to each estimate).
7. Compute the paired 95% confidence bands using the error estimates:

Lower bound=

$$(A+B)/2-\sqrt{C^2+D^2}$$

Upper bound=

$$(A+B)/2+\sqrt{C^2+D^2}$$

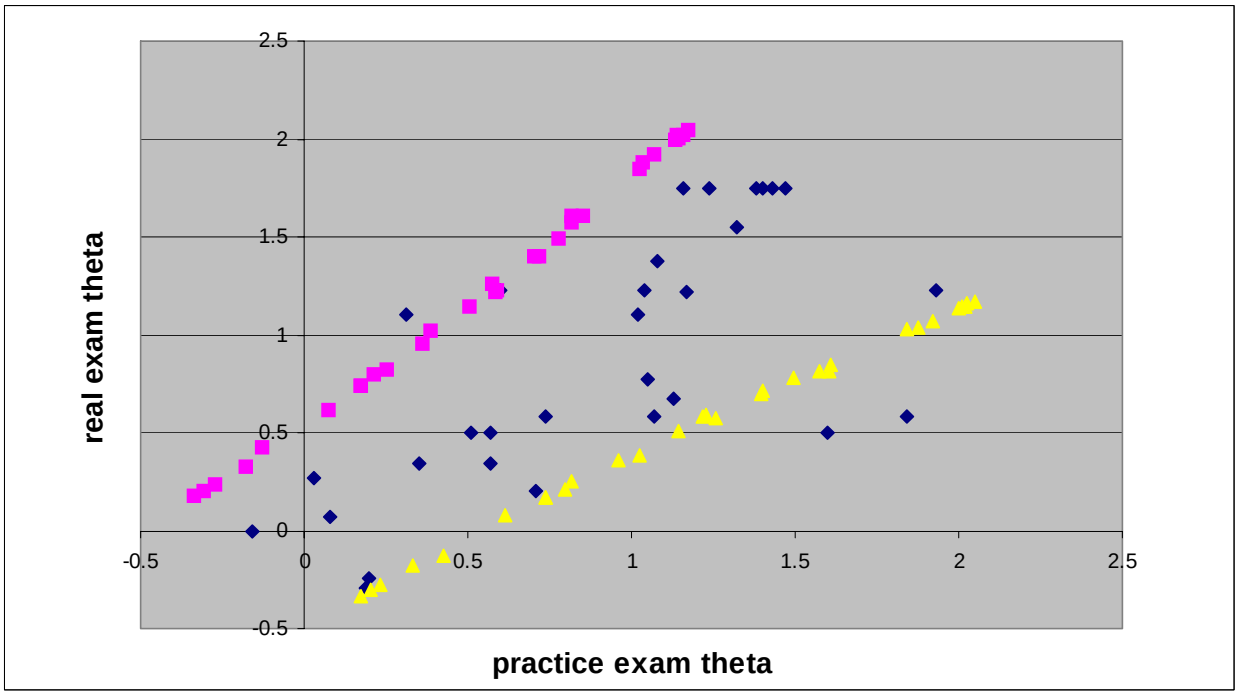
Where:

A = ability estimate in exam 1
B = adjusted ability estimate in exam 2
C = error estimates in exam 1
D = error estimates in exam 2

8. Plot the preceding information in a scatterplot. Overlay the following three pairs of variables in the scatterplot: 1) exam 1 ability estimates and exam 2 ability estimates, 2) lower bound values and upper bound values, and the pair of bound values reversing axes, 3) upper bound values and lower bound values. Bond and Fox (2001) made an illustration of the plotting this information in Excel. Nonetheless, the preceding procedures can also be implemented in other spreadsheet and statistics software applications.

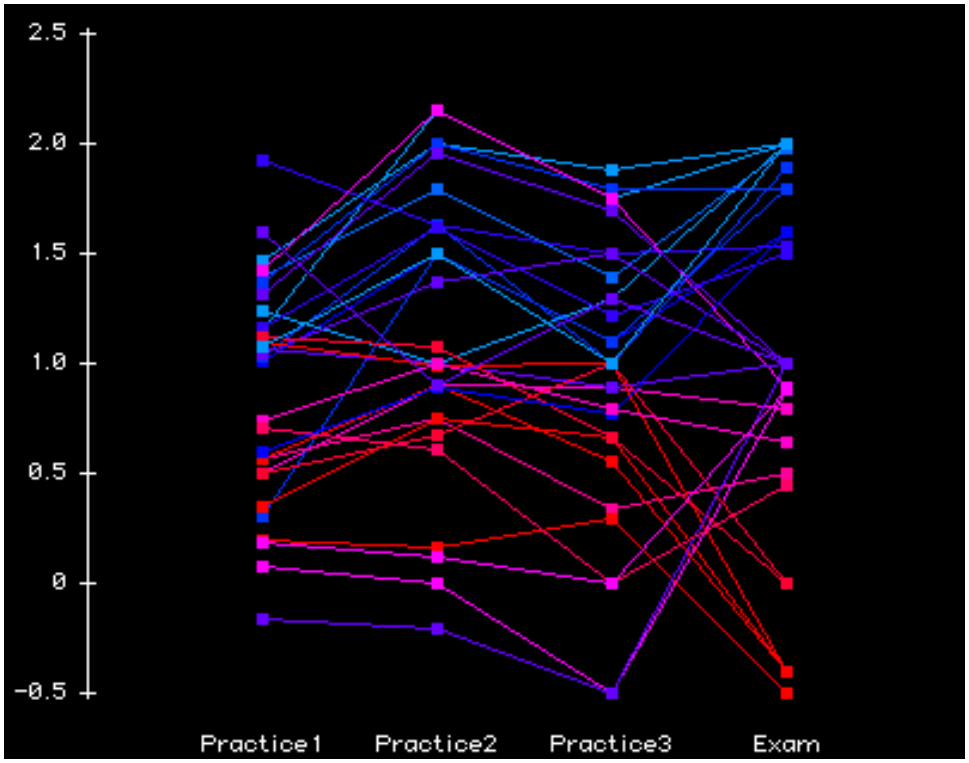
In this example, three observations are located outside the confidence band, which is about 5% of all observations. Note that two are over-estimated and one is under-estimated. Nevertheless, theta estimation is fairly consistent across the practice and real exams, and thus it is concluded that two exams seem to measure the same construct; theta scores obtained from one are comparable to that from the other.

Figure11. Scatterplot of practice exam theta by real exam theta with confidence intervals



This simple yet useful test equating method can be extended beyond two exams. In the following example four exams are equated by using the same subjects (see Figure 12). A scatterplot is useful in portraying two-dimensional data whereas a spin plot or a Trellis plot is applicable to three-dimensional data. For four-dimensional data, a parallel-coordinate plot, which can be implemented in DataDesk, SAS, Excel, and many other software applications, is strongly recommended.

Figure 12 Parallel-coordinate plot of four exams



In the current example that consists of three practice exams and one real exam, theta estimates are connected by line. While theta estimates from practice exam one to three do not seem to be varying substantively, this pattern of theta invariance falls apart in the actual exam. Theoretically speaking, theta estimation is independent from item attributes while item parameter estimation is independent from examinee ability given that the factor structure of all exams has one dimension only and they all measure the same construct. However, the pattern of these exams suggests that while the practice exam group may contain the same construct, it is likely that the real exam does not share the same construct as the practice exam group.

After we confirm that thetas yielded from the same subjects in two exams are fairly consistent, we can make even bolder inferences by equating items in two exams using the examinee thetas as the common point of reference. In psychometric software applications such as Winsteps and RUMM, the researcher can obtain the Item-Person Map (IPM), in which item difficulty and examinee ability values are expressed into a common scale, Logits, and thus both measures can be presented simultaneously. Figure 13a and Figure 13b are IPMs from two different exams with some common subjects but without common items.

Figure 13a Item-Person Map of Exam 1

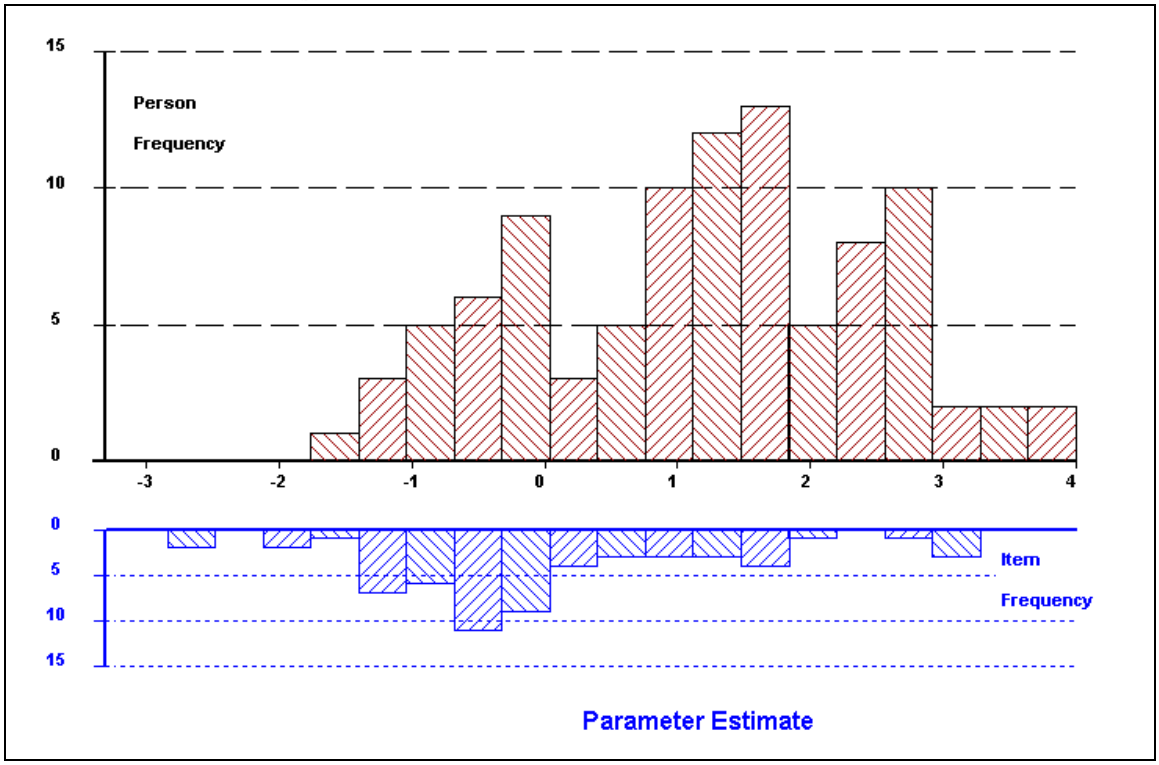
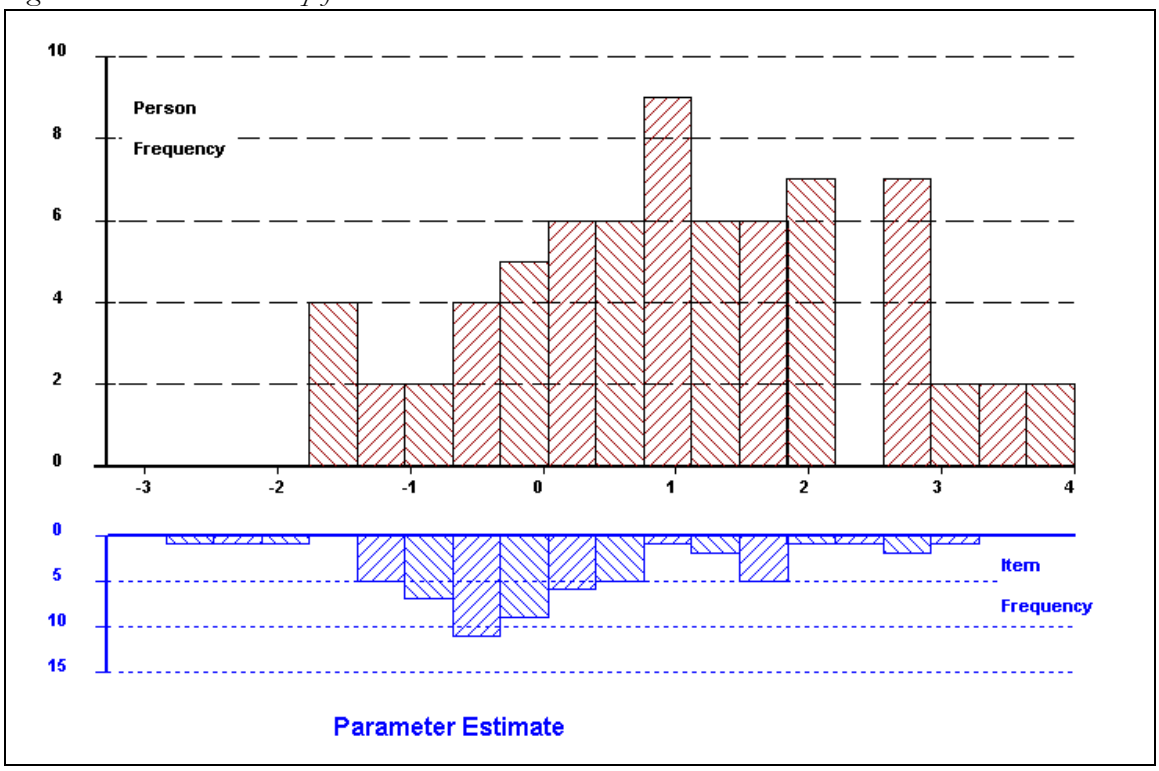
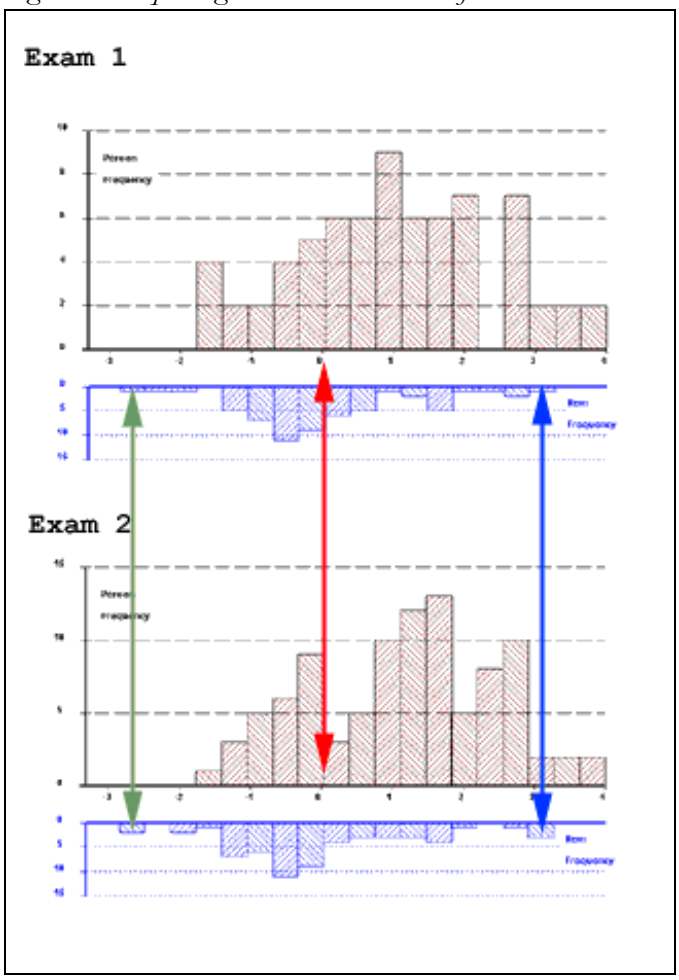


Figure 13b Item Person Map from Exam 2



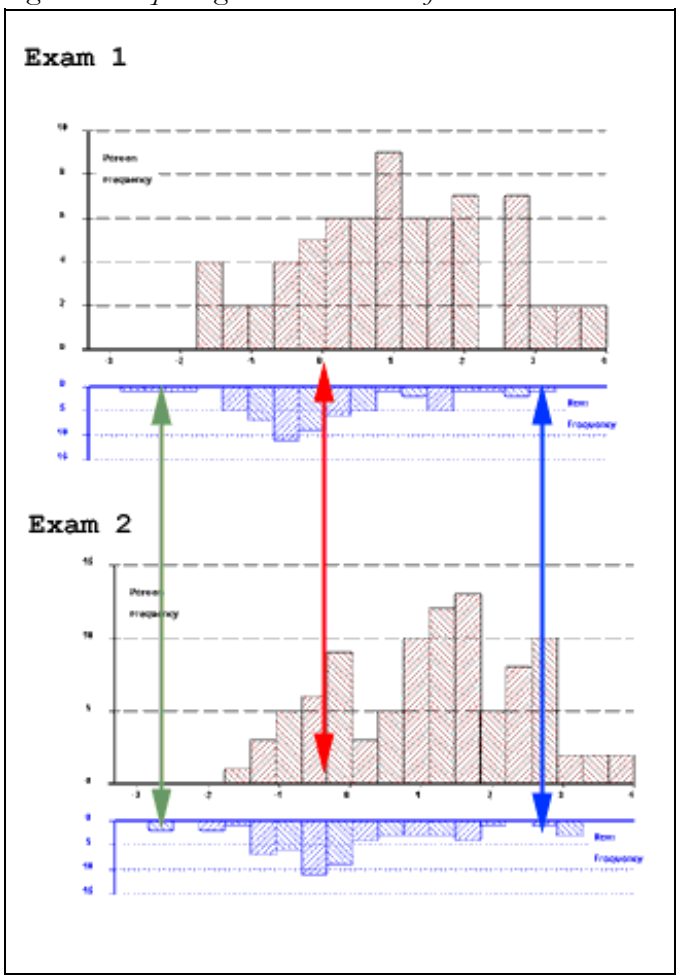
In the scatterplot analysis, the researcher should remove observations that do not yield consistent theta estimates. Afterwards, the average theta of each exam is computed and the difference between the two averages can be obtained from subtracting one from the other. If the difference is zero, then no adjustment in the two IPMs are needed. The two IPMs can be directly compared by using the center (zero) as the point of reference, as shown in Figure 14 (see the red line). In this case, the most difficult item in Exam 1 is comparable to the toughest item in Exam 2 (see the blue line), and the easiest item in Exam 1 is equivalent to the easiest one in Exam 2 (see the green line).

Figure 14. Equating two IPMs without adjustment



However, when the difference is not zero, some adjustment is necessary in order to make two sets of items commensurable. Figure 15 shows an example that the average theta of Exam 1 is greater than that in Exam 2 by .43. In this case, the IPM of Exam 2 will be shifted to the left. After the adjustment, the easiest item in Exam 2 is comparable to the second easiest item in Exam 1 (see the green line), and the hardest item in Exam 1 is equivalent to the second most difficult item in Exam 2 (see the blue line).

Figure 15. Equating two IPMs with adjustment



CONCLUSION

There is no single best test equating methodology. Different contexts call for different approaches. As demonstrated in the preceding examples, simultaneously equating with alternate forms is less resource-intensive because only one test administration is needed. Although across-sample equating with anchor items seems to be less efficient, the test developer can pick the anchor items with desirable attributes, such as stability throughout the period of item exposure. If resources are sufficient, test developers are encouraged to adopt this approach. However, please keep in mind that inserting anchors from an item bank into a test may be problematic when different IRT models are applied. For example, on one occasion item parameters are estimated in Bilog with 3P modeling and the person mean is set to zero. Later a set of anchor items are chosen from

this analysis but the Rasch (1P) model is used in Winsteps and also the item mean, instead of the person mean, is set to zero. Needless to say, the results would be highly misleading. Like across sample equating with anchors, the common subject approach also seems to be less efficient because it necessitates two test administrations. Nevertheless, this approach is helpful in providing evidence of construct validity that is not found in the two previous approaches.

REFERENCES

Andrich, D. (1988). *Rasch models for measurement*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills: Sage.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: LEA Publisher.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*(4), 227-246.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Editors). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In Robert Linn (Ed.) *Educational Measurement 3rd Edition*. (pp. 147-200). New York: American Council on Education, Macmillan.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices (2nd ed.)*. New York: Springer.

Linn, R. L. (1993). Linking results of distinct assessment. *Applied Measurement in Education, 6*(1), 83-102.

McDonald R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA Publisher.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Pomplun, M. R., Omar, H., & Custer, M. (2002). *A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model*. Paper presented at the American Education Researcher Association, New Orleans, LA.

Rasch Measurement Software and Publications. (2002). *Winsteps* [Computer software]. Chicago, IL: The Author.

Zimowski, M., Muraki, E., Mislevy, R.; & Bock D. (2003). *Bilog-MG* [Computer software]. Mooresville, IN: Scientific Software International.

Acknowledgement

Special thanks to Chen, Yi-Hsin for helping in compiling the table of comparing between Winsteps and Bilog.

Appendix Comparison between Winsteps and Bilog

	Winsteps	**Bilog**	**Notes**
Modeling	Rasch model	One-, two-, and three-parameter models	It is said that Rasch modeling is not the same as 1P modeling in Bilog. Rasch is a philosophy of psychometrics, in which data fits the model, not the model fits the data
Scale unit	Default: Logits	Default: Probits This is a rescaling by 1.7	You can enter scaling options in Winsteps and Bilog.
Assumptions	No assumption	Normal sample distribution	Winsteps makes no assumptions about parameter distributions. Bilog assumes normal sample distribution. This may squeeze or spread results particularly at the tails.
Estimation	Joint Maximum Likelihood Estimate (JMLE), also known as Unconditional Maximum Likelihood estimate (UCON).	Marginal Maximum Likelihood Estimate (MMLE) as the default. MMAP and Bayes are also available	MMLE assumes the conditional independence of responses to different items by persons of the same ability. UCON is more biased than conditional methods, but this bias is negligibly small and always less than the standard errors of the estimated measures. This usually has only decimal place effects.
Setting of origin	Item mean=0	Person mean=0 Person variance=1	
Test equating	Across-sample test equating by using anchored items	Multiple-form equating by using common or linking items	

	Winsteps	**Bilog**	**Notes**
Partial credit	Can handle both dichotomous and partial credit items, use step functions	Can handle binary responses only	The numbers in the step function output are difficulty indices in terms of logit, the natural log of the odds ratio. Going from one point to two points, and from two points to three points, will certainly increase the logit difficulty. Distances in logit are comparable. To be specific, if step3 - step2 = 0.1 and step2 - step1 = 0.1. The two "0.1" are considered the same quantities.
Fitness	Winsteps has two types of fitness indexes: INMSQ (Infit mean square) and OUTSQ (Outfit mean square). The INMSQ is usually more informative than the OUTSQ.	Fit statistic is expressed as Chi-square/degree of freedom, where Chi-square results are testing the fit between the expected and the observed.	

Citation

Yu, Chong Ho & Sharon E. Osborn Popp (2005). Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment Research & Evaluation*, 10(4). Available online: http://pareonline.net/getvn.asp?v=10&n=4

Authors

Chong Ho Yu, Ph.D.
PO Box 612
Tempe, AZ 85280

Sharon E. Osborn Popp, Ph.D.
College of Teacher Education and Leadership
PO Box 37100
Phoenix, AZ 85069-7100