Listening. Learning. Leading.®

Research Report
ETS RR–14-07

Enhancing the Equating of Item Difficulty Metrics: Estimation of Reference Distribution

Usama S. Ali

Michael E. Walker

June 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Enhancing the Equating of Item Difficulty Metrics: Estimation of Reference Distribution

Usama S. Ali[1] & Michael E. Walker[1,2]

1 Educational Testing Service, Princeton, NJ
2 Present address: College Board, New York, NY

Two methods are currently in use at Educational Testing Service (ETS) for equating observed item difficulty statistics. The first method involves the linear equating of item statistics in an observed sample to reference statistics on the same items. The second method, or the item response curve (IRC) method, involves the summation of conditional observed item statistics across the reference population total score frequencies. This article introduces a quick and effective method for obtaining the reference distribution for the transition from the linear equating method to the IRC method without recalculating all the item difficulties. More specifically, a mathematical formula is derived to estimate the score distribution of a reference group that maintains the current item difficulty scale. Future research is needed to compare the performance of the two approaches.

Keywords Linear approach; IRC approach; item difficulty equating; *p* value; delta; poststratification

doi:10.1002/ets2.12006

This article concerns two methods currently in use at Educational Testing Service (ETS) for equating observed item difficulty statistics. The first method involves the linear equating of item statistics in an observed sample to statistics on the same items for some reference population. The second method involves the summation of conditional (on total score) observed item statistics across the reference population total score frequencies (i.e., poststratification). This article focuses on the transition from the linear equating method to the poststratification method. More specifically, the article explains the process by which the reference population total score distribution can be estimated from a set of equated item statistics. This reference distribution is a key component of the poststratification method. A mathematical formula is established to retrieve the score distribution of a reference group that maintains the current item difficulty scale.

Notation

Before we present any equations using any of the symbols, we present this notation list.

Let i index the items to be analyzed, from 1 to I.

Let x represent the total score on the test, and let j index the total score levels, from 1 to J.

Let f_j represent the relative frequency of score x_j in the examinee group for which the score distribution is known (the observed group).

Let g_j represent the relative frequency of score x_j in the examinee group for which the score distribution is unknown (the reference group).

Let p_{ij} represent the proportion of correct answers to item i, for the observed-group examinees with total score x_j.

Let q_{ij} represent the proportion of correct answers to item i, for the reference-group examinees with total score x_j.

Let $\overline{p}_i$ represent the proportion of correct answers to item i in the full observed group.

Let $\overline{q}_i$ represent the proportion of correct answers to item i in the full reference group.

Item Difficulty Equating

The proportion of correct responses, or p value, is a commonly used index of item difficulty. For any given item, the p value depends upon the ability of the examinee group. For this reason, it is necessary in any testing program to set a difficulty

Corresponding author: U. Ali, E-mail: uali@ets.org

scale by defining the *p* value in some fixed reference population. Whenever *p* values are observed for any examinee group, they are transformed to the metric of this reference group.

One way to transform observed *p* values to the reference group metric is to obtain a set of items, across a wide range of difficulty, for which we have computed *p* values for the observed as well as the reference groups. We use these data to estimate a function relating observed *p* values to reference *p* values. We can then apply this function to other observed *p* values from the same examinee group to estimate the corresponding item difficulties in the reference group.

Tucker (1987) observed that the *p* value is an S-shaped function of person ability. This function can be approximated by a normal ogive. By applying an inverse normal transformation to the *p* value, we obtain an item difficulty index that is approximately linearly related to ability. At ETS, the resulting scale is called the *delta scale*. More specifically, the delta scale, centered at a value of 13 and with an effective range[1] of 6 (*very easy item*) to 20 (*very difficult item*), results from applying the following two-parameter normal-ogive transformation to the proportion correct item statistic for some item *i* (Holland & Thayer, 1985):

$$\Delta_i = 13 - 4\Phi^{-1}\left(\overline{p}_i\right), \tag{1}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the cumulative normal distribution. Once the item difficulty indices have been transformed in this way, the transformation relating observed to reference values may be approximated using the mean-sigma linear equating method. Equating observed delta indices to those in the reference group makes the item difficulty indices comparable across different examinee samples. The resulting equated delta values may be converted back to *p* values corresponding to the reference group.

The method described above constitutes a functional method, frequently used by testing programs, for adjusting item difficulty indices to the reference population. It involves administering a set of anchor items for which equated item statistics exist, along with any new items to a group of examinees. The resulting observed delta values are linked to the equated delta values through a linear (i.e., mean-sigma) procedure. Then the resulting linear parameters (i.e., slope and intercept) are used to obtain reference deltas for nonanchor items, which only have observed deltas.

One problem with the linear method is that the function relating observed to reference difficulty statistics is usually based on few (perhaps around 20) data points. Equating with such small numbers leads to relatively large error. For a given middle difficulty item, the item statistics estimated in this way from various samples of items can differ by as much as one delta point (roughly 10 points in terms of percent correct). For more difficult items, the range of estimated item difficulties can be even larger. The only way to decrease the error substantially is to greatly increase the number of items that are repeated across the two groups of examinees. For many reasons, such a route is not feasible.[2]

Fortunately, another method exists for adjusting *p* values. This method generates empirical item response curves (IRCs; i.e., proportion correct, or *p* values, conditioned on total score) for all items administered to a given examinee group. Given the total score distribution for the reference group, we can compute reference *p* values by multiplying the conditional values by the relative frequencies at the corresponding score points and then summing across the reference distribution.

As with the linear method, the error in this poststratification method can be reduced by increasing the sample size. Unlike with the linear method, with the poststratification method, people rather than anchor items constitute the sample. Thus, even with relatively few items, so long as the number of examinees is large enough, the resulting adjusted statistics will be very precise.

We might expect that for large-scale testing programs, the poststratification method is preferable to the linear method in terms of precision. If this is indeed the case, testing programs using the linear method may have to be changed to the poststratification method. To do so, it would be necessary to have the total score distribution in the reference group. For more established testing programs, that distribution may no longer be available. The program could always just choose the examinees from an arbitrary test administration to use as the reference group and set the item difficulty scale anew. That, however, would require resetting the statistical specifications for all tests, as well as recomputing the reference difficulty statistics for every item in the pool for every test title. Given the effort involved and the disruption caused by such a plan, we explored another alternative.

The remainder of this article illustrates how we can estimate the current reference distribution from a set of items that have both observed and reference item difficulty statistics. By doing so, we can maintain the current item difficulty scales while transitioning from the linear to the poststratification item difficulty equating method. The mechanism illustrated here is much less labor-intensive than resetting the difficulty scale to a more recent reference distribution. To estimate the reference score distribution, we work with the item means or *p* values for the reference group.

Estimation of Reference Distribution

Setup

Consider the situation in which the empirical IRCs for I test items are plotted against a J-point total score. This total score does not represent the sum of the I test items. Rather, the score is reported on a scale that is independent of the items generating the IRCs (e.g., we could use the scaled score typically associated with a large-scale testing program). The empirical IRCs delineate conditional p_{ij} values at each total score point x_j. If we know the empirical IRCs and the observed total score frequencies p_{ij}, then we can compute the $\overline{p}_i$ values in the observed population.

Let us start with the total score x_j and the proportion correct values p. We represent the height of the IRC for item i at total score j by p_{ij}. Given f_j, the relative frequency of scores x_j in the observed population, we can compute the observed p value $\overline{p}_i$ for item i by

$$\overline{p}_i = \sum_j f_j p_{ij} \tag{2}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

In a similar way, if we have the empirical IRCs and the reference population, we can estimate the $\overline{q}_i$ values that would result if the items were given to the reference population. We do this by substituting the relative frequencies for the reference population in Equation 2.

$$\overline{q}_i = \sum_j g_j q_{ij}. \tag{3}$$

Problem

We have observed values f_j for all score points x_j for some group of examinees (the total score distribution in the observed group). We want to estimate the g_j (the total score distribution in the reference group). The solution below presents the mathematical formula for deriving the reference distribution score frequencies g_j given the reference and observed p value statistics. The total test score considered in this context is the scaled score.

Solution

Let us start with a set of I items with item difficulty estimates corresponding to both the observed and reference populations. For any dichotomous item i, assume that the IRC (i.e., the p values conditioned on examinee total score) is invariant across different subpopulations. Thus, we can use the p_{ij} values (the conditional correct-answer proportions in the observed group) as estimates for the q_{ij} (the conditional correct-answer proportions in the reference group). We may want to smooth the IRC using some techniques such as loglinear presmoothing (see Holland & Thayer, 2000).

We also have estimates for the $\overline{q}_i$ values, obtained by transforming the equated delta values (found using the linear method described above).

Recall from Equations 2 and 3 that for any item i, $\overline{p}_i = \sum_j p_{ij} f_j$ and $\overline{q}_i = \sum_j q_{ij} g_j$.

Applying the second of these two equations at each score level j,

$$\begin{aligned} q_{11}g_1 + q_{12}g_2 + \ \ldots\ + q_{1J}g_J &= \overline{q}_{1(\text{ref})} \\ q_{21}g_1 + q_{22}g_2 + \ \ldots\ + q_{2J}g_J &= \overline{q}_{2(\text{ref})} \\ &\ldots \\ q_{I1}g_1 + q_{I2}g_2 + \ \ldots\ + q_{IJ}g_J &= \overline{q}_{I(\text{ref}).} \end{aligned} \tag{4}$$

To solve this set of equations, we need at least as many items as there are points on the total score scale (i.e., $I \geq J$), so that the number of equations is at least as great as the number of unknown values. The unknown values in the series of I simultaneous equations in Equation 4 are the relative frequencies at the total score points. So the matrix equation would be

$$\mathbf{Q}_{I\times J}\mathbf{g}_{J\times 1} = \overline{\mathbf{q}}_{I\times 1}.$$

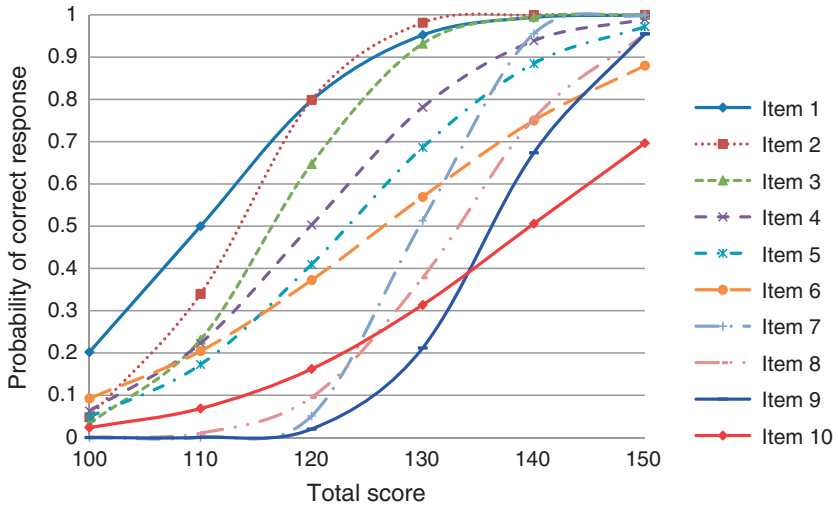

Figure 1 Empirical item response curves (IRCs) for 10 items over 6 score points.

In general, $i \neq j$, so we have to premultiply both sides of the equation by $\mathbf{Q}'$ to get

$$\mathbf{Q}'_{J\times I}\ \mathbf{Q}_{I\times J}\mathbf{g}_{J\times 1} = \mathbf{Q}'_{J\times I}\ \overline{\mathbf{q}}_{I\times 1}.$$

Then we can premultiply both sides by $(\mathbf{Q}'\mathbf{Q})^{-1}$ to get

$$\mathbf{g} = \left(\mathbf{Q}'\mathbf{Q}\right)^{-1}\mathbf{Q}'\overline{\mathbf{q}}. \tag{5}$$

The values in matrix $\mathbf{g}$ are the unknown reference group frequencies that we want to estimate.

Numerical Example

In Figure 1, 10 items are displayed ($I = 10$), each with probability of correct response (0 to 1) plotted against total score (100, 110, 120, 130, 140, and 150, i.e., $J = 6$). Each IRC was generated using a two-parameter normal-ogive model, whose mean and standard deviation are given in Table 1. The resulting conditional p_{ij} values are shown in Table 2, the item-by-score matrix.

The observed and reference score distribution are described by the relative frequencies that appear below the conditional p_{ij} (Table 2, the last two rows). We obtain the observed $\overline{p}_i$ (Table 2, second column) by multiplying the conditional by the corresponding relative frequencies and summing. The reference $\overline{q}_i$ (Table 2, third column) is computed in the same way. In this example, the p_{ij} are the same as q_{ij}. In general, we would assume that the conditional p values are the same in both populations.

In our case, the solution, using the derived formula in Equation 5, is presented in Table 3. These resulting reference relative frequencies are exactly the same as those shown in Table 2, last row.

Practical Considerations and Remarks

The current item analysis design for some testing programs uses an observed score model for item analysis and the linear model for delta equating. This article explains a quick and effective method for obtaining the reference distribution without recalculating all the item difficulties for all test titles. On the other hand, the application of the new approach for delta equating has several consequences and more challenges that we need to highlight and to solve before we can implement the new approach operationally.

First, the improved precision of the IRC approach compared to the linear model needs to be verified. An estimation of the variability in estimating the equated difficulty indices needs to be established (e.g., either analytically or using the bootstrap method) to compare the two competing approaches of delta equating. Also, we need to check the robustness of estimated item difficulty using the IRC approach against the poor estimation of item difficulty of one or two items.

Table 1 Parameters of Normal-Ogive Item Response Curves (IRCs)

	Parameters	
Item	Mean	SD
1	1.00	1.2
2	1.33	0.8
3	1.66	0.9
4	1.99	1.3
5	2.32	1.4
6	2.65	2.0
7	2.98	0.6
8	3.31	1.0
9	3.64	0.8
10	3.97	2.0

Table 2 The Probabilities of Correct Responses on Each Item at Different Total Scores

			p_{ij} (Empirical item response curves)					
Item	$\overline{p}_i$	$\overline{q}_i$	100	110	120	130	140	150
1	.8287	.9423	.2023	.5000	.7977	.9522	.9938	.9996
2	.8149	.9509	.0482	.3400	.7988	.9816	.9996	.0000
3	.7305	.9094	.0326	.2317	.6472	.9317	.9953	.9999
4	.6216	.8175	.0629	.2232	.5031	.7814	.9390	.9897
5	.5424	.7501	.0487	.1729	.4096	.6864	.8849	.9722
6	.4745	.6452	.0926	.2047	.3726	.5695	.7502	.8800
7	.3467	.6669	.0000	.0005	.0512	.5133	.9554	.9996
8	.2886	.5514	.0005	.0104	.0951	.3783	.7549	.9545
9	.1938	.4566	.0000	.0005	.0202	.2119	.6736	.9554
10	.2579	.4140	.0236	.0688	.1623	.3138	.5060	.6967
	f_j		.0214	.1359	.3413	.3413	.1359	.0241
	g_j		.0013	.0214	.1359	.3413	.3413	.1587

Table 3 The Resulting Reference Relative Frequencies per Score

x_j	g_j
100	.0013
110	.0214
120	.1359
130	.3413
140	.3413
150	.1587

Second, determining the timing of executing the IRC method during the operational cycle is essential. More specifically, is it possible to implement the IRC method during preliminary item analysis (PIA), conducted before test equating? Or, is the method only feasible for final item analysis (FIA), conducted after the test is equated? The timing is important because it is related to the more challenging problem of defining the total score upon which the item statistics are conditioned. Raw scores over different administrations are not comparable, such that the IRCs cannot be invariant across these different subpopulations. This leads us to the use of scaled scores, which are not available for PIA (because the current test has not yet been equated). We may consider conducting a preliminary equating to facilitate PIA; then we would need to investigate how much imprecision or bias such a procedure added to the item difficulty estimation.

Third, the previous challenge leads to the question: Which anchor item is not performing properly, based on which criteria, when evaluating the equating set for (score and) delta equating? In the linear procedure, these equating items are evaluated through delta plots that are typically used to identify items that performed differently across the two samples.

One delta plot depicts the relationship between the observed and output deltas with a bandwidth of 2 SD around the fitted line. The other plot illustrates the relationship of the difference between output and input equated deltas on one axis with the input delta on the other axis. The rule of thumb is to detect items satisfying $|\text{output } \Delta - \text{input } \Delta| \geq 2.0$. In the IRC approach, we can similarly illustrate the delta difference to evaluate the equating set.

Fourth, the minimum number of items (or test forms) needed to get robust estimates of the reference group distribution is still among the research questions to be answered for best practice. Other challenges are related to designing procedure that can be implemented easily.

In summary, the above mentioned issues that might affect the implementation of the reference group delta equating approach need further investigation. Our next step is to tailor a study to address some of these issues and find reliable and valid solutions.

Acknowledgments

We would like to thank Skip Livingston, Rebecca Zwick, Gautam Puhan, and Lili Yao for their suggestions, which have greatly improved this article.

Notes

1 Theoretically, delta ranges from $-\infty$ to $+\infty$. Practically speaking, we will rarely use items easier than a delta value of 6 ($p = .96$) or harder than a delta value of 20 ($p = .04$) on a test.

2 Equating error is inversely proportional to the square root of the sample size. Thus, to reduce the error by one half, we need to include four times as many anchor items (i.e., previously exposed items). Most testing programs work to minimize the number of reused items across test forms.

References

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Technical Report No. PSRTR-85-64). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Tucker, L. R. (1987). *Developments in classical item analysis methods* (Research Report No. RR-87-46). Princeton, NJ: Educational Testing Service.

Action Editor: Rebecca Zwick

Reviewers: Gautam Puhan and Lili Yao