Equating Tests under Neural Test Theory

SHOJIMA Kojiro

October 2007

Department of Test Analysis and Evaluation, Research Division, The National Center for University Entrance Examinations

Equating tests under neural test theory

Kojiro Shojima

Abstract

A method of equating tests under neural test theory (NTT) is proposed. It uses the concurrent calibration (CC) method in which the item reference profiles of only the items to be equated are updated. The CC method is very easy to use and is applicable to the common items and common examinees designs, and the mixed design of both these designs. In addition, it can deal with both horizontal and vertical equating conditions.

Key words: neural test theory, horizontal equating, vertical equating, equating indicator, concurrent calibration, missing data, common items design, common examinees design.

ニューラルテストモデルにおけるテスト等化

荘島宏二郎

要約

本研究では，ニューラルテスト理論 (neural test theory, NTT) において，共時推定法を用いてテスト同士を等化する方法を論じた．共時推定法は，等化指示子を用いて，等化すべき項目の項目参照プロファイルのみを更新していくという簡単な方法であり，共通項目計画，共通受験者計画，混合計画のいずれにも適用可能である．また，共時推定法は，水平的等化と垂直的等化の 2 つの状況に対応できるが，その際，手続きが少し異なる．今後の課題は，数値実験や実際のテスト運用の中で本方法の精度を確認することである．

キーワード: ニューラルテスト理論，水平的等化，垂直的等化，等化指示子，共時推定，欠測データ，共通項目計画，共通受験者計画．

Department of Test Analysis and Evaluation, Research Division, The National Center for University Entrance Examinations

1 Introduction

Neural test theory (NTT; Shojima, 2007a, 2007b) is an efficient statistical model for analyzing test data. The mechanism of the NTT model is based on that of the self-organizing map (SOM; Kohonen, 1995). The latent scale assumed in NTT is rank-ordered, which is the key difference from the scale assumed in item response theory (IRT; Lord, 1980; Hambleton & Swaminathan, 1985).

Tests are required to be standardized, and test standardization is composed of scaling and equating. Test scaling is necessary to examine the statistical features of the test to make it a reliable scale for measuring ability. Shojima (2007c, 2007d, 2007e) describes test scaling methods under NTT.

Test equating as another test standardization is also indispensable for comparing two or more tests on the same scale. The measured values of two persons weighed by different two weighing machines are always comparable because weighing machines are already equated (standardized). On the other hand, the scores of two examinees who take different two tests are not comparable as they are. The difference in their scores is a composite of the differences in the test difficulties and examinee abilities. In this paper, a method for equating tests under NTT is described.

Table 1 shows the data structure for test equating, where $\boldsymbol{U}_{gs}^{(z)}$ is the response matrix of group G_g to item set S_s. The superscript z is a missing indicator, which is a dichotomous variable coded 1 when the data is observed and 0 when the data is missing. In addition, N_g is the sample size of group G_g and n_s is the number of items in item set S_s.

Table 1: Data Structure for Equating

	Items	Item Set S_1	Item Set S_2	Item Set S_3
Examinees	$N \backslash n$	n_1	n_2	n_3
Group G_1	N_1	$\boldsymbol{U}_{11}^{(z)}$	$\boldsymbol{U}_{12}^{(z)}$	$\boldsymbol{U}_{13}^{(z)}$
Group G_2	N_2	$\boldsymbol{U}_{21}^{(z)}$	$\boldsymbol{U}_{22}^{(z)}$	$\boldsymbol{U}_{23}^{(z)}$
Group G_3	N_3	$\boldsymbol{U}_{31}^{(z)}$	$\boldsymbol{U}_{32}^{(z)}$	$\boldsymbol{U}_{33}^{(z)}$

The data structures used most frequently for test equating are the common items design and common examinees design (CID and CED; Hambleton & Swaminathan, 1985). The

CID is a state when the data structure of Table 1 is under the condition that

$$\boldsymbol{U}_{CID} = \begin{bmatrix} \boldsymbol{U}_{11}^{(1)} & \boldsymbol{U}_{12}^{(1)} & \boldsymbol{U}_{13}^{(0)} \\ \boldsymbol{U}_{21}^{(0)} & \boldsymbol{U}_{22}^{(0)} & \boldsymbol{U}_{23}^{(0)} \\ \boldsymbol{U}_{31}^{(0)} & \boldsymbol{U}_{32}^{(1)} & \boldsymbol{U}_{33}^{(1)} \end{bmatrix}. \tag{1}$$

This is the case where there are no examinees in group G_2 and no responses of groups G_1 and G_3 to item sets S_3 and S_1, respectively. In this case, item sets S_1 and S_2 are the base test items, item sets S_2 and S_3 are the items of the target test to be equated, and item set S_2 is the set of common items used for equating the target test onto the scale of the base test. Generally, the number of common items, n_2, is smaller than n_1 and n_3.

Next, the CED is a state when the data is observed as follows:

$$\boldsymbol{U}_{CED} = \begin{bmatrix} \boldsymbol{U}_{11}^{(1)} & \boldsymbol{U}_{12}^{(0)} & \boldsymbol{U}_{13}^{(0)} \\ \boldsymbol{U}_{21}^{(1)} & \boldsymbol{U}_{22}^{(0)} & \boldsymbol{U}_{23}^{(1)} \\ \boldsymbol{U}_{31}^{(0)} & \boldsymbol{U}_{32}^{(0)} & \boldsymbol{U}_{33}^{(1)} \end{bmatrix}, \tag{2}$$

where the responses for item set S_1 of group G_3, item set S_2 of all the examinees, and item set S_3 of group G_1 are missing. Then, item sets S_1 and S_3 are the base and target test items, respectively, and the examinees of group G_2 are called the common examinees. The sample size of the common examinees, N_2, is usually smaller than N_1 and N_3. In practical cases of test equating, a mixed design (MD) of CID and CED is frequently used. Its data structure is as follows:

$$\boldsymbol{U}_{MD} = \begin{bmatrix} \boldsymbol{U}_{11}^{(1)} & \boldsymbol{U}_{12}^{(1)} & \boldsymbol{U}_{13}^{(0)} \\ \boldsymbol{U}_{21}^{(1)} & \boldsymbol{U}_{22}^{(1)} & \boldsymbol{U}_{23}^{(1)} \\ \boldsymbol{U}_{31}^{(0)} & \boldsymbol{U}_{32}^{(1)} & \boldsymbol{U}_{33}^{(1)} \end{bmatrix}. \tag{3}$$

In addition, the data structure of the equivalent group design (EGD) is given as

$$\boldsymbol{U}_{EGD} = \begin{bmatrix} \boldsymbol{U}_{11}^{(1)} & \boldsymbol{U}_{12}^{(0)} & \boldsymbol{U}_{13}^{(0)} \\ \boldsymbol{U}_{21}^{(0)} & \boldsymbol{U}_{22}^{(0)} & \boldsymbol{U}_{23}^{(0)} \\ \boldsymbol{U}_{31}^{(0)} & \boldsymbol{U}_{32}^{(0)} & \boldsymbol{U}_{33}^{(1)} \end{bmatrix}, \tag{4}$$

where there are neither responses of the common items nor common examinees. Under EGD, the base test (item set S_1) and the target test (item set S_3) can be equated if and only if the ability levels of groups G_1 and G_3 are identical. However, this design is rarely used in practice because there is little chance of the ability levels of two groups being equal in a strict sense.

The base test must be scaled before the target test is equated. Under NTT, the state that a test is scaled means that the estimates of the item reference profiles (IRPs; Shojima,

2007a) of the base test items are obtained. The IRP directly shows the statistical features of each item, and the size of the IRP is identical to the number of latent ranks. The target test items are equated onto the base test scale whether or not the target test is already scaled.

2 Method

Although this study focused on test equating under the dichotomous neural test (DNT; Shojima, 2007a) model, the procedure can be directly transferred to the case under the graded neural test (GNT; Shojima, 2007b) model, which is a polytomous NTT model for ordered polytomous data, and it reduces to the DNT model when all items are binary.

In addition, the concurrent calibration (CC; Lord, 1980; Kim & Cohen, 1998) equating method is used. In the CC method, the parameters of the test items to be equated are estimated provided that those of the base test items are fixed. Although this method was originally proposed for test equating under item response theory (IRT; Lord, 1980; Hambleton & Swaminathan, 1985), the idea is applicable to the case under NTT. The CC method is feasible for any data structures of (1)-(3) because test equating by the CC method corresponds to the problem of estimating the IRPs of the items to be equated under the condition that the IRPs of the base test items are fixed.

Furthermore, there are two types of equating conditions: horizontal and vertical equating conditions. The former is the case for equating tests with nearly equal difficulties, and the latter is the case for equating ones with different difficulties.

2.1 Horizontal equating

Let us assume that item sets S_1 and S_2 are the base test items, which have already been scaled, and item sets S_2 and S_3 are the target test items. That is, item set S_3 is the set of items to be equated onto the base test scale. Let the response matrix of examinees be denoted by $\boldsymbol{U} = \{u_{ij}\}$ $(N \times n)$, where u_{ij} is the response of examinee i to item j, $N = N_1 + N_2 + N_3$, and $n = n_1 + n_2 + n_3$. The response matrix $\boldsymbol{U}$ is any of the three data structures (1)-(3). In addition, let $\boldsymbol{Z} = \{z_{ij}\}$ $(N \times n)$ be the missing indicator matrix (Shojima, 2007e), where z_{ij} is a dichotomous variable coded 1 when u_{ij} is observed and 0 when it is missing.

Next, let Q be the number of latent ranks of the base test scale. Then, the reference matrix of item sets S_1, S_2, and S_3 is

$$\boldsymbol{V} = \{v_{qj}\} = [\hat{\boldsymbol{V}}_1 \ \hat{\boldsymbol{V}}_2 \ \boldsymbol{V}_3] \quad (Q \times n), \tag{5}$$

where the elements corresponding to item sets S_1 and S_2 have already been estimated, and the elements for item set S_3 are unknown parameters. The q-th row vector in $\boldsymbol{V}$ is the rank reference vector (RRV) of rank R_q, and the j-th column vector in $\boldsymbol{V}$ is the item reference profile of item j. In addition, the equating indicator is defined as

$$\boldsymbol{e} = \{e_j\} = [\mathbf{0}'_{n_1}\ \mathbf{0}'_{n_2}\ \mathbf{1}'_{n_3}]' \quad (n \times 1), \tag{6}$$

where e_j is a dichotomous variable coded 1 when item j is to be equated and 0 otherwise.

With the conditions described above, test equating under NTT can easily be accomplished if only the IRPs of item set S_3 are updated provided that the base test items are fixed in the statistical learning process. That is, the procedure for updating the reference vectors is

$$\begin{aligned} &\text{For } (q{=}1;\ q \le Q;\ q = q+1) \\ &\text{— } \boldsymbol{v}_q^{(t,h)} = \boldsymbol{v}_q^{(t,h-1)} + h_{qw}(t)\{\boldsymbol{e} \odot \boldsymbol{z}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)})\}, \end{aligned} \tag{7}$$

where $\boldsymbol{u}_h^{(t)} = \{u_{hj}^{(t)}\}$ $(n \times 1)$ is the h-th row vector of $\boldsymbol{U}^{(t)}$ which is the input data in the t-th period, $\boldsymbol{z}_h^{(t)}$ is the missing indicator corresponding to $\boldsymbol{u}_h^{(t)}$, and $\boldsymbol{v}_q = \{v_{qj}^{(t,h-1)}\}$ $(n \times 1)$ is the RRV of latent rank R_q obtained by learning the supervising signal $\boldsymbol{u}_{h-1}^{(t)}$. In addition, the factor h_{qw} is defined as

$$h_{qw}(t|\alpha_t, \sigma_t^2) = \alpha_t \exp\Big\{-\frac{(R_q - R_w)^2}{2\sigma_t^2}\Big\}, \tag{8}$$

where

$$\alpha_t = \frac{T-t+1}{T}\alpha_1, \tag{9}$$

and

$$\sigma_t = \frac{(T-t)\sigma_1 + (t-1)\sigma_0}{T-1}. \tag{10}$$

From (7), the reference matrix in the t-th period is obtained as

$$\boldsymbol{V}^{(t)} = \{v_{qj}^{(t)}\} = [\hat{\boldsymbol{V}}_1\ \hat{\boldsymbol{V}}_2\ \boldsymbol{V}_3^{(t)}] \quad (Q \times n). \tag{11}$$

Consequently, the elements corresponding to the base test items in the reference matrix are invariant, while the elements for only the items to be equated are updated through the statistical learning process.

2.2 Vertical equating

The equating procedure when the difficulty levels of the tests are different is described in this section. If the difficulty range of the base test covers that of the target test, the procedure described for horizontal equating is applicable to this case. When the entire region of the difficulty range of the target test is outside that of the base test, it is necessary to add some upper or lower ranks to the original base test scale.

Let us suppose the condition where R_{Q+1}, R_{Q+2}, and R_{Q+3} as the upper ranks and R_0, R_{-1}, and R_{-2} as the lower ranks are added. Then, the reference matrix and the equating indicator becomes

$$\boldsymbol{V} = \{v_{qj}\} = \begin{bmatrix} \boldsymbol{v}'_{-2,1} & \boldsymbol{v}'_{-2,2} & \boldsymbol{v}'_{-2,3} \\ \boldsymbol{v}'_{-1,1} & \boldsymbol{v}'_{-1,2} & \boldsymbol{v}'_{-1,3} \\ \boldsymbol{v}'_{0,1} & \boldsymbol{v}'_{0,2} & \boldsymbol{v}'_{0,3} \\ \hat{\boldsymbol{V}}_1 & \hat{\boldsymbol{V}}_2 & \boldsymbol{V}_3 \\ \boldsymbol{v}'_{Q+1,1} & \boldsymbol{v}'_{Q+1,2} & \boldsymbol{v}'_{Q+1,3} \\ \boldsymbol{v}'_{Q+2,1} & \boldsymbol{v}'_{Q+2,2} & \boldsymbol{v}'_{Q+2,3} \\ \boldsymbol{v}'_{Q+3,1} & \boldsymbol{v}'_{Q+3,2} & \boldsymbol{v}'_{Q+3,3} \end{bmatrix} \quad \{(Q+6) \times n\}, \tag{12}$$

and

$$\boldsymbol{E} = \{e_{qj}\} = \begin{bmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \\ \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \\ \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \\ \mathbf{0}_{Q\times n_1} & \mathbf{0}_{Q\times n_2} & \mathbf{1}_{Q\times n_3} \\ \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \\ \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \\ \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \mathbf{1}'_{n_3} \end{bmatrix} \quad \{(Q+6) \times n\}. \tag{13}$$

Then, the reference matrix updating is executed as follows:

$$\begin{aligned} &\text{For } (q = -2;\ q \le Q+3;\ q = q+1) \\ &\text{— } \boldsymbol{v}_q^{(t,h)} = \boldsymbol{v}_q^{(t,h-1)} + h_{qw}(t)\{\boldsymbol{e}_q \odot \boldsymbol{z}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)})\}, \end{aligned} \tag{14}$$

where $\boldsymbol{e}_q$ is the q-th row vector of $\boldsymbol{E}$. From the above equation, the reference matrix in the t-th period becomes

$$\boldsymbol{V}^{(t)} = \{v_{qj}^{(t)}\} = \begin{bmatrix} \boldsymbol{v}_{-2,1}^{(t)\prime} & \boldsymbol{v}_{-2,2}^{(t)\prime} & \boldsymbol{v}_{-2,3}^{(t)\prime} \\ \boldsymbol{v}_{-1,1}^{(t)\prime} & \boldsymbol{v}_{-1,2}^{(t)\prime} & \boldsymbol{v}_{-1,3}^{(t)\prime} \\ \boldsymbol{v}_{0,1}^{(t)\prime} & \boldsymbol{v}_{0,2}^{(t)\prime} & \boldsymbol{v}_{0,3}^{(t)\prime} \\ \hat{\boldsymbol{V}}_1 & \hat{\boldsymbol{V}}_2 & \boldsymbol{V}_3^{(t)} \\ \boldsymbol{v}_{Q+1,1}^{(t)\prime} & \boldsymbol{v}_{Q+1,2}^{(t)(t)\prime} & \boldsymbol{v}_{Q+1,3}^{(t)\prime} \\ \boldsymbol{v}_{Q+2,1}^{(t)\prime} & \boldsymbol{v}_{Q+2,2}^{(t)(t)\prime} & \boldsymbol{v}_{Q+2,3}^{(t)\prime} \\ \boldsymbol{v}_{Q+3,1}^{(t)\prime} & \boldsymbol{v}_{Q+3,2}^{(t)\prime} & \boldsymbol{v}_{Q+3,3}^{(t)\prime} \end{bmatrix} \quad \{(Q+6) \times n\}. \tag{15}$$

It is obvious from the above procedure that horizontal equating under NTT is a special case of vertical equating when no rank is added to the base test scale.

In vertical equating, both updated and fixed elements are mixed in the IRP of each base test item. Accordingly, when the monotonically increasing constraint (Shojima, 2007a, 2007b) is imposed on the IRPs, it is necessary to make the IRPs of the base test items monotonically increasing even between the updated and fixed elements. In this case, the following step should be inserted just before the counter t is incremented by one. That is,

$$
\begin{aligned}
&\text{For } (j{=}1;\ j \leq n_1 + n_2;\ j = j + 1) \\
&\quad \text{For } (q{=}0;\ q \leq -2;\ q = q - 1) \\
&\quad \text{— If } v_{q-1,j}^{(t)} \geq v_{qj}^{(t)}, \text{ then } v_{q-1,j}^{(t)} = v_{qj}^{(t)}. \\
&\quad \text{For } (q = Q;\ q \leq Q + 2;\ q = q + 1) \\
&\quad \text{— If } v_{q+1,j}^{(t)} \leq v_{qj}^{(t)}, \text{ then } v_{q+1,j}^{(t)} = v_{qj}^{(t)}.
\end{aligned}
\tag{16}
$$

3 Discussion

This paper describes a very simple procedure for test equating under neural test theory (NTT) using the concurrent calibration method, in which the item reference profiles (IRPs) of the items to be equated are updated, provided that the IRPs of the base test items are invariant. With this method, the IRP estimates of the target test items are already equated onto the base test scale.

Although this method can flexibly deal with any equating design if the data structure is not extremely inadequate for test equating, the reliability of the equating is dependent on a large sample size of the common examinees or a large number of common items. Further research is required to examine how many common items and examinees are necessary to make the equating accurate. In addition, studies for determining how many ranks should be added to the original scale in vertical equating are also necessary.

Test equating can be numerically or statistically executed by applying the results of this study, and for years test equating has been an essential technique for administrating tests under NTT. However, it is still difficult to manage tests year after year. As it is already known that item parameters of item response theory (IRT) change over the years, the IRPs estimated at a certain point will not be invariant against time. It is essential for all those who have administrated tests under NTT to share their experienes.

References

Hambleton, R. K. & Swaminathan, H. (1985) *Item response theory*. Kluwer-Nijhoff.

Kim, S.-H. & Cohen, A. S. (1998) A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, **22**, 131-143.

Kohonen, T. (1995) *Self-organizing maps*. Springer.

Lord, F. M. (1980) *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Shojima, K. (2007a) Neural test theory. *DNC Research Note*, 07-02.

Shojima, K. (2007b) The graded neural test model: A neural test model for ordered polytomous data. *DNC Research Note*, 07-03.

Shojima, K. (2007c) Maximum likelihood estimation of latent rank under the neural test model. *DNC Research Note*, 07-04.

Shojima, K. (2007d) Chi-square goodness-of-fit test under the neural test model. *DNC Research Note*, 07-05.

Shojima, K. (2007e) Estimation for neural test models with missing data. *DNC Research Note*, 07-09.