Research Report

Linking Parameter Estimates Derived From an Item Response Model Through Separate Calibrations

Shelby J. Haberman

December 2009

ETS RR-09-40

Linking Parameter Estimates Derived From an Item Response Model Through Separate Calibrations

Shelby J. Haberman

ETS, Princeton, New Jersey

December 2009

Abstract

A regression procedure is developed to link simultaneously a very large number of item response theory (IRT) parameter estimates obtained from a large number of test forms, where each form has been separately calibrated and where forms can be linked on a pairwise basis by means of common items. An application is made to forms in which a two-parameter logistic model is applied to dichotomous items and a general partial credit model is applied to polytomous items.

Key words: Generalized partial credit model, two-parameter logistic model, regression analysis

Acknowledgments

The author thanks Neil Dorans, Alina von Davier, and Dan Eignor for helpful comments.

Linking test forms by use of item response theory (IRT) is a familiar activity when one reference form is equated to one base form (Hambleton, Swaminathan, & Rogers, 1991, ch. 9). In practice, testing programs often link one test form to another in circumstances in which multiple test forms are involved. In typical cases, new forms are equated based on one or more old forms, and these old forms have in turn been linked to earlier forms. If modifications in linking procedures are required for any reason, then their implementation can be an arduous task. In this report, an approach based on linear models is considered that permits simultaneous linking of a large number of forms without the need to produce a long sequence in which one form is linked to one or more previously used forms. The suggested approach is a generalization of the log-mean mean procedure (Mislevy & Bock, 1990) briefly described in Kolen and Brennan (2004).

Section 1 describes the model used in the analysis of IRT true-score equating. Section 2 describes the required regression analysis. Section 3 provides some concluding remarks.

1 The Model

To describe the general problem under study, consider the following situation. A given test has administrations 1 to T, where $T > 1$. In these T administrations, a finite and nonempty set J of v items is used, but not all examinees receive all items. Associated with Item j in J are response scores 0 to $r_j - 1$, where $r_j \geq 2$ is an integer. Associated with an examinee is a real proficiency variable θ. If an examinee has proficiency θ, then the probability $P_j(k|\theta) > 0$ that the examinee receives score k, $0 \leq k \leq r_j - 1$, on item j satisfies

$$\log[P_j(k|\theta)/P_j(k-1|\theta)] = Da_j(\theta - b_j + d_{jk}),$$

where a_j is a real and normally positive, b_j and d_{jk} are real, and $\sum_{k=1}^{r_j-1} d_{jk} = 0$ (Muraki, 1997). Thus one has a generalized partial credit model that reduces to a two-parameter logistic model if item j has $r_j = 2$ categories. The item discrimination a_j, the item difficulty b_j, and the location parameters d_{jk} are unknown, except that d_{jk} obviously is 0 in the dichotomous case of $r_j = 2$. The multiplier D is a known constant. It may be 1, 1.7, or 1.702. The values 1.7 and 1.702 are employed so that parameters from IRT models based on the logistic distribution function will be similar in value to IRT parameters based on the normal distribution function (Lord & Novick, 1968, p. 400). At administration t, $1 \leq t \leq T$, examinee i, $1 \leq i \leq N_t$, is administered a nonempty subset J_{it} of the items in J. The response score for item j is then X_{ijt}. It is assumed that, conditional on the proficiency θ_i of examinee i, the X_{ijt}, j in J_{it}, are independent. In addition, at administration t, the examinee population is assumed to have a normal proficiency distribution $N(B_t, A_t^2)$ with mean B_t and standard deviation A_t. To identify parameters, let $A_1 = 1$ and $B_1 = 0$, so that the proficiency distribution for Administration 1 is a standard normal distribution. This convention is reasonable if the form used in Administration 1 is regarded as the base form.

Marginal maximum-likelihood estimation may be employed to determine the parameters A_t, B_t, a_j, b_j, and d_{jk}; however, this approach is challenging with conventional software if J includes thousands of items. A less computationally demanding approach involves separate calibrations for each form. For each Administration t, let the set J_t include each Item j such that the number N_{jt} of examinees i with j in J_{it} exceeds some minimum threshold $m_j > 0$. To ensure that parameters can be estimated, assume that the number v_t of elements of J_t is at least 3, and assume that each Item j in the set J is in J_t for some Administration t. A scaled version of the parameters a_j, b_j, and d_{jk} for each Item j in the set J_t of items associated with Administration t may be obtained with conventional software such as Parscale for estimation by maximum marginal likelihood. In conventional application of such software, the marginal distribution of the proficiency distribution is a standard normal distribution. Such a marginal distribution can be achieved by use of a linear transformation. The proficiency θ_{it} of examinee i in administration t can be converted to the scaled proficiency $\theta'_{it} = A'_t\theta_{it} + B'_t = (\theta_{it} - B_t)/A_t$, where $A'_t = 1/A_t$ and $B'_t = -B_t/A_t$, so that θ'_{it} has a standard normal distribution. One may then apply a conventional analysis to the data from Administration t. In this analysis, the conditional probability $P'_j(k|\theta')$ that $X_{ijt} = k$ given $\theta'_{it} = \theta'$ is $P_j(k|A_t\theta' + B_t)$, so that

$$\begin{aligned}\log[P'_j(k|\theta')/P'_j(k-1|\theta')] &= Da_j(A_t\theta' + B_t - b_j + d_{jk}) \\ &= Da'_{jt}(\theta' - b'_{jt} + d'_{jkt}),\end{aligned}$$

where

$$a'_{jt} = A_t a_j = (A'_t)^{-1}a_j,$$

$$b'_{jt} = (b_j - B_t)/A_t = A'_t b_j + B'_t,$$

and

$$d'_{jkt} = d_{jk}/A_t = A'_t d_{jk}.$$

Conventional analysis with programs such as Parscale provides maximum marginal-likelihood estimates $\hat{a}'_{jt}$, $\hat{b}'_{jt}$, and $\hat{d}'_{jkt}$ for a'_{jt}, b'_{jt}, and d'_{jkt}, respectively, for each Item j in J_t for each Administration $t \geq 1$. In typical linking attempts, for each Administration $t > 1$ and each Item j in J_t, a chained approach is used to estimate the parameters A_t, B_t, A'_t, B'_t, a_j, b_j, and d_{jk}. This approach requires availability of sequences of common items. Let K_t be the set of Items j common to both Administration t and to some Administration $s < t$, so that j is in both J_t and J_s. The chained approach can only be used to link all Administrations t for $1 \leq t \leq T$ if K_t is nonempty for $2 \leq t \leq T$. In other words, to each Administration t must correspond an Administration $s < t$ such that these two administrations share common items.

2 Regression Analysis

Even if, for some Administration $t > 1$, the set K_t of common items is empty, it still may be possible to estimate all required parameters by use of a three-stage regression analysis. In the first stage, the item discrimination a_j is estimated for each Item j in J. In the second stage, the item difficulty b_j is estimated for each Item j. In the third stage, the location parameter d_{jk} is estimated for each score k, $0 \leq k \leq r_j - 1$, for each Item j.

In the first stage, assume, as is normally the case, that each estimated item discrimination $\hat{a}'_{jt}$, j in J_t, is positive. The equations

$$\log a'_{jt} = \log A_t + \log a_j$$

lead to the regression model in which $\log \hat{A}_t$ and $\log \hat{a}_j$ are selected to minimize

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\log \hat{a}'_{jt} - \log \hat{A}_t - \log \hat{a}_j]^2$$

subject to the constraint that $\hat{A}_1 = 1$. The minimization problem is commonly encountered in the analysis of variance when an incomplete two-way layout is considered and an additive model is employed in which the variables represented by rows and columns are treated as nominal variables. The T administrations correspond to rows and the v items in J correspond to columns. The layout is incomplete because not all combinations of administrations and items are observed.

Efficient computation of least-squares estimates requires some care due to the very large number of items to be considered. For each Item j, let U_j be the set of Administrations t such that j is in J_t, and let u_j be the number of elements of U_j. For Administrations t and t', let $H_{tt'}$

be the set of Items j such that j is in both J_t and $J_{t'}$ (j is a common item for Administrations t and t'), and let G_t be the set of positive integer $t' \leq T$ such that $H_{tt'}$ is not empty, so that Administrations t and t' share common items. Solution of the normal equations shows that

$$\log \hat{a}_j = u_j^{-1} \sum_{t \in U_j} [\log \hat{a}'_{jt} - \log \hat{A}_t] \tag{1}$$

and

$$\log \hat{A}_t - v_t^{-1} \sum_{t' \in G_t} \log \hat{A}_{t'} \sum_{j \in H_{tt'}} u_j^{-1} = v_t^{-1} \sum_{j \in J_t} \left[\log \hat{a}'_{jt} - u_j^{-1} \sum_{t' \in U_j} \log a'_{jt'} \right] \tag{2}$$

(Scheffé, 1959, p. 114). Clearly $\hat{A}'_t = 1/\hat{A}_t$.

In the second stage, b_j and B_t are estimated by use of the equation

$$b'_{jt} A_t = -B_t + b_j.$$

The estimates $\hat{B}_t$ and $\hat{b}_j$ are selected to minimize

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\hat{b}'_{jt} \hat{A}_t + \hat{B}_t - \hat{b}_j]^2$$

subject to the constraint $\hat{B}_1 = 0$. Again the regression analysis corresponds to an additive model for an incomplete two-way layout. Thus

$$\hat{b}_j = u_j^{-1} \sum_{t \in U_j} [\hat{b}'_{jt} \hat{A}_t + \hat{B}_t] \tag{3}$$

and

$$\hat{B}_t - v_t^{-1} \sum_{t' \in T} \hat{B}_{t'} \sum_{j \in H_{tt'}} u_j^{-1} = -v_t^{-1} \sum_{j \in J_t} \left[\hat{b}'_{jt} \hat{A}_t - u_j^{-1} \sum_{t' \in U_j} \hat{b}'_{jt'} \hat{A}_{t'} \right]. \tag{4}$$

It then follows that $\hat{B}'_t = -\hat{B}_t / \hat{A}_t$.

In the third stage, the $\hat{d}_{jk}$ are selected to minimize

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\hat{d}'_{jkt} \hat{A}_t - \hat{d}_{jk}]^2 = \sum_{j \in J} \sum_{t \in U_j} [\hat{d}'_{jkt} \hat{A}_t - \hat{d}_{jk}]^2.$$

In this case, one can proceed as in the analysis of variance for a one-way layout with independent variable corresponding to items and with replications corresponding to administrations in which the item appears. Thus

$$\hat{d}_{jk} = u_j^{-1} \sum_{t \in U_j} \hat{d}'_{jkt} \hat{A}_t. \tag{5}$$

If $T = 2$, then the regressions reduce to log-mean mean equating (Mislevy & Bock, 1990). The set $K_2 = H_{12}$ consist of the common items j in both J_1 and J_2. Then $\hat{A}_2$ is the geometric mean of the ratios $\hat{a}'_{j2}/\hat{a}'_{j1}$ for j in K_2. In addition, $\hat{B}_2$ is the arithmetic mean of $\hat{b}'_{j1} - \hat{b}'_{j2}\hat{A}_2$ for j in K_{12}. If j is in J_1 but not in J_2, so that Item j is only encountered at Administration 1, then $\hat{a}_j = \hat{a}'_{j1}$, $\hat{b}_j = \hat{b}'_{j1}$, and $\hat{d}_{jk} = \hat{d}'_{jk1}$. If j is in J_2 but not in J_1, so that Item j is only encountered at Administration 2, then

$$\hat{a}_j = \log \hat{a}'_{j2} - \log \hat{A}_2,$$

$$\hat{b}_j = \hat{b}'_{j2}\hat{A}_2 + \hat{B}_2,$$

and

$$\hat{d}_{jk} = \hat{d}'_{jk2}\hat{A}_2.$$

If j is in K_2, so that Item j is a common item, then

$$\log \hat{a}_j = \frac{\log \hat{a}'_{j1} + \log \hat{a}'_{j2} - \log \hat{A}_2}{2},$$

$$\hat{b}_j = \frac{\hat{b}'_{j1} + \hat{A}_2\hat{b}'_{j2} + \hat{B}_2}{2},$$

and

$$\hat{d}_{jk} = \frac{\hat{d}'_{jk1} + \hat{d}'_{jk2}\hat{A}_2}{2}$$

(Kolen & Brennan, 2004, p. 162).

The regression approach successfully defines estimates if, and only if, the set G of pairs $\{(j,t) : j \in J_t, 1 \le t \le T\}$ satisfy the inseparability requirement (Goodman, 1968) used in the case of two-way contingency tables with omitted cells. The inseparability requirement is essentially the requirement that all items can be linked together by use of common items for a sequence of administrations. Thus for each j and j' in J must correspond a positive integer w such that, for Administrations $t(z)$, $1 \le z \le w$, j is in $J_{t(1)}$, j' is in $J_{t(w)}$, and $H_{t(z)t(z+1)}$ is nonempty if z is a positive integer less than w. For a simple example, let $T = 3$, and let J be the set of integers from 1 to 5. Let $J_1 = \{1,2,3\}$, let $J_2 = \{2,3,4\}$, and let $J_3 = \{3,4,5\}$. Then G is inseparable. For example, consider $j = 1$ and $j' = 5$. Then for $w = 2$, $t(1) = 1$ and $t(2) = 3$, j is in $J_{t(1)}$, j' is in $J_{t(3)}$, and $H_{t(1)t(2)} = \{3\}$ is not empty. On the other hand, let $T = 3$, let J be the set of integers from 1 to 7, let $J_1 = \{1,2,3\}$, let $J_2 = \{2,3,4\}$, and let $J_3 = \{5,6,7\}$. The inseparability condition fails, for H_{13} and H_{23} are empty. Thus j in J_1 or J_2 cannot be linked to j' in J_3.

In practice, the problem of minimization of the sums of squares requires some care when the number of items in J is very large. It is desirable that a computer routine apply (2) and (4) to solve $T-1$ simultaneous equations rather than use a completely general approach with $T+v-1$ simultaneous equations to solve, $T-1$ equations for the $T-1$ administrations other than the initial administration and v equations for the v items in J. In SAS, the GLM procedure may be applied with the ABSORB option for the variable that specifies administrations as long as the data are sorted by order of administration. Elementary procedures are then needed to apply (1), (3), and (5), for GLM does not obtain $\log \hat{a}_j$, $\hat{b}_j$, and $\hat{d}_{jk}$ if the ABSORB option is used.

A change in the definition of the base form has no real impact on results. Let the base form be Form s for some positive integer $s \leq T$. The proficiency parameter θ_{it} is converted to the proficiency parameter $\theta^*_{it} = (\theta_{it} - B_s)/A_s$, the item difficulty b_j is converted to $(b_j - B_s)/A_s$, the item discrimination a_j is converted to $a^*_j = A_s a_j$, and the location d_{jk} becomes $d^*_{jk} = d_{jk}/A_s$. Thus

$$a^*_j(\theta^*_{it} - b^*_j + d^*_{jk}) = a_j(\theta_{it} - b_j + d_{jk}).$$

The mean of θ^*_{it} is $B^*_t = (B_t - B_s)/A_s$, and the standard deviation of θ^*_{it} is $A^*_t = A_t/A_s$. Consider minimization of

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\log \hat{a}'_{jt} - \log \hat{A}^*_t - \log \hat{a}^*_j]^2$$

subject to the constraint that $\hat{A}^*_s = 1$, minimization of

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\hat{b}'_{jt} \hat{A}^*_t + \hat{B}^*_t - \hat{b}^*_j]^2$$

subject to the constraint $\hat{B}^*_s = 0$, and minimization of

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\hat{d}'_{jkt} \hat{A}^*_t - \hat{d}^*_{jk}]^2.$$

Then $\hat{A}^*_t = \hat{A}_t/\hat{A}_s$, $\hat{a}^*_j = \hat{A}_s \hat{a}_j$, $\hat{B}^*_t = (\hat{B}_t - \hat{B}_s)/\hat{A}_s$, $\hat{b}^*_j = (\hat{b}_j - \hat{B}_s)/\hat{A}_s$, and $\hat{d}^*_{jk} = \hat{d}_{jk}/\hat{A}_s$.

3 Conclusion

The approach proposed in this report is readily applied even when the linkages between forms are quite complex. Common software programs such as Parscale and SAS may be used in calculations. If outliers are a concern, then standard methods of residual analysis may be employed to identify unusually large residuals from each required regression analysis (Draper & Smith, 1998). If needed, it is possible to remove selected items from the equating computations. Because the estimated difficulty $\hat{b}'_{jt}$ is relatively unstable if the estimated discrimination a'_{jt} is unusually small, it may also be desirable to remove items from equating if $\hat{a}'_{jt}$ is unusually small, say less than 0.2.

The proposed approach does satisfy the basic requirement that all parameters are estimated with increasing accuracy if sample sizes become large at each administration and if all model assumptions are satisfied. The approach does not exploit information concerning accuracy of parameter estimates or correlations of parameter estimates, so it is not fully efficient in a statistical sense. More efficient approaches require more computational resources than are typically available in operational programs.

Once the estimates $\hat{A}_t$, $\hat{B}_t$, $\hat{a}_j$, $\hat{b}_j$, and $\hat{d}_{jk}$ are available, a variety of methods can be employed to equate scores from different administrations (Kolen & Brennan, 2004, ch. 6). In the case of IRT true-score equating, let scores be provided for Administration t based on the subset F_t of items, where F_t is included in J_t. Thus Examinee i has a raw score

$$S_{it} = \sum_{j \in F_t} X_{ijt}$$

with range from 0 to $s_t = \sum_{j \in F_t}(r_j - 1)$. The test characteristic curve for Administration t is estimated to be

$$\hat{T}_t(\theta) = \sum_{j \in G_t} \sum_{k=1}^{r_j} (k-1)\hat{P}_{jk}(\theta),$$

where

$$\log[\hat{P}_j(k|\theta)/\hat{P}_j(k-1|\theta)] = D\hat{a}_j(\theta - \hat{b}_j + \hat{d}_{jk}).$$

The raw score of 0 for Administration t is converted to a raw score of 0 for Administration 1, while the raw score of s_t for Administration t is converted to a raw score of s_1 for Administration 1. For $0 < s < s_t$, a raw score of s for Administration t is converted to a raw score of $\hat{T}_1(\hat{T}_t^{-1}(s))$ for Administration 1.

The approach applied to the two-parameter logistic model and general partial credit model in this report applies with little change if other models are used. For example, it is a simple matter to modify analysis for use with a three-parameter logistic model or for a two-parameter normal ogive model. Note that in the case of a three-parameter logistic model, an added stage would be employed in which the guessing parameter for an Item j would be averaged over the reported values for all forms in which it appears.

References

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.

Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association*, *63*, 1091–1131.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer-Verlag.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.