Using SAS® to Implement Simultaneous Linking in Item Response Theory

Lili Yao, Shelby Haberman and Jun Xu, Educational Testing Service

ABSTRACT

In item-response theory (IRT), item parameters estimated from examinee responses characterize performance of items from a test form used in administration of an educational assessment (Hambleton, Swaminathan, & Rogers, 1991, ch. 1). These parameters are specific to the proficiency distribution of the examinees for that administration. When an assessment must be administered at many different times, many different test forms must be prepared. Typically the proficiency distributions of examinees differ for tests administered at different times, so that item parameters for different test forms are not comparable. Achievement of comparability relies on items common to more than one test form. These items permit linkage of the parameters for all test forms so that these parameters apply to a common proficiency distribution of test takers. Simultaneous linking (Haberman, 2009) uses regression analysis to perform the required linking. This approach was originally applied to the generalized partial credit model (Muraki, 1997) for tests that measure only one skill. This paper provides application of simultaneous linking to a between-item version (Adams, Wilson, & Wang, 1997) of the generalized partial credit model. This between-item model is a special case of multi-dimensional IRT (Haberman, 2013). SAS provides an efficient approach to implementation of simultaneous linking. SAS contributes here not only to solution of large regression problems but also to needed automation of the analysis of the many test administrations. For each administration, item identifiers must be retrieved, control statements must be produced for analysis of responses of test takers to all test items they receive, test data must be placed in a format suitable for an external Fortran program for item-response analysis, and program results must be entered into SAS. Given the large number of forms, macros in SAS must be developed to automate the required analysis. After all forms have been examined, linkage of items requires solution of two very large least-squares problems. The ABSORB option of the GLM (Generalized Linear Models) procedure provides an efficient approach to calculation of the desired regression estimates. Linking items from a real assessment with a very large number of administrations illustrates use of SAS in simultaneous linking.

INTRODUCTION

When educational assessments seek to measure multiple skills by use of separate test sections for each skill, between-item models (Adams, Wilson, & Wang, 1997) can be an appropriate multi-dimensional item-response model for analysis of the examinee responses to the items in the assessment. When these assessments have major consequences for examinees and test-takers can take the tests at many different times, reduction of security risks requires that many different forms should be used and that form content must be difficult to predict prior to the time of test administration. To ensure fairness, these test forms must be linked to adjust test scores to reflect inevitable variations in difficulty. The requirements of security and fairness together lead to complex linkage of forms in which selected items appear in more than one form in an unpredictable manner. In this paper, regression analysis is used to link simultaneously a large number of forms for an assessment which measures multiple skills and satisfies conditions for use of a between-item model for item responses. If only two forms are involved, then the procedure for simultaneous linking resembles log-mean mean linking (Mislevy & Bock, 1990).

The GLM procedure in SAS for general linear models may be used to implement the proposed least-squares analysis. When thousands of items are found in the test forms to be linked, the ABSORB option in GLM procedure greatly simplifies computations. Given the large number of forms, macros in SAS must be developed to automate the required analyses to provide input for the GLM procedure and to use output from GLM.

To illustrate the proposed method, data are used from 62 administrations from a testing program that measures 4 skills. Only 2 skills are directly linked by common items. In all, 8,836 distinct items appear in these administrations, and about 540,000 examinees are in the sample. This example effectively illustrates feasibility of simultaneous linking even with a test with a complex design.

The paper is organized as follows. Description of the model used appears in the second section. In the third section, regression analysis provides a method to link items. In the fourth section, the illustrative example demonstrates application of the SAS GLM procedure for simultaneous linking. In the fifth section, examples are provided of SAS code employed. Concluding remarks appear in the last section.

THE BETWEEN-ITEM MIRT MODEL

In educational testing, item-response theory provides a method to link many distinct test forms from many different administrations when no test taker receives more than a small fraction of the test items encountered in the entire collection of tests. In the tests considered in this paper, a test may consist of several sections, where each section assesses a different skill. In item-response theory, it is assumed that examinee performance on the test reflects an unobserved latent vector with one element for each skill tested. Each element of the latent vector measures proficiency on a single skill. In the between-item model employed, the latent proficiency vector has a nonsingular multivariate normal distribution with unknown mean and unknown covariance matrix. This vector may have a different distribution in each test administration. Conditional on the latent vector, the responses of a test taker to different test items are independent. To permit parameter identification, for some base administration, it is assumed that the latent variable associated with a given skill has mean 0 and conditional variance of 1 given any other latent variables. These conventions reflect the MIRT program (Haberman, 2013) employed to estimate model parameters.

The generalized partial credit model (Muraki, 1997) and the between-item multidimensional IRT model (Adams et al., 1997) determine the relationship of an item response associated with a specific skill and the latent variable associated with that skill. Each item has a finite number of nonnegative integer-valued item scores. The item corresponds to an unknown real item discrimination parameter α and a vector of item threshold parameters β_x, where x is a positive integer no greater than the maximum possible score for the item. For each positive item score x, given the latent vector for a test taker, the logarithm of the ratio of the conditional probability that the item score is x and the conditional probability that the item score is x-1 is a linear function of the latent proficiency for the skill associated with the item. The linear function has slope (item discrimination parameter) α and intercept (item threshold parameters) β_x.

Linking test forms requires the assumption that the relationship of an item response to the latent vector does not depend on the form used or the administration. Under this assumption, linking relies on common items, namely items that appear in more than one test administration. If common items are properly arranged for a specific skill, then the available information permits identification of all item parameters associated with that skill. In addition, for each administration, the available information identifies the mean of the associated latent variable and the conditional variance of that variable given the other latent variables.

SIMULTANEOUS LINKING

In very complex cases, computational constraints dictate calculation of separate estimates (calibrations) of item parameters and correlations of latent variables for each administration. These computations result in estimated item parameters not on a consistent scale, for identification of item parameters in each calibration requires the assumptions that 0 is the mean of each latent variable and 1 is the conditional variance of each latent variable given the other latent variables. For each administration, the original latent variable θ for a skill is standardized to yield an adjusted latent variable $\theta' = (\theta - \mu)/\sigma$, where μ and σ are, respectively, the mean and conditional standard deviation of θ for the administration. The conditional standard is for θ given the other latent variables. In terms of θ', the corresponding adjusted

item parameters are the adjusted item discrimination $a' = \sigma a$ and the adjusted item threshold $\beta'_x = \beta_x + a\mu$. To link estimated adjusted item parameters from separate calibrations to provide estimated item parameters on a consistent scale requires a statistical procedure. Traditional approaches (Stocking & Lord, 1983) are based on methodology for linking two test forms. In contrast, simultaneous linking (Haberman, 2009) provides an approach to this problem that treats all items and administrations at once by use of least squares. Initial applications of simultaneous linking have treated cases in which only one skill appears. In this paper, the approach employed applies to tests with multiple skills. The new approach, which relies on item threshold parameters instead of item difficulty parameters, is less sensitive than the older approach to items with low discrimination.

Simultaneous linking involves 2 large linear regressions. In the first, the predicted variables is the logarithm $\log \hat{\alpha}'$ of the observed estimated adjusted item discrimination $\hat{\alpha}'$, and the predictors are the logarithm $\log \hat{\alpha}$ of the estimated item discrimination $\hat{\alpha}$ and the logarithm $\log \hat{\sigma}$ of the estimated conditional standard deviation $\hat{\sigma}$. The constraint is added that $\hat{\sigma}$ is 1 for the base administration. The second regression has predicted variable the observed estimated adjusted item threshold $\hat{\beta}'_x$ and predictors the estimated item threshold $\hat{\beta}_x$ and the product $\hat{a}\hat{\mu}$ of the previously-found estimated item discrimination $\hat{a}$ and the estimated mean $\hat{\mu}$. The constraint is added that $\hat{\mu}$ is 0 for the base administration. In the end, all estimated item parameters and all estimated means and conditional standard deviations are obtained.

In analysis of variance, these regressions are associated with additive models for incomplete two-way layouts. The items correspond to rows, and the administrations correspond to columns. The layout is incomplete because not all combinations of administrations and items are observed. The standard criterion R^2 from regression analysis provides one measure of effectiveness of the proposed linking approach.

The GLM procedure in SAS provides a convenient tool to implement these two regressions. In GLM, the ABSORB option applies to the variable that specifies item code (accession ID) and the variable that specifies the category of the item. Data must be sorted by item code and item category. To treat the large number of test forms, SAS macros automate the required separate calibrations and processing of GLM results to yield estimated item parameters.

SAS CODES

In the interest of space, SAS codes are only provided for the use of the GLM procedure. For more detail, contact the first author at lyao@ets.org. It is assumed that SAS macros have already been used to run the MIRT program (Haberman, 2013) for the separate calibration of each administration and to transfer the item parameters produced by the program to SAS data sets. In the first application of GLM, the data set test.sortslope contains an entry for each item to be linked in each administration. The associated information is the administration code (Adm), the associated skill (Skill), the item code (Accession), the estimated adjusted item discrimination $\hat{\alpha}'$(Slope), and the logarithm $\log \hat{\alpha}'$ of the estimated adjusted item discrimination (Loga). The data have been sorted by item code. Administration codes, accession codes, and skills are modified to avoid providing information about the testing program not needed for the example. Only the two skills directly linked are included in the file. The first two observations are displayed in Table 1.

Table 1. The first two observations in the input file for regression analysis for the slopes.

Adm	Skill	Accession	Slope	Loga
Admin4	Skill2	Item1	0.3465457889	-1.059740322
Admin4	Skill2	Item2	0.45739793	-0.782201523

The following code for GLM is used for the regression of log estimated adjusted item discriminations on log estimated item discriminations and log estimated conditional standard deviations.

```
/* Slope analysis */
ods output "Solution"=test.glmslope;/* Output SAS file for estimated log
                                      standard deviations*/
proc glm data=test.sortslope;          /* GLM procedure for the item
                                      discriminations */
absorb accession;                    /* Absorb option for the accession Ids */
class adm skill;  /* Adm and Skill are categorical variables */
model loga=adm*skill/solution;  /* Set up the glm procedure for loga.
                                   The adm*skill specification leads to
                                   different standard deviations for the
                                   same administration for different skills.
                                   The option solution results in output of
                                   logarithms of estimated conditional
                                   standard deviations. */

run;
ods output close;
```

The file test.glmslope includes the variables Dependent, Parameter, Estimate, Biased, StdErr, tValue, and Probt. Only Parameter and Estimate are used in the analysis. For an administration Adm and skill Skill, the variable parameter is a character variable of the form adm*skill Adm Skill. The variable Estimate is the estimated logarithm $\log \hat{\sigma}$. Table 2 provides results for the administration Admin1 and skills Skill1 and Skill2 for Parameter and Estimate,

Table 2. The first two observations in the output file for regression analysis for the slopes.

Parameter	Estimate
adm*skill Admin1 Skill1	0.0228834893
adm*skill Admin1 Skill2	0.1980624561

SAS macros are used to convert the content of test.glmslope to estimated conditional standard deviations and to find the estimated item discriminations. A file test.bdata is constructed with the accession code (Accession), the skill (Skill), the estimated item discrimination $\hat{\alpha}$ (A), the logarithm $\log \hat{\alpha}$ of the estimated item discrimination (Loga), the administration code (Adm), the estimated adjusted item threshold $\hat{\beta}'_x$ (Intercept), and the item category x (Cat). The first two observations are displayed in Table 3.

Table 3. The first two observations in the input file for regression analysis for the intercepts.

Adm	Skill	Accession	A	Loga	Intercept	Cat
Adm 1	Skill 2	Item 1	0.7321378387	-0.31178647	1.8730353122	1
Adm 2	Skill 2	Item 2	0.3947289439	-0.92955597	0.3934502256	1

The following code for GLM is used for the regression of log estimated adjusted item discriminations on adjusted estimated item threshold and predictors are the estimated item threshold and the product of the previously-found estimated item discrimination and the estimated mean.

```
/* Intercept analysis */
ods output "Solution"=test.glmint; /* Results go to SAS data set
                                      test.glmint.*/
proc glm data=test.bdata;    /* The GLM procedure for the item thresholds */
absorb accession cat;           /* The Absorb options is for the accession
                                   Ids and Cat */
class adm skill;                /* Adm and Skill are categorical variables */
model intercept=a*skill*adm/solution;/* Set up the glm procedure for
                                        the item thresholds. The
```

```
                                        a*skill*adm specification leads to
                                        different means for the same
                                        administration for different skills and
                                        indicates that means are multiplied by
                                        estimated item discriminations (a). The
                                        option solution leads to output of
                                        estimated means. */

run;
ods output close;
```

The file test.glmint includes the variables Dependent, Parameter, Estimate, Biased, StdErr, tValue, and Probt. Only Parameter and Estimate are used in the analysis. For an estimated item discrimination (a), an administration Adm and skill Skill, the variable parameter is a character variable of the form a*adm*skill Adm Skill. The variable Estimate is the estimated mean $\hat{\mu}$. Table 4 provides results for the first administration for Parameter and Estimate,

Table 4. The first two observations in the output file for regression analysis for the intercepts.

Parameter	Estimate
a*adm*skill Admin1 Skill1	-0.202007902
a*adm*skill Admin1 Skill2	-0.192567947

SAS macros are used to find estimated item thresholds.

EXAMPLE

The testing program examined assesses 4 skills, but common items are only present for the first two skills. For the first skill, items are almost always dichotomous. For the second skill, more than 90% of items are dichotomous but cases with 3 or 4 item scores do occur. For the third skill, all items have 5 item scores. For the last skill, items have 10 scores.

For the logarithms of estimated item discriminations, R^2 = 0.987 and adjusted R^2 = 0.928 for the first skill, while R^2 = 0.988 and adjusted R^2= 0.944 for the second skill. These values are quite high even in the case of adjusted R^2. In the case of estimated item thresholds, R^2= 0.993 and adjusted R^2= 0.960 for the first skill, and R^2= 0.996 and adjusted R^2= 0.979 for the second skill. The higher R^2 statistics associated with threshold parameters may reflect their higher variability relative to log item discriminations. In summary, least square is very effective in placing item parameters for different administrations on a common scale despite the complexity of the linkage.

CONCLUSIONS

This paper demonstrates use of SAS to implement simultaneous linking for separate calibrations of test forms by between-item models. Use of linear least squares in simultaneous linking contributes greatly to the practicality of the approach when very large numbers of items and administrations must be linked, for the ABSORB option permits employment of the GLM procedure in SAS. Use of an example with 62 administrations and four skills, only two of which are directly linked, shows that simultaneous linking provides a practical approach to very difficult linking problems.

REFERENCES

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). *The multidimensional random coefficients*

multinomial logit model. Applied Psychological Measurement, 21, 1-23. doi: 10.1177/0146621697211001

Haberman, S. J. (2009). Linking parameter estimates derived from an item response model through separate calibrations (Research Rep. No. RR-09-40). Princeton, NJ: ETS. doi: 10.1002/j.2333-8504.2009.tb02197.x

Haberman, S. J. (2013). A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm (Research Report Series No. RR-13-32). Princeton, NJ: ETS. doi: 10.1002/j.2333-8504.2013.tb02339.x

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

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. doi: 10.1007/978-1-4757-2691-6_9

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 2, 201-210. doi: 10.1177/014662168300700208

ACKNOWLEDGEMENT

Any opinions expressed in this publication are those authors and not necessarily of Educational Testing Service.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the first author at

Lili Yao
Educational Testing Service
6097345450
lyao@ets.org

TRADEMARKS