ED 464 143 TM 033 826

AUTHOR	de la Torre, Jimmy; Patz, Richard J.
TITLE	Item Response Theory Equating Using Bayesian Informative Priors.
PUB DATE	2001-04-11
NOTE	11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Achievement Tests; *Bayesian Statistics; *Equated Scores; *Estimation (Mathematics); *Item Response Theory; *Junior High School Students; Junior High Schools; Markov Processes; *Mathematics Tests; Maximum Likelihood Statistics; Monte Carlo Methods; Test Items
IDENTIFIERS	Calibration; Comprehensive Tests of Basic Skills; Item Parameters

ABSTRACT

This paper seeks to extend the application of Markov chain Monte Carlo (MCMC) methods in item response theory (IRT) to include the estimation of equating relationships along with the estimation of test item parameters. A method is proposed that incorporates estimation of the equating relationship in the item calibration phase. Item parameters from a previous calibration of one test form are used to construct informative prior distributions to be used in the new simultaneous calibration of the two test forms. Data were from the standardization of the Comprehensive Test of Basic Skills, Fifth Edition for an eighth-grade mathematics test, with 3,171 examinees used to obtain initial estimates for the anchor test (set A) parameters, 2,000 examinees in the equating process, and 4,000 examinees used to evaluate the effectiveness of the various methods. The new method was compared to traditional methods based on marginal maximum likelihood calibration followed by a Stocking and Lord (M. Stocking and F. Lord, 1983) linear transformation. Results indicate that the new approach can lead to modest improvement in equating accuracy. Under this approach, the predicted scores for a validation group have higher correlations and lower root mean square errors in comparison to observed scores. (SLD)

Item Response Theory Equating Using Bayesian Informative Priors[1]

Jimmy de la Torre
University of Illinois Urbana Champaign
jdelator@s.psych.uiuc.edu

Richard J. Patz
CTB/McGraw-Hill
rpatz@ctb.com

April 11, 2001

Introduction

A variety of estimation procedures, both maximum likelihood and Bayesian in nature, have been used for parameter estimation in item response theory (IRT; Lord, 1980) settings. There are also a wide variety of test equating methods in use for placing different test forms on common scales. Calibration (i.e., item parameter estimation) and equating are typically conducted in sequence, as part of a "divide-and-conquer" approach to making inferences from educational assessments. After item parameters are estimated, they are treated as fixed and known for the purpose of deriving equating relationships (see, for example, Patz and Junker, 1999b). More comprehensive and unified analyses using item response theory models has been greatly facilitated with the development of Markov chain Monte Carlo (MCMC) estimation techniques (e.g., Albert, 1992; Patz and Junker, 1999a, 1999b; Patz, Junker, and Johnson, in press).

This paper seeks to extend the application of MCMC methods in IRT to include the estimation of equating relationships along with the estimation of test item parameters. A very common method for IRT equating is based on linear transformation that minimizes expected differences in test characteristic curves between the original ("anchor") calibration and the newly estimated item parameters (Stocking and Lord, 1983). This common approach falls under the "divide-and-conquer" category described above, in that item parameters are treated as fixed and known when the linear transformation constants are estimated. A drawback of this approach is that the combined uncertainty attributable to both item parameter estimation and equating parameter estimation is difficult to assess and control.

[1] Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 2001, Seattle, WA

In this paper we propose and explore an alternative method that incorporates estimation of the equating relationship in the item calibration phase. Item parameters from a previous calibration of one test form are used to construct informative prior distributions to be used in the new simultaneous calibration of the two test forms. We implement the approach using Markov chain Monte Carlo estimation techniques (Patz and Junker, 1999b), and we examine its effectiveness using equating data from a national standardization of an eighth grade mathematics test. We compare our results to those obtained using traditional equating techniques.

Data

We examine our calibration and equating methods using data collected from the standardization of the CTB/McGraw-Hill's Comprehensive Test of Basic Skills, Fifth Edition (CTBS/5 TerraNova). In particular, an eighth-grade mathematics form is examined. The long form ("Complete Battery") of this test may be logically divided up into two subsets of test items—a 31-item operational survey (Set A) and an additional set of 25-items (Set B). We treat Set A as the anchor test and Set B as test to be equated to have same metric as Set A.

A total of 9,171minees completed this long test form as a part of the CTBS/5 standardization process. For the purpose of this paper, we randomly divided these examinees into 3 groups. In calibrating the anchor set, a group of 3171 examinees were used to obtain initial estimates of the Set A item parameters. Another group of 2000 examinees were used in the equating process. Finally, a separate group of 4,000 examinees was used to evaluate the effectiveness of the various methods.

Procedure

The 56 test items were of multiple-choice type format with 4 options. Given the test format, the three-parameter logistic model was deemed appropriate. Two methods of marginal maximum likelihood (MML) estimation were employed in the calibration stage: EM and MCMC. The EM estimates were obtained using the software PARDUX (Burket, 1998). The code for the MCMC estimation was written in S-Plus, and is a minor variant of the algorithm and software of Patz and Junker (1999b). To obtain approximate MML item parameter estimates under the Bayesian MCMC scheme, flat prior distributions for item parameters were used, and Metroplis-Hastings candidate distributions yielding efficient acceptance rates were used. These are displayed in Table 1. The posterior distributions of the parameters were estimated from the last 24,000 of 25,000 iterations, and posterior means were used as approximate modal parameter estimates.

Table 1: MCMC prior and candidate distributions in the calibration stage

Parameter	Prior	Candidate Distribution
θ	N(0,1)	$N(\theta_0, 0.5)$
α	Uniform	$\exp(N(\alpha_0, 0.003))$
β	Uniform	$N(\beta_0, 0.003)$
γ	Uniform	$\mathrm{Beta}(800\gamma_0, 800(1-\gamma_0))$

In equating Set B to Set A, three methods were investigated: 1) MML estimation followed by Stocking and Lord (1983) linear transformation (MML-SL); 2) Bayesian estimation utilizing informative prior distributions for anchor parameters (BIP); and 3) the Bayesian informative prior approach followed by Stocking and Lord linear transformation (BIP-SL). Each of these three equating methods were applied using MML-estimated anchor parameters from both the MCMC and EM approaches described above. Examinees used for this equating step were 2000 examinees who were not utilized in the calibration step for anchor parameters.

The MML-SL method was performed using PARDUX. The procedure involves simultaneous estimation of item parameters for both Set A and Set B, followed by a linear transformation that achieves minimal weighted discrepancy between the newly estimated Set A TCC and the Set A TCC based on the anchor parameters (Stocking and Lord, 1983). The transformation constants when applied to Set B put these item parameters in the metric of the anchor parameters.

The Bayesian method also involves simultaneous estimation of item parameters for both Set A and Set B, but using informative rather than flat prior distributions for the Set A item parameters. Flat priors were used for Set B item parameters. For the candidate distributions, the same control parameters used in calibration stages were employed. Table 2 lists the prior and candidate distributions used in the Bayesian procedures.

Table 2: Bayesian prior and candidate distributions in the equating stage

	Parameter	Prior	Candidate Distribution
Set A	θ	$N(0,1)$	$N(\theta_0,0.5)$
	α	$\exp(N(\alpha^*,0.1))$	$\exp(N(\alpha_0,0.003))$
	β	$N(\beta^*,0.25)$	$N(\beta_0,0.003)$
	γ	$\text{Beta}(200\gamma^*,200(1-\gamma^*))$	$\text{Beta}(800\gamma_0,800(1-\gamma_0))$
Set B	θ	$N(0,1)$	$N(\theta_0,0.5)$
	α	Uniform	$\exp(N(\alpha_0,0.003))$
	β	Uniform	$N(\beta_0,0.003)$
	γ	Uniform	$\text{Beta}(800\gamma_0,800(1-\gamma_0))$

* indicates parameter as estimated in anchor calibration.

The Markov chains used for estimation consisted of 25,000 iterations, the last 24,000 of which were used in determining the posterior distributions of the item parameters. Similar to the calibration stage, the items parameters were estimated using the posterior means. To investigate whether further improvements in equating can be achieved, Stocking and Lord transformation constants between anchor parameters and BIP equating estimates were also computed.

In the final stage of the study, the quality of the equating relationship derived from each combination of the two calibration methods and three equating methods was examined. Quality of equating was measured in terms of observed score prediction accuracy (Set B to Set A) for a new set of 5000 examinees. In this validation process, expected a posteriori proficiency estimates were obtained given observed responses on Set B and item parameters estimated for Set B under each equating method. To arrive at more precise EAP ability estimates, the first two moments of the prior distribution of ability, which is assumed to be normal, were empirically determined following the method described by Mislevy and Bock (1982). With the Set B-estimated abilities and the Set A anchor parameters, observed scores on Set A were predicted for each examinee. The accuracy of the prediction was assessed by comparing the predicted scores and observed scores using Pearson's correlation and root mean square error.

Results

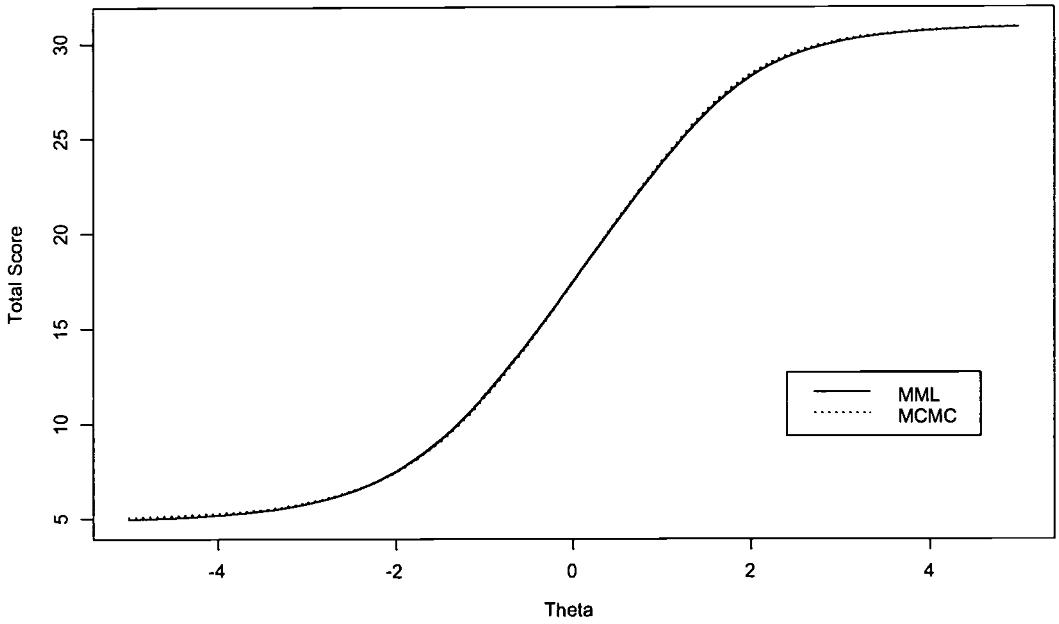
Calibration Stage

The MML anchor parameter estimates obtained under EM and MCMC algorithms are presented in table 3. The correlation coefficients between the estimates are 0.997 (α), 0.994 (β), and .901 (γ). Figure 1 shows that as far as the test characteristic curves are concerned, the two sets of estimates are almost identical.

Table 3: Item Parameter Estimates of the Anchor Set

	MML-EM Estimates			MML-MCMC Estimates		
Item	α	β	γ	α	β	γ
1	1.00	0.05	0.35	1.03	0.05	0.35
2	0.84	0.30	0.44	0.83	0.26	0.43
3	1.16	-0.05	0.14	1.19	-0.04	0.15
4	0.78	-0.73	0.00	0.82	-0.63	0.05
5	0.54	0.10	0.00	0.58	0.21	0.05
6	0.63	0.67	0.13	0.64	0.65	0.13
7	1.00	0.79	0.31	1.00	0.77	0.30
8	0.69	0.58	0.18	0.69	0.55	0.17
9	1.05	-0.52	0.13	1.07	-0.52	0.12
10	1.10	0.22	0.26	1.13	0.22	0.26
11	0.54	-1.31	0.20	0.52	-1.61	0.07
12	1.17	-0.80	0.07	1.20	-0.78	0.08
13	0.74	-0.98	0.05	0.79	-0.87	0.11
14	0.48	-0.15	0.00	0.54	0.07	0.08
15	1.02	-1.54	0.20	0.98	-1.71	0.05
16	1.00	-1.44	0.00	1.05	-1.32	0.09
17	0.66	-0.88	0.00	0.71	-0.73	0.07
18	0.45	-0.41	0.00	0.49	-0.19	0.07
19	1.21	-0.90	0.20	1.23	-0.89	0.20
20	1.24	1.45	0.12	1.29	1.43	0.12
21	1.33	0.26	0.27	1.35	0.26	0.27
22	0.90	-0.24	0.13	0.92	-0.22	0.14
23	0.33	0.78	0.18	0.36	0.77	0.18
24	0.65	0.34	0.14	0.67	0.34	0.14
25	0.75	-0.26	0.30	0.75	-0.31	0.28
26	1.48	0.88	0.12	1.51	0.87	0.13
27	0.93	0.57	0.17	0.94	0.56	0.17
28	0.78	1.02	0.24	0.79	1.00	0.23
29	1.29	1.46	0.09	1.31	1.45	0.09
30	1.41	1.49	0.22	1.45	1.47	0.22
31	0.76	1.68	0.17	0.76	1.67	0.17
Mean	0.90	0.08	0.15	0.92	0.09	0.16
SD	0.30	0.90	0.11	0.30	0.89	0.09

Figure 1: Test characteristic curves from the anchor parameter estimates Using EM and MCMC



Equating Stage

Table 4: Mean and standard deviation of the parameter estimates of Set A using different calibration and equating methods

Calibration Method		Equating Method					
		MML-SL		BIP		BIP-SL	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
MML-EM	α	0.96	0.34	0.93	0.34	0.93	0.34
	β	0.12	0.89	0.06	0.91	0.07	0.91
	γ	0.18	0.10	0.16	0.11	0.16	0.11
MML-MCMC	α	0.98	0.35	0.94	0.34	0.95	0.35
	β	0.12	0.87	0.08	0.91	0.08	0.90
	γ	0.18	0.10	0.16	0.10	0.16	0.10

Table 4 shows the mean and standard deviation for Set A item parameter estimates obtained under the six calibration/equating conditions. Comparing table 4 and table 3, we note that, not surprisingly, BIP equating yields Set A item parameter estimates from the equating group that are more similar to the parameter estimates in the calibration group. Additional evidence that equated parameter estimates using BIP are closer to the anchor parameters is given by the root mean square differences listed on Table 5. This tables shows that the root mean square differences are smaller for MCMC equating and this is true for all parameters. Using the S-L transformation on the MCMC estimates does not appear to have a notable, systematic effect on these root mean square differences.

Table 5: Root mean square differences between the anchor and equating parameter estimates of Set A using different calibration and equating methods

		Equating Method		
Calibration Method		MML-SL	BIP	BIP-SL
	α	0.165	0.126	0.127
MML-EM	β	0.155	0.134	0.132
	γ	0.070	0.048	0.048
	α	0.173	0.128	0.132
MML-MCMC	β	0.155	0.073	0.070
	γ	0.070	0.012	0.012

Table 6: Mean and standard deviation of the parameter estimates of Set B using different calibration and equating methods

		Equating Method					
Calibration Method		MML-SL		BIP		BIP-SL	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
	α	1.03	0.37	1.00	0.40	1.00	0.41
MML-EM	β	0.26	0.81	0.08	0.85	0.09	0.85
	γ	0.21	0.06	0.14	0.10	0.14	0.10
	α	1.05	0.38	1.00	0.40	1.01	0.41
MML-MCMC	β	0.25	0.79	0.08	0.84	0.08	0.84
	γ	0.21	0.06	0.14	0.11	0.14	0.11

The means and standard deviations of the item parameter estimates for Set B are presented in Table 6. In general, the parameter estimates are higher mean but lower standard deviation when equating is carried out using MML-SL. The discrepancy in the mean estimate is most obvious for the difficulty parameter while the difference in the spread is most apparent for the guessing parameter. The S-L transformation has only a very small effect on the mean and variance of the BIP estimates.

Validation Stage

Table 7: Correlation of the predicted and actual total scores on Set A

		Equating		
		MML-SL	BIP	BIP-SL
Calibration	MML-EM	0.848	0.849	0.849
	MML-MCMC	0.848	0.850	0.850

A first measure of the quality of equating was based on the corrclation between the actual and predicted total score on Set A. Table 7 shows that the highest correlation can be attained when MCMC is used in both the calibration and equating stages. No additional improvement can be observed with the use of the S-L transformation. However, for most practical purposes, all the correlation coefficients can be considered the same.

Table 8: RMSE of the predicted and actual total scores on Set A

		Equating		
		MML-SL	BIP	BIP-SL
Calibration	MML-EM	3.37	3.27	3.27
	MML-MCMC	3.37	3.28	3.27

In addition to the correlation coefficient, the root mean square error between the actual and predicted scores on the test can indicate how well the two sets of test are equated. The root mean square errors in Table 8 suggest that following: a smaller error is obtained when the equating is performed using BIP; the S-L transformation can reduce the error further; and the best result can be obtained by calibrating and equating using MCMC and then linearly transforming the estimates. These results are not apparent when root mean square error is taken at the item level. (See Table 9.) Whereas, the root mean square error using ML-SL equating is about 1.03 times larger compared to the root mean square error using BIP-SL at the test level, it is roughly equal to 1at the item level.

Table 9: RMSE of the predicted and actual item response on Set A

		Equating		
		MML-SL	BIP	BIP-SL
Calibration	MML-EM	0.430	0.429	0.429
	MML-MCMC	0.431	0.430	0.429

Discussion

This paper introduces an approach to equating using Bayesian informative prior distributions, and examines how this method compares to traditional methods based on marginal maximum likelihood calibration followed by a Stocking and Lord (1983) linear transformation. Results indicate that the BIP approach can lead to modest improvement in equating accuracy. Under BIP equating the predicted scores for a validation group have higher correlations and lower root mean square errors in comparison to observed scores. In this context, following BIP equating with a Stocking and Lord transformation appears to have only minimal impact on results.

There are a number of important questions left unexamined in this paper. Foremost, is the identification of optimal prior distributions to be used in the BIP equating procedure. In addition, it is plausible to assume that the BIP procedure might be most useful in contexts where the initial calibration group is similar in nature to the validation group but fundamentally different from the equating group. This might be the case, for example, in a testing program that administers Form A in year 1, Form B in year 2, and establishes the link between Forms A and B in a separate, smaller scale equating study.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs Sampling. *Journal of Educational Statistics,* 17, 251-269.

Burket, G. (1998). PARDUX [computer program]. Unpublished.

Kolen, M.J., Brennan, R.L. (1995). *Test Equating: Methods and Practices*. New York Springer-Verlag.

Lord, F. M. (1980). Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Earlbaum

Mislevy, R.J., & Bock, R.D. (1982) Adaptive EAP estimation of ability in microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.

Patz, R.J., & Junker, B.W. (1999a) A straightforward approach to Markov chain Monte Carlo methods for item response theory. *Journal of Educational and Behavioral Statistics*, 24, 146-178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366.

Patz, R. J., Junker, B. W., & Johnson, M. (in press). The hierarchical rater model. *Journal of Educational and Behavioral Statistics.*

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 7, 201-210.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

TM033826

I. DOCUMENT IDENTIFICATION:

Title: Item Response Theory Equating Using Bayesian Informative Priors

Author(s): Jimmy de la Torre & Richard J. Patz

Corporate Source: University of Illinois at Urbana-Champaign

Publication Date: April 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: JIMMY DE LA TORRE / GRADUATE STUDENT	
Organization/Address: University of Illinois at Urbana-Champaign	Telephone: (217) 265-0384	FAX:
	E-Mail Address: jdelator@s.psych.uiuc.edu	Date: 4/2/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com