Research Report

Cognitive Diagnosis for NAEP Proficiency Data

Xueli Xu

Matthias von Davier

Research & Development

May 2006
RR-06-08

Cognitive Diagnosis for NAEP Proficiency Data

Xueli Xu and Matthias von Davier
ETS, Princeton, NJ

May 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

More than a dozen statistical models have been developed for the purpose of cognitive diagnosis. These models are supposed to extract a much finer level of information from item responses than traditional unidimensional item response models. In this paper, a general diagnostic model (GDM) was used to analyze a set of simulated sparse data and real data from National Assessment of Educational Progress (NAEP) assessments. The purpose of this study was to find out (a) whether the parameters can be recovered for a sparse data matrix in the framework of the GDM and (b) how to estimate group characteristics for large survey data, such as NAEP, in the framework of the GDM. The results of the simulation study show that GDM parameters can be recovered satisfactorily. The GDM under single group and multiple group assumptions were employed to fit NAEP assessment data. The results under these two assumptions and between the GDM and the statistics derived using the operational NAEP model were compared. The findings indicate that it is possible to conduct cognitive diagnosis for NAEP proficiency data.

Key words: General diagnostic model (GDM), model parameters, group characteristics, item responses, large survey data

Acknowledgements

The authors would like to thank Catherine McClellan and Shelby Haberman for helpful suggestions; we also extend our thanks to Sandip Sinharay, Kentaro Yamamoto, Dan Eignor, and Issac Bejar for their reviews and comments.

1. Introduction

Cognitive diagnosis of item response data is a popular topic these days. Compared to traditional unidimensional item response models, cognitive diagnosis models are developed to extract much finer information from item responses than the traditional approach. Once the relationship between cognitive attributes and items is specified, the cognitive diagnosis models describe the functional effect of these items on the inference of individual skill mastery. By developing such models, one can provide information about skill mastery status in multiple dimensions of skills that may help students and teachers to choose appropriate training programs. The cognitive diagnosis models are mainly developed for individual inferences.

With the ambition of extracting more information from item responses, however, the parameter estimates of cognitive diagnosis models may suffer from instabilities because a large number of parameters are estimated simultaneously. Even if the item parameters are assumed to be known, the potentially large number of person parameters based on a limited set of items may reduce the accuracy of individual estimates. In contrast to potentially unreliable individual estimates, reliably estimating group characteristics in terms of marginal skill mastery probabilities might be feasible because information is aggregated across individuals. Thus, the population parameters and the subgroup characteristics, in addition to the item parameters, are the foci in this paper.

Because using cognitive diagnosis models to make inferences about population and subgroup characteristics is a promising approach, this study investigated using such a model to analyze the data from subgroup-oriented educational surveys, such as the National Assessment in Educational Progress (NAEP). A compensatory version of the general diagnostic model (GDM; von Davier & Yamamoto, 2004) is used in this study since this class of models already includes many standard item response theory (IRT) models and their extensions to confirmatory multidimensional IRT models, as well as skill profile models (von Davier, 2005).

This paper is organized as follows. Section 2 introduces the GDM and discusses some of its properties. A simulation study is described in Section 3 to demonstrate the capability of parameter recovery in a sparse data matrix in the framework of the GDM. Section 4 and 5 present the analysis and results for two NAEP assessment data sets. In Section 5, a brief discussion is provided.

2. The General Diagnostic Model and Its Properties

During the last decade, more than a dozen models have been developed for cognitive diagnosis. (For a comprehensive review, see Roussos, 1994; Junker & Sijtsma, 2001.) The GDM (von Davier, 2005) is a model that contains many logistic-type models as special cases. These logistic-type models include latent class models of various types, the compensatory fusion model, and a two-parameter logistic (2PL) model among others (von Davier & Yamamoto, 2004). The rest of this section will focus on the introduction of the GDM and a discussion of its properties.

Let the matrix $Q = \{q_{ik}\}$ specify the correspondence between items and attributes. The entry $q_{ik} = 1$ means that the attribute k is measured by the item i, and $q_{ik} = 0$ otherwise. With $q_{ik} \in \{0, 1\}$, let the number of skill levels be $L_k = m + 1$ and choose skill levels $\alpha_k \in \{(0 - c), (1 - c), \ldots, (m - c)\}$ for some constant c. In the GDM, these skill levels are real valued constants that can be chosen by the users to match their hypothesis or emulate a certain model. The partial credit version of the GDM (von Davier, 2005) for a polytomous response $x \in \{0, 1, 2, \ldots, m_i\}$ is

$$P(X = x|\beta_i, \boldsymbol{\alpha}, \boldsymbol{q_i}, \boldsymbol{\gamma_i}) = \frac{\exp[\beta_{xi} + \sum_{k=1}^{K} x\gamma_{ik}q_{ik}\alpha_k]}{1 + \sum_{y=1}^{m_i} \exp[\beta_{yi} + \sum_{k=1}^{K} y\gamma_{ik}q_{ik}\alpha_k]}.$$

In this model, K attributes are assessed and the latent vector is denoted by $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_K\}$. The dichotomous Q matrix specified above consists of K attribute columns and I item rows. The parameters to be estimated are item parameters (such as β_is and $\boldsymbol{\gamma}_i$s) and population parameters (i.e., the point mass for latent attribute patterns). Solution to the maximum likelihood may be obtained by means of the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981; von Davier & Yamamoto, 2004; Dempster, Laird, & Rubin, 1977; Muraki, 1992), which, since the complete-data density function belongs to the exponential family, is guaranteed to converge to the local maxima of the parameters values. Thus, the estimates of item parameters and population parameters can achieve asymptotic consistency when both the number of examinees and the number of items are large. These properties were also empirically demonstrated in the von Davier (2005) study.

An EM algorithm was developed for the partial credit version of the GDM and implemented in the software *mdltm* (von Davier, 2005). The software provides the users with the following estimates: (a) the posterior distribution of attribute patterns for groups specified in advance, (b) the posterior distribution of attribute patterns for each individual, (c) the estimates of item parameters of interest, and (d) the conditional probability of response categories for each item.

The first two types of estimates are, in fact, the population parameters. The posterior distribution of latent attribute patterns for groups is obtained only when a multiple group assumption is used in the analysis.

3. Simulation Study: Sparse Data Matrix

In the simulated data, both the generating model and the fitted model were the GDM. Forty data sets were generated without missing values, each of which included 2,880 examinees and 36 items and assessed up to four attributes. Each attribute was assumed to have two levels, denoted as 0 and 1. Thus, the examinees were classified in one of $2^4 = 16$ latent attribute patterns. Three degrees of missing response rates were used in this simulation study, 10%, 25%, and 50%. The missing responses were randomly assigned to items for each examinee using the corresponding proportion. For example, in the case where 50% of data was missing, 18 out of 36 items were omitted for each simulated examinee.

For this recovery study, the item parameters of interest included the βs and γs in the GDM, and the population parameters were the point masses for 16 latent attribute patterns. After three overall groups of 40 data sets each with different levels of missing data were produced, the software *mdltm* (von Davier, 2005) was used to fit the data. The recovery of the item parameters and the population parameters were studied by calculating the bias and root mean squared error (RMSE) across the 40 data sets for each of the parameters of interest. For any parameter in this model, $\zeta = (\beta,\ \gamma,$ or the point mass for latent attribute patterns), the bias of the estimate $\hat{\zeta}$ is obtained by $bias(\hat{\zeta}) = \sum_{s=1}^{40} \frac{\hat{\zeta}_s - \zeta}{40}$. The RMSE of $\hat{\zeta}$ follows the form

$$RMSE(\hat{\zeta}) = \sum_{s=1}^{40} (\hat{\zeta}_s - \zeta)^2/40,$$

where s is the index for simulations and $\hat{\zeta}_s$ is the estimate in simulation s.

The recovery for each item parameter in different missing data conditions can be seen in Tables 2 through 7. In particular, Tables 2 and 3 show the bias and RMSE of the estimated parameters in the case where 10% of the data was missing. Tables 4 and 5 give the bias and RMSE of the estimates in the 25% missing condition. The bias and RMSE of the estimates in the 50% missing condition are shown in Tables 6 and 7. The recovery for the point mass of each latent attribute pattern can be seen in Tables 8 through 10 for these three missing conditions, respectively. In particular, the Means column represents the estimates for the point mass, while the Truth column shows the true values for the point mass. The Diff. column in Tables 8 through 10 is the difference between the estimates and the true values, while the RMSE column is the RMSE of the estimates. Table 1 summarizes the recovery by averaging the biases and RMSEs across all the item parameters or across all the population parameters.

Tables 1 through 10 show that the RMSE of the item parameters was consistently higher in the cases where 50% of the data was missing than in the cases where 10% or 25% of the data was missing. However, even the highest observed RMSE in the 50% missing condition is still tolerable. Overall, the largest RMSE is 0.23 for the item parameter estimates, which happened in the case where 50% of the data is missing.

The tables also show that the RMSE for the point mass of the latent attribute patterns is very small for all levels of missing data. This finding means that even with a severely sparse data (the 50% missing condition), the missing data does not have a significant effect on the estimates of the skill distribution in the population. This suggests it is feasible to consider estimating the population parameters in more sparse data matrices, such as those exhibited in NAEP data.

Table 1.

Summary of the Accuracy of Estimates

		10% missing (32 items left)	25% missing (25 items left)	50% missing (18 items left)
Item parameters	Average bias	0.001	0.002	0.005
	Average RMSE	0.071	0.083	0.119
Population parameters	Average bias	0.000	0.000	0.000
	Average RMSE	0.004	0.004	0.007

Table 2.
Bias of Slopes and Difficulties for 10% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.006	†	-0.0093	0.0022
2	†	-0.0226	0.0005	†	0.0002
3	†	-0.0086	0.0099	†	-0.0049
4	-0.0198	†	†	0.0155	0.0048
5	†	-0.0101	-0.0032	†	-0.0015
6	0.009	†	†	0.0063	-0.0122
7	†	†	†	†	-0.0016
8	0.0053	†	†	0.0007	-0.0149
9	0.0063	-0.0359	0.012	†	0.005
10	†	0.0065	0.0026	0.0048	0.0023
11	†	†	†	†	-0.0117
12	†	-0.0139	-0.0032	†	-0.0055
13	†	0.0113	0.0066	-0.02	0.0005
14	†	†	†	-0.0186	0.0015
15	†	†	0.01	†	0.0178
16	0.0216	†	0.004	0.0131	-0.0018
17	†	0.0145	-0.0036	†	0.002
18	†	†	†	-0.0099	0.0054
19	†	0.0194	0.0032	-0.0158	0.0122
20	0.0069	†	-0.0061	0.013	0.0038
21	†	†	0.0073	0.0115	0.0111
22	-0.0075	†	†	0.0002	-0.003
23	†	0.0119	-0.0048	0.0237	-0.0353
24	0.0318	†	0.0202	-0.0105	0.0199
25	0.0002	†	†	-0.0033	0.0127
26	-0.0024	†	-0.0085	†	0.0083
27	0.0164	†	†	0.001	0.0051
28	†	0.0118	0.000	†	0.0001
29	0.0005	†	-0.0003	†	-0.0075
30	-0.0068	†	0.0067	-0.028	-0.005
31	†	†	-0.0021	†	0.0197
32	†	0.0137	-0.0046	†	-0.0003
33	0.0001	-0.0177	†	†	-0.0021
34	†	†	-0.0105	†	-0.0158
35	†	-0.0108	0.001	0.0012	-0.007
36	-0.001	0.003	0.0189	†	-0.005

Table 3.
RMSE for 10% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.0693	†	0.0689	0.0519
2	†	0.0716	0.0637	†	0.0436
3	†	0.0657	0.0642	†	0.0480
4	0.0624	†	0	0.0554	0.0466
5	†	0.0659	0.0621	†	0.0702
6	0.0623	†	0	0.0654	0.0609
7	†	0	†	0	0.0376
8	0.0798	†	0	0.0811	0.0858
9	0.0826	0.0965	0.0655	†	0.0644
10	†	0.0865	0.0904	0.0798	0.0780
11	†	0	†	0	0.0469
12	†	0.0783	0.0639	†	0.0561
13	†	0.0884	0.0839	0.0874	0.0842
14	†	0	†	0.0470	0.0481
15	†	0	0.0654	†	0.0677
16	0.0776	†	0.0806	0.0908	0.0552
17	†	0.0714	0.0535	†	0.0522
18	†	0	†	0.0499	0.0470
19	†	0.0900	0.0661	0.0814	0.0672
20	0.0534	†	0.0671	0.0776	0.0722
21	†	0	0.0669	0.0628	0.0619
22	0.07299	†	0	0.0610	0.0612
23	†	0.0898	0.0833	0.0841	0.0927
24	0.0862	†	0.1047	0.0875	0.1034
25	0.0628	†	0	0.0710	0.0614
26	0.0640	†	0.0941	†	0.0760
27	0.0744	†	0	0.0769	0.0422
28	†	0.0649	0.0687	†	0.0509
29	0.0665	†	0.0596	†	0.0468
30	0.0532	†	0.0680	0.0720	0.0615
31	†	0	0.0541	†	0.0460
32	†	0.0808	0.0615	†	0.0697
33	0.0633	0.0670	†	0	0.0562
34	†	0	0.1611	†	0.1690
35	†	0.0911	0.0861	0.0858	0.0684
36	0.0670	0.0766	0.0780	†	0.0671

Table 4.
Bias of Slopes and Difficulties for 25% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.0115	†	-0.0219	0.0007
2	†	-0.019	0.0054	†	-0.0102
3	†	-0.0166	0.011	†	-0.0066
4	-0.0136	†	†	0.0062	0.005
5	†	0.0073	-0.0121	†	-0.0111
6	0.0087	†	†	0.0108	-0.0135
7	†	†	†	†	-0.0012
8	-0.0076	†	†	0.0121	-0.0144
9	0.0089	-0.0019	-0.0096	†	-0.0061
10	†	0.0042	0.0127	0.0066	-0.0011
11	†	†	†	†	-0.0112
12	†	-0.0052	-0.0027	†	0.0045
13	†	0.0002	0.005	-0.0143	-0.0137
14	†	†	†	-0.023	0.0058
15	†	†	0.018	†	0.0246
16	0.0201	†	-0.013	0.0335	-0.0186
17	†	0.0174	-0.0045	†	-0.0025
18	†	†	†	-0.0065	0.011
19	†	0.0072	0.0191	-0.0308	0.0026
20	0.0297	†	0.0006	0.0001	0.0165
21	†	†	0.004	0.0208	0.014
22	-0.015	†	†	-0.0117	0.0157
23	†	0.0301	-0.0111	0.0164	-0.0266
24	0.0465	†	0.0177	0.0075	0.0242
25	0.0048	†	†	-0.0047	0.0172
26	-0.0005	†	-0.003	†	0.0064
27	0.0047	†	†	0.0182	-0.0061
28	†	0.0333	-0.0028	†	0.0069
29	-0.0104	†	-0.0002	†	-0.001
30	0.0073	†	-0.0071	-0.0176	-0.0124
31	†	†	0.0032	†	0.0173
32	†	0.0366	-0.015	†	-0.0038
33	0.0007	-0.0134	†	†	0.0011
34	†	†	-0.0186	†	-0.0183
35	†	-0.0186	0.0006	0.0159	-0.001
36	-0.0045	0.0127	0.01	†	0.0058

Table 5.
RMSE for 25% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.0784	†	0.0936	0.0474
2	†	0.0915	0.0814	†	0.0444
3	†	0.0693	0.0592	†	0.0522
4	0.079	†	†	0.0631	0.0460
5	†	0.0712	0.0784	†	0.0688
6	0.073	†	†	0.0789	0.0652
7	†	†	†	†	0.0384
8	0.0876	†	†	0.1058	0.0917
9	0.0843	0.1057	0.0885	†	0.0708
10	†	0.1159	0.0932	0.1051	0.0879
11	†	†	†	†	0.0468
12	†	0.1032	0.091	†	0.0632
13	†	0.1017	0.0836	0.0982	0.0907
14	†	†	†	0.0638	0.0473
15	†	†	0.0796	†	0.0838
16	0.0807	†	0.088	0.1131	0.0737
17	†	0.0848	0.0723	†	0.0645
18	†	†	†	0.0546	0.0488
19	†	0.105	0.0717	0.1061	0.0718
20	0.0858	†	0.0874	0.0854	0.0832
21	†	†	0.0762	0.059	0.0674
22	0.0767	†	†	0.086	0.0713
23	†	0.1106	0.0807	0.0916	0.0974
24	0.1184	†	0.1071	0.1042	0.1372
25	0.0787	†	†	0.0755	0.0736
26	0.0778	†	0.1056	†	0.0880
27	0.085	†	†	0.0783	0.0504
28	†	0.0879	0.0623	†	0.0574
29	0.0708	†	0.0696	†	0.0580
30	0.0603	†	0.086	0.0769	0.0777
31	†	†	0.0683	†	0.0481
32	†	0.0977	0.0822	†	0.0632
33	0.0696	0.0821	†	†	0.0720
34	†	†	0.2192	†	0.2148
35	†	0.1173	0.0957	0.094	0.0946
36	0.0875	0.0872	0.0844	†	0.0675

Table 6.
Bias of Slopes and Difficulties for 50% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.015	†	-0.0066	0.0047
2	†	-0.0435	0.0047	†	0.0098
3	†	0.0132	-0.0046	†	-0.0145
4	0.0046	†	†	0.012	0.0078
5	†	-0.0208	0.0065	†	0.0047
6	0.0077	†	†	-0.0136	-0.0264
7	†	†	†	†	0.006
8	0.0176	†	†	0.0097	0.0035
9	0.0138	-0.0196	0.0149	†	-0.004
10	†	0.0479	0.0014	0.0133	-0.0109
11	†	†	†	†	-0.0174
12	†	-0.0073	-0.0054	†	-0.0003
13	†	0.0205	0.0054	-0.0305	-0.0011
14	†	†	†	-0.0111	0.002
15	†	†	0.0133	†	0.0308
16	0.0364	†	0.0071	0.0527	-0.03
17	†	0.0002	0.0016	†	0.016
18	†	†	†	-0.0027	0.0001
19	†	0.0294	-0.0098	-0.0137	0.0202
20	0.0443	†	0.0122	0.0298	0.0239
21	†	†	0.0001	0.0316	-0.0011
22	-0.0006	†	†	0.0259	-0.0166
23	†	0.044	-0.0285	0.0542	-0.0348
24	0.0841	†	0.0575	-0.008	0.067
25	0.0166	†	†	-0.0036	0.007
26	0.0126	†	-0.0023	†	0.0018
27	0.0142	†	†	0.0327	-0.0102
28	†	0.0044	0.0027	†	0.009
29	0.0133	†	-0.0044	†	0.0057
30	0.0247	†	-0.014	-0.0202	-0.0174
31	†	†	-0.0057	†	0.0279
32	†	0.036	-0.0116	†	-0.0014
33	0.0224	-0.021	†	†	-0.0037
34	†	†	-0.0131	†	-0.0197
35	†	-0.0116	0.0088	-0.0007	-0.0334
36	-0.0086	-0.0146	0.0286	†	-0.005

Table 7.
RMSE for 50% Missing Data

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	β_i
1	†	0.1056	†	0.0953	0.0824
2	†	0.1272	0.1044	†	0.07
3	†	0.1251	0.1041	†	0.0805
4	0.101	†	†	0.0995	0.0661
5	†	0.1387	0.1376	†	0.0958
6	0.1129	†	†	0.1259	0.0751
7	†	†	†	†	0.0626
8	0.1517	†	†	0.1455	0.1305
9	0.1158	0.1347	0.1217	†	0.0918
10	†	0.1658	0.1417	0.1329	0.1157
11	†	†	†	†	0.0643
12	†	0.1354	0.1076	†	0.0799
13	†	0.1811	0.1612	0.1326	0.1092
14	†	†	†	0.0634	0.0676
15	†	†	0.1114	†	0.096
16	0.148	†	0.1301	0.1688	0.0992
17	†	0.0998	0.1152	†	0.0676
18	†	†	†	0.0774	0.0528
19	†	0.1874	0.1333	0.1347	0.1037
20	0.1228	†	0.1173	0.1489	0.1135
21	†	†	0.1062	0.1005	0.0705
22	0.1002	†	†	0.1144	0.0928
23	†	0.1718	0.1874	0.1692	0.1427
24	0.2093	†	0.1616	0.1759	0.1807
25	0.1045	†	†	0.1111	0.0772
26	0.1224	†	0.1328	†	0.1039
27	0.1055	†	†	0.1154	0.0552
28	†	0.1346	0.1043	†	0.0847
29	0.0947	†	0.0817	†	0.0687
30	0.1272	†	0.1375	0.1244	0.0925
31	†	†	0.0802	†	0.0669
32	†	0.1634	0.1357	†	0.0835
33	0.1071	0.1174	†	†	0.1035
34	†	†	0.2305	†	0.2308
35	†	0.19	0.156	0.1725	0.1085
36	0.1017	0.1556	0.1105	†	0.0885

Table 8.
Population Parameter Recovery for 10% Missing Data

Patt.	Means	Truth	Diff.	Diff./S.E.	RMSE
0	0.2605	0.2604	0.0001	0.1140	0.0049
1	0.0514	0.0521	-0.0007	-0.7156	0.0043
2	0.0512	0.0521	-0.0009	-0.8105	0.0048
3	0.0102	0.0104	-0.0002	-0.4335	0.0026
4	0.0522	0.0521	0.0001	0.1124	0.0034
5	0.0103	0.0104	-0.0001	-0.1894	0.0026
6	0.0103	0.0104	-0.0001	-0.2652	0.0025
7	0.0520	0.0521	-0.0001	-0.1538	0.0038
8	0.0527	0.0521	0.0006	0.6472	0.0040
9	0.0107	0.0104	0.0003	0.5079	0.0023
10	0.0109	0.0104	0.0005	0.9385	0.0023
11	0.0528	0.0521	0.0008	0.9267	0.0038
12	0.0100	0.0104	-0.0004	-0.6184	0.0029
13	0.0521	0.0521	0.0000	0.0434	0.0047
14	0.0514	0.0521	-0.0007	-0.7187	0.0042
15	0.2613	0.2604	0.0009	0.9530	0.0043

Table 9.
Population Parameter Recovery for 25% Missing Data

Patt.	Means	Truth	Diff.	Diff./S.E.	RMSE
0	0.2598	0.2604	-0.0006	-0.4967	0.0054
1	0.0514	0.0521	-0.0006	-0.5357	0.0054
2	0.0518	0.0521	-0.0003	-0.1763	0.0066
3	0.0100	0.0104	-0.0004	-0.6526	0.0029
4	0.0518	0.0521	-0.0003	-0.3205	0.0038
5	0.0113	0.0104	0.0009	1.2607	0.0033
6	0.0101	0.0104	-0.0003	-0.5259	0.0023
7	0.0521	0.0521	0.0000	-0.0198	0.0056
8	0.0527	0.0521	0.0006	0.5695	0.0047
9	0.0104	0.0104	0.0000	0.0515	0.0025
10	0.0115	0.0104	0.0011	1.4175	0.0036
11	0.0532	0.0521	0.0011	1.0532	0.0047
12	0.0095	0.0104	-0.0010	-1.2084	0.0036
13	0.0517	0.0521	-0.0004	-0.3029	0.0054
14	0.0512	0.0521	-0.0009	-0.7858	0.0052
15	0.2614	0.2604	0.0010	0.8891	0.0052

Table 10.
Population Parameter Recovery for 50% Missing Data

Patt.	Means	Truth	Diff.	Diff./S.E.	RMSE
0	0.2584	0.2604	-0.0021	-1.3744	0.0070
1	0.0527	0.0521	0.0006	0.4371	0.0064
2	0.0516	0.0521	-0.0005	-0.2983	0.0077
3	0.0100	0.0104	-0.0004	-0.3261	0.0061
4	0.0521	0.0521	0.0000	0.0115	0.0063
5	0.0096	0.0104	-0.0008	-0.6357	0.0058
6	0.0103	0.0104	-0.0002	-0.1627	0.0044
7	0.0514	0.0521	-0.0007	-0.3862	0.0078
8	0.0540	0.0521	0.0019	1.2356	0.0072
9	0.0100	0.0104	-0.0004	-0.4797	0.0041
10	0.0100	0.0104	-0.0005	-0.4001	0.0051
11	0.0547	0.0521	0.0026	1.4168	0.0086
12	0.0111	0.0104	0.0007	0.5641	0.0056
13	0.0524	0.0521	0.0003	0.1326	0.0101
14	0.0514	0.0521	-0.0007	-0.3504	0.0093
15	0.2606	0.2604	0.0001	0.0778	0.0077

4. Cognitive Diagnosis of NAEP Assessment Data

NAEP data is unique in many ways. Two features of concern in the study are the high degree of sparseness in the data and the group level characteristics of all estimates. To avoid fatigue effects and to maintain a certain level of motivation, each NAEP student is administered only 20-25 items on average. Such a short test obviously will not lead to an accurate estimate of the latent ability for each individual. Fortunately, the group level characteristics, instead of the individual ability estimates, are the targets of inference in NAEP. By aggregating the estimates of the latent ability across individuals when working with the posterior distribution of the latent ability, it is possible to obtain a consistent estimate of the latent ability distribution for subgroups of interest (Mislevy, Beaton, Kaplan, & Sheehan, 1992).

An immediate question for cognitive diagnosis of NAEP data is how to estimate the distribution of latent attribute patterns for subgroups. There are three solutions available to this question. Let $P(\mathbf{Y}|\alpha)$ be the likelihood function under a specified model and $\alpha \sim G(\cdot)$ be the prior distribution of this latent space defined by the latent attribute patterns. The simplest solution is to specify a single distribution $G(\cdot)$ for the latent attribute patterns and then obtain the estimates

for subgroups by

$$P(\alpha) = \sum_{i \in \ subgroup \quad k} \omega_i P(\alpha_i | \mathbf{Y}_i),$$

where ω_i is the weight for each examinee in this subgroup. Though this solution allows one to borrow information from other observations because it uses all the data to estimate the latent distribution, this approach will give biased estimates for subgroups when the skill distribution of the subgroups of interest differ to a large extent. The second approach is to specify prior distributions for each subgroup separately and then obtain the subgroup estimates by summing over the posteriors of each examinee in that subgroup. This approach is equivalent to multiple group IRT (Bock & Zimowski, 1997), but in this case it was applied to the diagnosis model. This approach provided a consistent estimate of the latent distribution for each subgroup. However, if there are a huge number of subgroups to estimate, this approach will lead to "the curse of dimensionality." That is, the number of parameters will increase as the number of subgroups increases so that there is not enough data to estimate the parameters. The third and a more complex approach is to specify the prior distribution dependent on the covariates of interest. That is, $P(\alpha|\mathbf{z}) \sim f(\mathbf{z}'\beta, \alpha)$, where $\mathbf{z}$ is the covariate vector. Then the estimates for each subgroup are calculated by aggregating the posterior $P(\alpha_i|\mathbf{Y}_i, \mathbf{z}_i)$ over the students in that subgroup of interest. This approach allows one to utilize the information from all of the data $\mathbf{Y}$ and $\mathbf{z}$ as well as to consider the differences among subgroups. The current study focused on the first two solutions because they are simple to implement. The third approach emulates what is operationally done in NAEP estimation of a latent regression model, commonly referred to as conditioning modeling, in the context of large scale educational assessment.

The GDM with a single group assumption (first approach) and a multiple group assumption (second approach) were used to analyze a Reading assessment data set, and only a single group assumption was used to analyze a Math assessment data set. A Q matrix does not exist for NAEP current assessment data. This study used the scales defined in the *Reading Framework of the National Assessment of Educational Progress, 1992-2000* (National Assessment Governing Board [NAGB], n.d.) and the *Mathematics Framework for the 2005 National Assessment of Educational Progress* (NAGB, 2004) as cognitive attributes, and the correspondence between items and scales defined the Q matrices used here. The data and results are described in the following subsections.

4.1 Reading Assessment: Data Description

Grade 12 data from the 2002 NAEP Reading assessment were analyzed using the GDM. This assessment included three subscales: reading for literary experience, reading for information, and reading to perform a task (National Center for Education Statistics, 2003). In this study, a simple Q matrix was chosen in the way that these three subscales were taken as three attributes and the items had entries only on the scale they belonged to. For example, if an item was measured in the reading for literary experience scale, then the item had 1 on this scale and 0 on the other two attribute scales (i.e., reading for information and reading to perform a task). The data contained responses to 112 items in total by 14,724 students. The students were selected by a stratified sampling plan, and the allocation of items to students followed the balanced incomplete block design (NAEP Glossary of Terms, n.d.). One of the salient features of NAEP is that each student takes only part of the complete set of items. In particular for this assessment, each student took only 20 items on average. Thus, the structural missing values dominate the huge data matrix, since each examinee takes on average only about one sixth of the items from the total of 112 items. In the NAEP Reading design, students commonly receive test booklets with items from only one scale. Fortunately, some booklets contain items from at least two scales, allowing the model to tap into information on two attributes from examinees using those booklets.

4.2 Reading Assessment: Analysis and Results

The data were analyzed under both a single group assumption and a multiple group assumption. In the single group assumption, there was only one prior distribution for the latent attribute patterns, and this prior distribution was updated through the iterations in the EM algorithm. Three models were used to first fit the data under the single group assumption: the 2PL IRT model (Birnbaum, 1968), the three-skill-three-level GDM, and the three-skill-four-level GDM. The number of levels for each attribute was chosen to take both model complexity and model inference into account. In particular, a model with five or more levels for each attribute will result in too large a number of latent attribute patterns to estimate. A model with two levels for each attribute in this case seemed to make strong assumptions on mastery and nonmastery.

As mentioned earlier, the three cognitive skills are reading for literary experience, reading for information, and reading to perform a task. The three or four levels for each attribute were prespecified. For example, in this analysis, the three levels were represented by -1.41421, 0.00000,

and 1.41421, while the four levels were represented by -1.73205, -0.57735, 0.57735, and 1.73205. These levels are the profile scores for each attribute, and they are specified by the users. In this analysis, we selected these levels around 0 symmetrically. These three models were compared in terms of the log-likelihood and Bayesian information criterion (BIC; Schwartz, 1978) index. Results indicated that the three-skill-four-level model fitted the data better than the 2PL model and the three-skill-three-level model in terms of both log-likelihood and BIC (see the first three rows in Table 11). Therefore, the analysis focused on the three-skill-four-level model. This model was then applied in the multiple group analysis. With the multiple group assumption, separate prior distributions of latent attribute patterns were specified for different subgroups. For instance, with a gender subgroup analysis, two separate prior distributions were used, one for the male group and one for the female group. Thus, these two separate marginal distributions for the latent attribute patterns were updated along with the EM algorithm. The likelihood and BIC index for male and female groups are listed in the gender group row in Table 11. The results for separate racial group analysis are also listed in that table. The four racial groups analyzed were White, Black, Hispanic, and Asian. The results showed that multiple group analysis for gender and race gave larger log-likelihood than single group analysis. However, the racial group analysis resulted in larger BIC than single group analysis, while gender group analysis resulted in smaller BIC than single group analysis. We have to investigate further to understand the results.

Table 11.
Model Comparison: Reading Assessment

Models	# of parameters	Log-likelihood	BIC
Under single group assumption			
2PL IRT	328	-150234.69	303617.3
3-skill-3-level GDM	290	-150355.23	303493.7
3-skill-4-level GDM	327	-149987.27	303112.8
Under multiple group assumption			
Racial group	516	-149576.38	303513.0
Gender group	390	-149701.72	302699.2

The latent attribute distributions for each of the racial groups resulting from the single group assumption and the multiple group assumption were plotted and compared in Figures 1 through 3. Figure 1 presents the results for Skill 1. Figure 2 corresponds to Skill 2 and Figure 3 to Skill 3.

Within each figure, four graphs represent the four different racial groups. Each graph

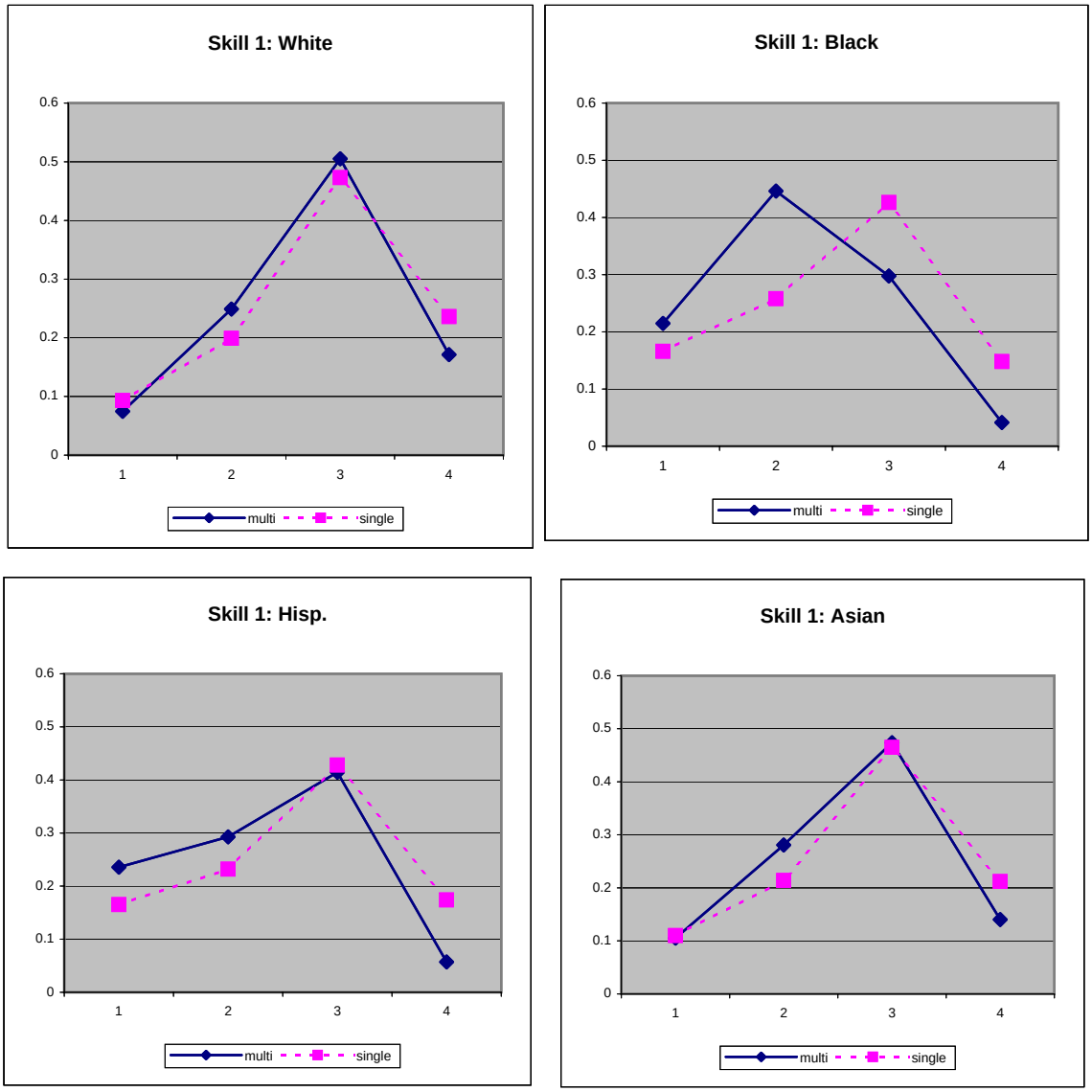

***Figure* 1. Empirical score distributions for racial groups for Skill 1.**

compares the skill distribution obtained under a single group assumption (dotted line) and a multiple group assumption (solid line) and shows that these two assumptions have different outcomes with respect to the estimates of skill distributions for different racial groups. In particular, the two group assumptions do not exhibit much difference in the estimates for the White and Asian students, but they demonstrate dramatic differences for the Black and Hispanic students. For example, the Black students have dramatically different skill distributions under these two assumptions for all three skills, while the Hispanic students have substantially different

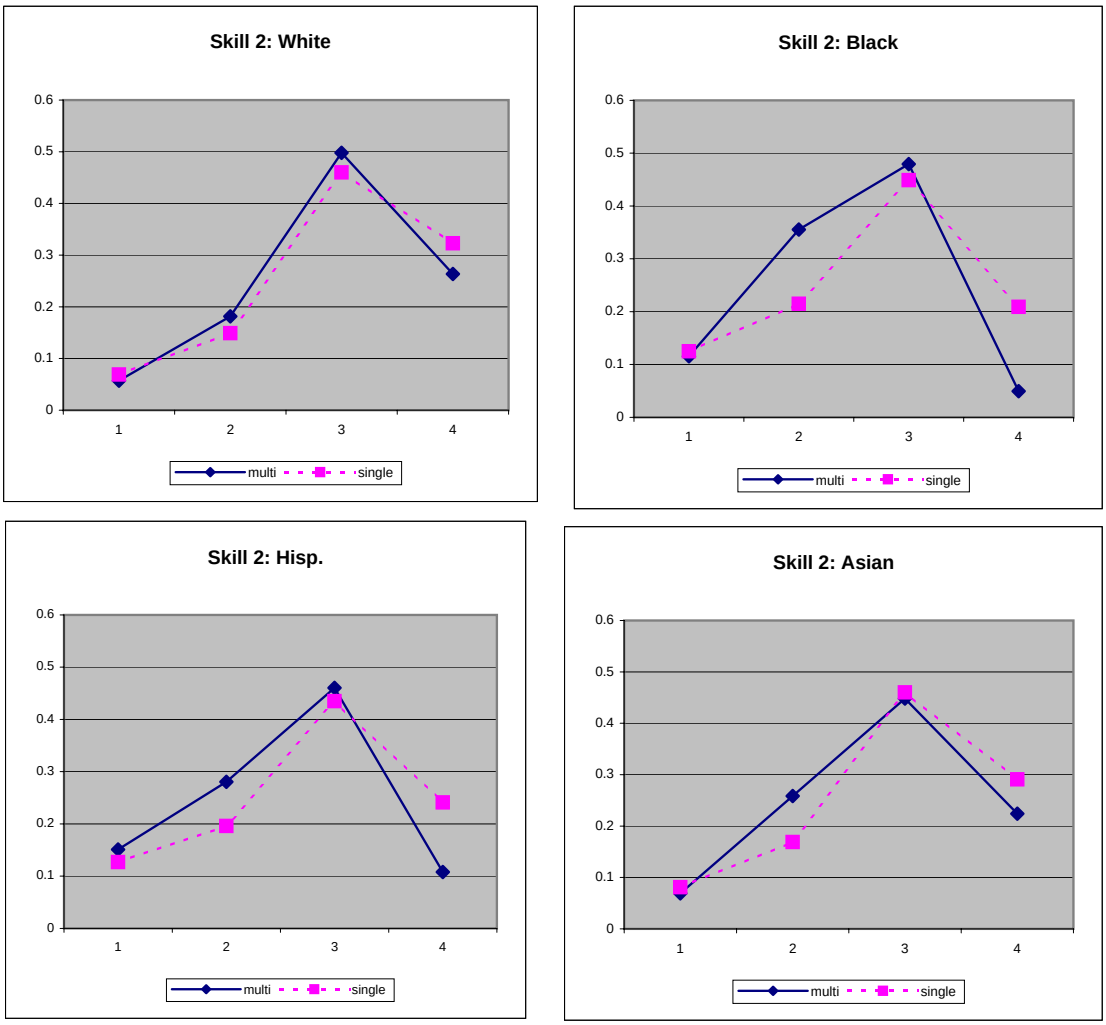

***Figure* 2. Empirical score distributions for racial groups for Skill 2.**

skill distributions for Skill 3 under the two assumptions. It seems that the single group model leverages the differences among groups at the extreme points. Take Skill 3 as an example (see Figure 3). Under the single group assumption, the high percentages of the White and Asian groups in the highest levels of this skill compensate for the lower percentages of the Black and Hispanic groups. The multiple group model breaks apart this compensating effect and results in a decrease of the estimates for this level. In short, this finding shows that the single group assumption might generate biases in estimating the skill distributions for those subgroups that are not considered in building the population model. The multiple group model will provide more accurate estimates for the groups that are represented explicitly in the population model. In this

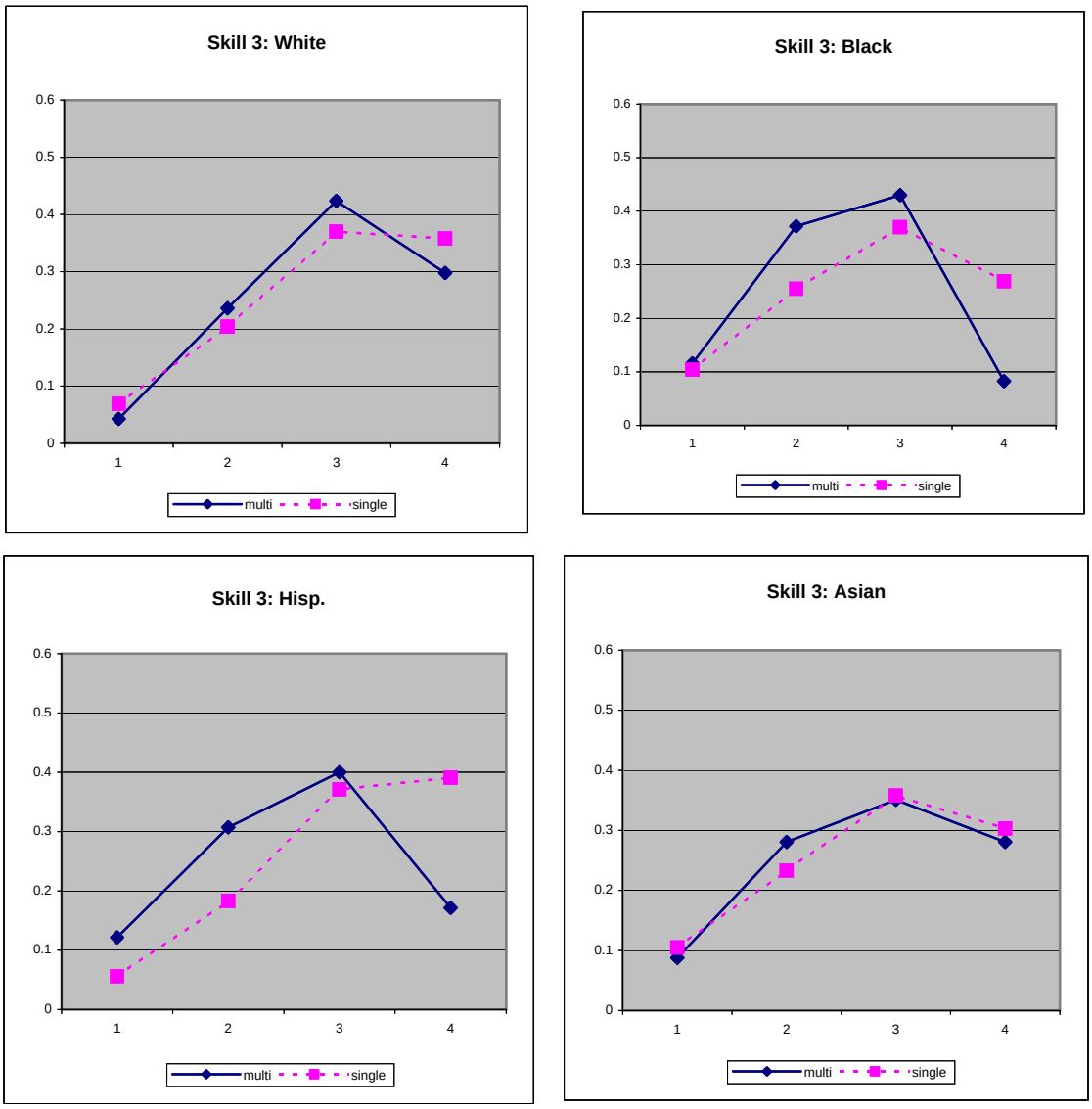

***Figure* 3. Empirical score distributions for racial groups for Skill 3.**

study, the background variables, such as gender or racial groups, were considered one at a time. No interactions or more than one background variable were considered simultaneously. A direct deduction from our finding is that the inference concerning the interaction between background variables might not be as accurate as one might expect. Thus, more complicated models need to be investigated, and that is one focus of our future research.

In addition, the group characteristics (such as mean and standard deviation) for racial groups derived from the GDM analysis under the multiple group assumption and the operational NAEP

analysis are compared in Table 12. Given that the GDM uses a slightly different base model (2PL instead of 3PL) for most dichotomous items and given the limited number of located classes in the GDM used here, the derived group characteristics based on the GDM analysis are quite close to the operational NAEP analysis.

Table 12.
Comparison Between GDM and NAEP Operational Analysis

	GDM Mean	NAEP Mean	Difference	GDM Std	NAEP Std
Skill 1					
White	0.1714	0.1764	0.0050	0.9436	0.9191
Black	-0.5274	-0.5240	0.0034	0.9420	0.9161
Hisp.	-0.3936	-0.4041	-0.0105	1.0250	0.9913
Asian	0.0333	0.0209	-0.0124	0.9806	0.9246
Skill 2					
White	0.1494	0.1703	0.0209	0.9476	0.9198
Black	-0.4253	-0.4678	-0.0425	0.8802	0.9058
Hisp.	-0.3566	-0.3965	-0.0399	1.0124	0.9989
Asian	-0.0107	-0.0453	-0.0345	0.9851	0.9097
Skill 3					
White	0.1384	0.1536	0.0152	0.9674	0.9878
Black	-0.4490	-0.4866	-0.0376	0.9267	0.9738
Hisp.	-0.2778	-0.2893	-0.0115	1.0455	1.0900
Asian	-0.0219	-0.0715	-0.0496	1.0694	1.0468

4.3 Math Assessment: Data Description

Four subscales were measured in the 2005 NAEP grade 12 math assessment. These four subscales were number properties and operations, measurement and geometry, data analysis and probability, and algebra. In addition, each item was classified as being in one of three categories of complexity: low, medium, and high (NAGB, 2004). The items of low complexity relied heavily on recognition of learned concepts and principles. The items in the moderate complexity category involved more flexible thinking than those in the low complexity category. The items in the high complexity category made heavy demands on abstract reasoning, planning, analysis, judgment, and creative thought. Thus, an item belonged to one of the four subscales and one of the three categories. Though there were 180 items in this assessment, each student took only 30-35 items, which was only about one sixth of the whole set. Unlike the Reading assessment, each student took items from all subscales. This design will likely improve the estimates of the skill mastery

for individuals compared to those from the Reading assessment. However, this small number of items is still unreliable in providing consistent individual estimates on the student level. Instead, the group characteristics, such as the skill distribution functions, should be estimated.

4.4 Math Assessment: Analysis and Results

As in the case of Reading assessment, the current Math assessment was not developed for cognitive diagnosis purposes. In the current study, the four scales and the three categories were considered as cognitive attributes. Three Q matrices were formatted from the item specifications defined in the framework. The first Q matrix contained the four subscales only, on which the items had Entries 1 or 0, depending on the specification in the framework. The second Q matrix included both subscales and complexity categories, in which the items had Entries 1 or 0, according to the item specifications. The third Q matrix was a simpler version of the second one, developed by deleting the low complexity category. Two, three, or four levels were assumed for each attribute in these Q matrices with consideration of identifiability. For instance, for the second and third Q matrices, the attributes with more than two levels could lead to a huge number of latent classes to estimate. The maximum number of levels for attributes in the first Q matrix was four. So in this study, three models were examined and compared: the four-attribute-four-level GDM, the six-attribute-two-level GDM, and the seven-attribute-two-level GDM.

For this analysis, the whole data set of 9,347 examinees was randomly divided into two smaller data sets for the purpose of cross-validation. Each of them contained about 4,600 examinees. One of the smaller data sets (denoted as data set A) was used to check the model fits of these three models, and from this a model was chosen to analyze the data. The other data set (denoted as data set B) was used to examine whether the item parameter estimates were close to those in the first data set under the model chosen. The goodness-of-fit results for data set A can be seen in Table 13.

All these models were used under the single group assumption because there are too many population parameters to estimate otherwise. It shows that the seven-skill-two-level model is better than the other two models in terms of log-likelihood, and the four-skill-four-level model is better than the other two according to BIC index. Since this study was only an exploratory study of the possibility of using cognitive diagnosis analysis for NAEP assessment data, and we wanted both scales and complexity categories in this model, we chose the seven-skill-two-level model. We

Table 13.

Model Comparison: Math Assessment

Models	# of parameters	Log-likelihood	BIC
4-skill-4-level GDM	678	-92961.9838	191652.0
7-skill-2-level GDM	730	-92783.8280	191735.1
6-skill-2-level GDM	632	-93531.1921	192401.8

used this model to analyze data set B to check the stability of item parameter estimates. The absolute differences between the item parameter estimates in the two data sets are summarized in Table 14.

Table 14.
Frequency of Differences Between the Estimates Obtained From Two Data Sets

Bin	Slope I	Slope II	Intercept	Total number
0	80	80	126	286
0.05	35	26	28	89
0.1	30	25	22	77
0.3	33	42	50	125
0.5	0	5	7	12
More	2	2	10	14

The values of these differences are divided into six bins: 0, 0-0.05, 0.05-0.1, 0.1-0.3, 0.3-0.5, and more. Three types of estimates are listed in this table. The Slope I column represents the slope parameter estimates for the subscales, the Slope II column stands for the slope parameter estimates for the complexity categories, and the Intercept column means the intercept parameter estimates. The entries under Slope I, Slope II, and Intercept are the frequencies of differences falling into different categories. For instance, 80 in the first row in the Slope I column is the frequency of those differences having a value of zero. According to this table, 452 item parameters out of a total of 603 have differences of less than 0.1 across the two data sets. Considering the small relative ratio of the number of examinees per parameter for these two data sets, 4668/603 and 4679/603, the parameter estimates are quite stable across data sets. This result was consistent with what we found in the simulation study, where the item parameters can be

recovered satisfactorily.

Table 15 lists the marginal probabilities of attribute mastery (including subscales and complexity categories) for subgroups of interest, such as gender, racial groups, limited English proficiency (LEP), and individual education plans (IEPs) for disabled students.

Table 15.
Marginal Skill Mastery Given Background Characteristics: Using Anchor Points -1.73205 and 1.73205

Subgroups	Scale I	Scale II	Scale III	Scale IV	Low	Medium	High
Male	0.471	0.485	0.495	0.493	0.394	0.392	0.389
Female	0.458	0.466	0.478	0.476	0.354	0.352	0.350
White	0.495	0.512	0.523	0.519	0.432	0.430	0.427
Black	0.363	0.348	0.369	0.364	0.173	0.170	0.169
Hispanic	0.389	0.382	0.396	0.390	0.218	0.216	0.215
Asian	0.507	0.544	0.544	0.559	0.514	0.512	0.509
IEP (Y)	0.279	0.253	0.278	0.265	0.106	0.104	0.105
IEP (N)	0.474	0.502	0.509	0.519	0.523	0.522	0.519
IEP (Missing)	0.479	0.491	0.502	0.499	0.392	0.390	0.387
LEP (Y)	0.309	0.296	0.309	0.313	0.134	0.132	0.131
LEP (N)	0.472	0.482	0.494	0.491	0.383	0.381	0.378
LEP (Fomerly)	0.480	0.498	0.502	0.515	0.443	0.442	0.438

Table 15 shows that the female group had smaller probabilities of mastery in all attributes than the male group, which is consistent with the results from operational NAEP analysis. The White and Asian students tended to have larger probabilities of mastery than the Hispanic and Black students. Students who did not have an IEP or who were not identified as having LEP tended to have greater probability to master these attributes than those who had an IEP or who were identified as having LEP. These conclusions are very similar to those found in operational NAEP analysis.

5. Discussion

As observed by von Davier and Yamamoto (2004), the GDM is an extension of the partial credit model (PCM; Masters, 1982) derived by decomposing the unidimensional latent trait into item-dependent linear combinations of K underlying discrete traits. This makes the model similar to that of confirmatory factor analysis, while operating with categorical item response variables, rather than continuous variables. These underlying latent traits are what we were interested in

knowing for each examinee and for each group. Though the goodness of fit and model selection criteria (such as BIC and Akaike information criterion [AIC]) are still important for model selection in fitting latent trait models, they may not be the only criteria used for model selection when the interest lies in identifying the underlying latent traits of students and subpopulations. Cognitive diagnosis modeling provides a tool to extract the needed information.

Through simulation and real data implementations, this paper demonstrated satisfactory recovery of item parameters when using the GDM to fit a spare data matrix, and it demonstrated possible applications with NAEP assessment data. At the present, only simple Q matrices are being used in the analysis of NAEP data because the NAEP assessments were not designed for cognitive diagnosis purposes. A single group assumption and a multiple group assumption for the population distribution are used to fit the Reading assessment data. It is shown that these two assumptions lead to different conclusions for certain subgroups, such as Hispanic and Black students, whose latent attribute distributions are more likely leveraged by the performances of White and Asian students under the single group assumption than the multiple group assumption. These observations are compatible with those from operational NAEP analysis.

Group characteristics for policy relevant reporting variables are the focus of NAEP. The effectiveness and accuracy of the estimates of subgroup characteristics lie in carefully choosing predictors and deliberately confining the interpretations to a certain extent. The inclusion of a latent regression model to predict skill distributions conditioned on group membership increases model complexity proportional to the total number of subgroups. Previous criticism of operational NAEP procedures addressed a similar issue. One major point of this criticism was whether the same or almost the same accuracy of subgroup distribution estimates can be achieved with latent regression models of much lower levels of complexity for the operational NAEP analysis. This question urgently needs to be addressed if models are used to predict subgroup distributions on a multitude of discrete skills rather than a few continuous proficiency variables. Our future research will focus on the best trade-off between model complexity and accuracy of inference.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–449.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hamilton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer-Verlag.

von Davier M.(2005). *A general diagnostic model applied to language testing data* (ETS RR-05-16). Princeton, NJ: ETS.

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis.* Paper presented at the 4th Spearman Invitational Conference, Philadelphia, PA.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Junker, B., & Sitjsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparameteric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mislevy, R., Beaton, A. E., Kaplan, B. & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–162.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

NAEP glossary of terms. (n.d.). Retrieved April 25, 2006, from http://nces.ed.gov/nationsreportcard/glossary.asp

National Assessment Governing Board. (n.d.). *Reading framework for the National Assessment of Educational Progress: 1992-2000.* Retrieved April 26, 2006, from http://www.nagb.org/pubs/92-2000read/toc.html

National Assessment Governing Board. (2004). *Mathematics framework for the 2005 National Assessment of Educational Progress.* Retrieved April 26, 2006, from

http://www.nagb.org/pubs/m_framework_05/toc.html

National Center for Education Statistics. (2003). *The nation's report card: Reading 2002* (NCES 2003-521). Retrieved April 28, 2006, from http://nces.ed.gov/nationsreportcard/pdf/main2002/2003521.pdf

Roussos, L. (1994). *Summary and review of cognitive diagnosis models.* Unpublished manuscript.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.