

**STRATEGIES FOR ADDRESSING  
HIGH-DIMENSIONAL COGNITIVELY DIAGNOSTIC  
ASSESSMENT PROBLEMS**

**BY YAN SUN**

**A dissertation submitted to the  
School of Graduate Studies  
Rutgers, The State University of New Jersey  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
Graduate Program in Education**

**Written under the direction of**

**Jimmy de la Torre**

**and approved by**

---

---

---

---

**New Brunswick, New Jersey**

**October, 2019**

**© 2019**

**Yan Sun**

**ALL RIGHTS RESERVED**

## **ABSTRACT OF THE DISSERTATION**

# **Strategies for Addressing High-Dimensional Cognitively Diagnostic Assessment Problems**

**by Yan Sun**

**Dissertation Director: Jimmy de la Torre**

In recent years, cognitive diagnosis models (CDMs) have sparked the interest of educational measurement researchers and practitioners because of its capability to provide formative information on student mastery or nonmastery of a set of fine-grained skills. One of the advantages of CDMs is that, by treating latent variables as discrete, usually binary, CDMs can accommodate higher dimensional latent space than multidimensional latent trait (e.g., multidimensional item response theory) models. Theoretically, the number of attributes that can be estimated by a CDM is unlimited; however, in practice, this number may not exceed 20 due to a number of computational issues. This constraint limits the use of CDMs in scenarios where a comprehensive diagnosis of a complete knowledge space, such as large-scale diagnostic assessments or retrofitting summative assessments

using CDMs, is of interest.

In this dissertation, a series of strategies are proposed to address issues in classifying examinees' proficiency profiles for high-dimensional testing data. In particular, these strategies can be used in situations where attributes can be partitioned into non-overlapping knowledge subsets. An approach, called the accordion procedure (AP), is proposed to address the high dimensionality estimation problem by focusing only on the attributes of one particular subset at a time, while the attributes of each of the remaining subsets are collapsed to create composite nuisance attributes. Simulation studies are conducted to examine the performance of AP compared to the complete profile estimation procedure in terms of classification accuracy and computation time. A real data illustration is also provided by retrofitting extant large-scale assessment data using AP.

To provide appropriate actionable feedback, one important prerequisite is ensuring the CDMs fitted to test data yield accurate classifications of examinees' proficiency profiles. However, due to various reasons (e.g., short test, poor item quality), tests sometimes do not provide sufficient information to classify examinees accurately.

When a test is not sufficiently informative, other sources of information might be needed to improve the classification accuracy. Thus, in the second study, covariates are incorporated in the context of AP using a four-step latent regression approach to supplement the information obtained from CDMs. The four-step approach is shown to be computationally more manageable when data are high-dimensional, as well as more flexible when specifications of each step need to be adjusted. Simulation and real-data studies are conducted to examine the performance of the proposed approach.

Cognitive diagnosis computerized adaptive testing (CD-CAT) has been proposed to

administer a test more efficiently by selecting the optimal set of items for each examinee. However, when the number of skills of interest is large, practical issues, such as calibration of item pools and item selection method, emerge.

The third study aims to propose a series of strategies to make high-dimensional CD-CAT feasible, namely, an item pool calibration method, item selection method, and examinees' prior distribution estimation method. Simulation studies are conducted to evaluate the performance of the proposed strategies.

In summary, the issues associated with using CDMs in high-dimensional situations are addressed in this dissertation. Several strategies are proposed primarily with the aim of obtaining accurate classification results to ensure that the feedback and remedial procedures are informative and effective.

## **Acknowledgements**

I would like to thank everyone who, in one way or another, advised, supported and criticized this work. First and foremost, I thank my advisor Dr. Jimmy de la Torre who continued advising me to conduct research after he had left Rutgers. Without Jimmy's solid academic training, financial support while I was visiting Hong Kong, constructive criticism and great new ideas in research, I couldn't imagine myself finishing this dissertation.

I appreciate the help I got from the great faculty members and staffs we have at GSE. Dr. Chia-Yi Chiu gave me the very first research project. Together with my colleague Yanhong, we were able to publish this work on one of the top journals in the field. I always thought it might be faster if she just publishes it on her own, but she was willing to spend time on enlightening the young scholars like me who knew barely nothing at the time. I also thank Dr. Drew Gitomer for his great insights in the connections between cognitive theory and modeling. Also, whenever you need help from Drew, you know that he's got your back. I thank Colleen McDermott for her always quick response and patience to every question I got.

I thank my mentor at ACT, Dr. Pravin Chopade, who joined my dissertation committee later. I had a great time in the summer of 2017 at Iowa City and eventually turned the research project into my dissertation.

I would love to thank my colleagues: Dr. Charles Iaconangelo, Dr. Mehmet Kaplan, Dr. Wenchao Ma and his wife Wenjing Guo, Dr. Nathan Minchen, Dr. Kevin Carl Santos, Dr. Miguel Sorrel Luján , Na'ama Av-Shalom, Yanhong Bian, and Yuan-Pei Chang for the help, advice, jokes, and company.

I thank my parents Yuping Sun and Xijun Zhao who never have a chance to go to college, but understand the power of knowledge. They probably don't understand what I'm doing in graduate schools (my mom would still ask me about how my homework was just a few years ago), but they supported it unconditionally. I thank my girl friend, Yu Bai, whom I shared smiles and tears with over the years, and I hope she will successfully defend her dissertation soon enough.

# Table of Contents

<b>Abstract</b> . . . . .	ii
<b>Acknowledgements</b> . . . . .	v
<b>List of Tables</b> . . . . .	x
<b>List of Figures</b> . . . . .	xii
<b>1. Introduction</b> . . . . .	1
1.1. References . . . . .	7
<b>2. The Accordion Procedure: A Method for Accommodating a Large Number of Attributes in Cognitive Diagnosis Modeling</b> . . . . .	10
2.1. Theoretical Framework . . . . .	13
2.1.1. CDMs . . . . .	13
2.1.2. Item Parameter Estimation of the G-DINA model . . . . .	15
2.2. The Proposed Procedure . . . . .	17
2.2.1. Collapsing a Q-matrix . . . . .	21
2.2.2. Item and Person Parameter Estimation of AP . . . . .	23
2.3. Simulation Studies . . . . .	23
2.3.1. Simulation Study 1 . . . . .	24



2.3.2. Simulation Study 2 . . . . .	32
2.4. Real-Data Illustration . . . . .	34
2.5. Conclusions and Discussions . . . . .	39
2.6. References . . . . .	42
<b>3. Improving Attribute Classification Accuracy in High Dimensional Data: A</b>	
<b>Four-Step Latent Regression Approach . . . . .</b>	<b>44</b>
3.1. CDMs . . . . .	46
3.1.1. AP . . . . .	48
3.2. Incorporating Covariates in CDMs . . . . .	49
3.3. Incorporating Covariates for the Accordion Procedure . . . . .	52
3.3.1. Classification Error Probabilities and Correction Weights . . . . .	53
3.3.2. Four-Step Approach . . . . .	55
3.4. Simulation Study . . . . .	57
3.4.1. Design . . . . .	57
3.4.2. Results . . . . .	62
3.5. Real-Data Illustration . . . . .	68
3.6. Discussions and Conclusions . . . . .	74
3.7. References . . . . .	78
<b>4. Strategies for Implementing CD-CAT in High-Dimensional Testing Situations</b>	<b>81</b>
4.1. Cognitive Diagnosis Models . . . . .	84
4.2. CD-CAT . . . . .	85

4.2.1.	Entry Level . . . . .	85
4.2.2.	Item Selection Method . . . . .	86
4.2.3.	Termination Rule . . . . .	88
4.3.	Calibrating Item Pools with High Dimensionality . . . . .	89
4.4.	The Modified-GDI . . . . .	90
4.4.1.	Estimating Prior Distributions using Covariates . . . . .	92
4.5.	Simulation Studies . . . . .	94
4.5.1.	Simulation Study 1 . . . . .	94
4.5.2.	Simulation Study 2 . . . . .	101
4.6.	Discussion . . . . .	107
4.7.	References . . . . .	110
<b>5.</b>	<b>Summary . . . . .</b>	<b>113</b>
5.1.	References . . . . .	117

## List of Tables

2.1. Q-matrix for $D = 2, K(d) = 3$ . . . . .	22
2.2. The Collapsed Q-matrix for $D = 2, K(d) = 3$ with $d = 1$ as the Target . .	22
2.3. Q-matrix for $D = 2, K(d) = 5, J = 50$ . . . . .	26
2.4. $CVC_d, CAC, CT(sec.)$ for $D = 2, K(d) = 5$ . . . . .	28
2.5. $CVC_d, CAC, CT(sec.)$ for $D = 2, K(d) = 8$ . . . . .	29
2.6. $CVC_d, CAC, CT(sec.)$ for $D = 4, K(d) = 5$ . . . . .	30
2.7. $CVC_d, CAC, CT(sec.)$ for $D = 4, K(d) = 8$ . . . . .	31
2.8. $RE_{CVC_d}, RE_{CAC},$ and $RE_{CT}$ for $D = 2, K(d) = 5$ . . . . .	32
2.9. $CVC_d, CAC, CT(sec.)$ for $D = 3, K(d) = 5$ . . . . .	34
2.10. Attributes Identified for the TIMSS 2007 Fourth-Grade Mathematics Book- lets 4 and 5 . . . . .	36
2.11. Q-matrix for the TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5	37
2.12. Sample Information . . . . .	38
2.13. $VAR_d, AAR, RE_{CT}$ for TIMSS 2007 Fourth-Grade Mathematics Data . .	39
3.1. Q-matrix for $D = 2, K(d) = 5, J = 20$ . . . . .	61
3.2. $CVC_d$ for $D = 4, K(d) = 5$ . . . . .	64
3.3. $CVC_d$ for $D = 4, K(d) = 8$ . . . . .	65

3.4. Attributes Identified for TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5 . . . . .	70
3.5. Q-matrix for TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5 . . . . .	71
3.6. McFadden's $R^2$ and $r_{pb}$ of Attribute Classification and Overall Mathematics Performance . . . . .	74
4.1. Posterior Probabilities and Success Probabilities for Latent Class $l'$ and $l(d)$	92

## List of Figures

2.1. Item <i>M031172</i> from TIMSS 2007 Fourth-Grade Mathematics . . . . .	20
3.1. Classification Certainty for $D = 4$ , $K(d) = 5$ , $N = 2000$ , Short Test, Low Item Quality . . . . .	66
3.2. Classification Certainty for $D = 4$ , $K(d) = 5$ , $N = 2000$ , Short Test, Medium Item Quality . . . . .	67
3.3. Classification Certainty for $D = 4$ , $K(d) = 5$ , $N = 2000$ , Short Test, High Item Quality . . . . .	67
3.4. Classification Certainty for $\alpha_{(2)3}$ , $\alpha_{(3)3}$ , and $\alpha_{(1)3}$ . . . . .	75
4.1. Proportion of Examinees Met Minimax Values Prior to CD-CAT, $D = 2$ , G-DINA . . . . .	98
4.2. Proportion of Examinees Met Minimax Values Prior to CD-CAT, $D = 2$ , DINA . . . . .	99
4.3. Test Length for Case 1, $D = 2$ , $K(d) = 5$ , G-DINA . . . . .	100
4.4. Test Length for Case 1, $D = 2$ , $K(d) = 8$ , G-DINA . . . . .	101
4.5. Test Length for Case 2, $D = 2$ , $K(d) = 5$ , G-DINA . . . . .	104
4.6. Test Length for Case 2, $D = 2$ , $K(d) = 8$ , G-DINA . . . . .	105
4.7. Shifts of Posterior Distribution for $\alpha = (1, 0, 1, 1, 1)$ across Testing Stages: Uniform Prior . . . . .	106

4.8. Shifts of Posterior Distribution for $\alpha = (1, 0, 1, 1, 1)$ across Testing Stages:	
Informative Prior . . . . .	107

# Chapter 1

## Introduction

In traditional summative assessments, which are mostly rooted in classical test theory or item response theory (IRT), students' abilities are evaluated by the summed scores or locations on a continuous proficiency scale. These assessments are useful in ranking students' proficiency levels, thus can be used for summative purposes, such as selection of candidates, college entrance admission, or decision on scholarships (de la Torre & Minchen, 2014). In contrast, cognitively diagnostic assessments (CDAs) are formative assessments designed to provide information regarding students' cognitive strengths and weaknesses on a set of fine-grained skills (de la Torre & Minchen, 2014; DiBello, Roussos, & Stout, 2007). CDAs can be useful in providing immediate diagnostic feedback, based on which adjustments to classroom instructions or remedial measures can be made.

Used in conjunction with CDAs, cognitive diagnosis models (CDMs) are a family of psychometric models that aim to classify examinees' proficiency profiles of skills (general referred to as attributes) based on their item responses (for some examples of CDMs, see de la Torre, 2011; DiBello et al., 2007; Hartz & Roussos, 2008; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008). Similar to IRT models, CDMs are also latent variable models that assume each examinee is associated with unobserved latent variables. IRT models differ from CDMs in that the

former work with continuous latent variable, whereas the latter assume the latent variable to be discrete. In most of the existing CDMs, attributes are treated as binary variables; therefore, if a test involves  $K$  attributes, the maximum number of latent classes is  $2^K$ , provided all latent classes are permissible, and each class represents a unique attribute profile. Hence, CDMs are also categorized as restricted latent class models (Haertel, 1989).

One of the key components of CDMs is the attribute-item association, called the Q-matrix (Tatsuoka, 1983, 1985). The Q-matrix is essentially a loading indicator matrix specifying which attributes each item measures (Rupp & Templin, 2008). The development of the Q-matrix for a CDA often involves collaborative undertaking between content and psychometrics experts, among others. As an example, Tjoe and de la Torre (2014) detailed the necessary steps in identifying required attributes and developing a Q-matrix for a proportional reasoning test for middle school students.

Over a dozen CDMs have been developed in the past with different assumptions on the underlying cognitive processes. For example, the deterministic inputs, noisy “AND” gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model assumes an examinee needs to possess all required skills to answer an item correctly, whereas the deterministic inputs, noisy “OR” gate (DINO; Templin & Henson, 2006) model assumes an examinee only needs at least one of the required attributes to answer an item correctly. Other models assume that the response to an item is an additive process of the required attributes. As an example, the additive-CDM (A-CDM; de la Torre, 2011) formulates the probability of success of an item as the sum of the effects of the required attributes. In addition to the above specific CDMs, general CDMs have also been proposed, namely, the generalized-DINA (G-DINA; de la Torre, 2011) model, the general diagnostic model (GDM; von Davier, 2008), and the



log-linear CDM (LCDM; Henson et al., 2009) to accommodate more complex underlying processes. General CDMs do not impose constraints on item parameters so that the probability of success of each distinguishable latent group can be freely estimated.

Previous literature has offered a number of parametric approaches to fit CDMs, which include the marginal maximum likelihood estimation with expectation maximization (MMLE-EM; de la Torre, 2009, 2011; von Davier, 2008) algorithm, joint maximum likelihood estimation algorithm (JMLE; Chiu, Köhn, Zheng, & Henson, 2016; de la Torre, 2009), and Markov chain Monte Carlo (MCMC; de la Torre, 2009; Hartz, 2002; Henson et al., 2009; Templin & Henson, 2006) algorithm. In addition, clustering (Chiu, Douglas, & Li, 2009) and nonparametric methods (Chiu & Douglas, 2013; Chiu, Sun, & Bian, 2018) have also been proposed for classifying examinees' latent classes without fitting a CDM. Regardless of which approach is used to classify examinees, a common problem persists - the curse of dimensionality (Bellman, 1957). In the CDM context, this refers to the problem that the more attributes a test involves, the more difficult it is to classify examinees accurately, because the number of latent classes grows exponentially with the number of attributes. For parametric approaches, although theoretically, the number of attributes that can be analyzed by a CDM is unlimited, in practice this number may not exceed 20 due to a number of computational limitations. Expanding the capacity of CDMs to accommodate higher-dimensional data will dramatically improve their use in practice.

Traditionally, most psychometric models rely solely on item responses. However, some research exploited ancillary information about examinees to improve the precision of person and item parameter estimates, particularly for IRT models (see Ackerman & Davey, 1991; de la Torre & Patz, 2005; Kahraman & Kamata, 2004; Mislevy, 1987; Mislevy &

Sheehan, 1989). Although not as extensively explored in the context of CDMs, two latent regression approaches to incorporate covariates can be found in the literature, namely, a one-step approach (e.g., Ayers, Rabe-Hesketh, & Nugent, 2013; Park & Lee, 2014; Park, Xing, & Lee, 2018) or a multi-step approach (Iaconangelo, 2017). The one-step approach provides unbiased parameter estimates for both CDMs and latent regressions, however, the one-step approach is not as flexible as the multi-step approach when adjustments to the CDMs or the latent regressions need to be made (Iaconangelo, 2017).

Adopting the core concept of IRT-based computerized adaptive testing (IRT CAT), cognitive diagnosis computerized adaptive testing (CD-CAT; for examples, see Cheng, 2009; Hsu, Wang, & Chen, 2013; Kaplan, de la Torre, & Barrada, 2015; McGlohen & Chang, 2008) is a recent development in the realm of CDMs. Akin to IRT CAT, CD-CAT can increase the testing efficiency by administering items tailored to each examinee's proficiency profile. The entry level, item selection method, and termination rule are considered as key components in an IRT CAT system (Weiss & Kingsbury, 1984), which can be fully adopted in the context of CD-CAT. For example, various item selection methods for CD-CAT can be found in the literature, which choose the optimal set of items for each examinee based on certain psychometric indices, such as the Kullback-Leibler (KL) divergence-based methods (Cheng, 2009; Cover & Thomas, 1991; Kaplan et al., 2015; Xu, Chang, & Douglas, 2003), Shannon entropy (SHE; Xu et al., 2003) method, and G-DINA model discrimination index (GDI; Kaplan et al., 2015) method.

High-dimensional data are not rare in educational applications, especially in large-scale assessments. For example, the assessment framework of the Trends in International Mathematics and Science Study (TIMSS) 2007 identified 38 skills in three domains for

the fourth-grade mathematics (Mullis et al., 2005). Although the TIMSS 2007 was not designed for formative purposes, researchers have re-analyzed part of the item response data containing 15 skills using the DINA model, and extracted rich diagnostic information that can be used to inform classroom instructions (Lee, Park, & Taylan, 2011). Because the number of latent classes grows exponentially with the number of attributes, estimating CDMs using the complete set of 38 skills involved in the TIMSS 2007 dataset will involve evaluating more than 200 billion latent classes, which is impossible for the existing estimation methods to handle. Thus, there is still a need for methods that can accommodate a large number of skills to improve the practicability of CDMs in large-scale educational applications.

As the dimensionality of a test increases, it becomes more difficult to accurately classify examinees' attribute profiles, which consequently affects the use and interpretation of diagnostic results. Although a few previous research has considered incorporating ancillary information into CDMs, few of them can be adopted without modifications in high-dimensional testing situations. For example, estimation of a one-step approach becomes unfeasible when a test measures more than 20 attributes. Moreover, few research has considered expanding the usage of CD-CAT in high-dimensional testing situations. In fact, most of the previous research in CD-CAT only considers up to eight attributes (e.g., Cheng, 2009; Hsu et al., 2013; Kaplan et al., 2015; Liu, You, Wang, Ding, & Chang, 2013). Entry level, item pool calibration, and item selection method in high-dimensional CD-CAT are issues waiting to be addressed.

This dissertation consists of three studies aiming to provide strategies to solve estimation problems, improve examinees' classification accuracy, and make adaptive testing

feasible for high-dimensional assessment situations. Study I, given in Chapter 2, proposes an efficient procedure to estimate examinees' proficiency classes in situations where attributes can be partitioned into non-overlapping subsets. The performance of the proposed procedure is evaluated empirically by both simulated and real-data studies. Study II, given in Chapter 3, is grew out of the issues found in Chapter 2: that is, classification accuracy suffers when tests do not provide enough information. Aside from improving the quality of the test, an alternative to enhance the information obtained from CDMs is to incorporate other sources of information. Hence, in Chapter 3, a multi-step procedure incorporating covariates is proposed to improve the classification accuracy, the performance of which is examined by both simulation and real-data studies. Last, Study III, given in Chapter 4, explores the implementation of CD-CAT in the context of high-dimensional testing situations. A series of strategies, based on the techniques developed in the previous studies, such as high-dimensional item pool calibration method and estimation of examinees' prior distributions, are proposed to make high-dimensional CD-CAT feasible and more efficient. In the last chapter, summaries of key findings from each study and remarks on the contributions of this dissertation are provided.

## 1.1 References

- Ackerman, T. A., & Davey, T. C. (1991). *Concurrent adaptive measurement of multiple abilities*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification*, *30*, 195–224.
- Bellman, R. (1957). *Dynamic programming*. Princeton: Princeton University Press.
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.
- Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225–250.
- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.
- Chiu, C.-Y., Köhn, H.-F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika*, *81*, 1069–1092.
- Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*, 355–375.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.
- de la Torre, J., & Minchen, N. D. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89–97.
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311.
- DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 970–1030). Amsterdam: North-Holland Publications.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.
- Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). Retrieved from

<http://hdl.handle.net/2142/87393>.

- Hartz, S. M., & Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series, 2008*.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.
- Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563–582.
- Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models* (Doctoral dissertation). Retrieved from <https://doi.org/doi:10.7282/T3W95D95>.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.
- Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement, 28*, 407–426.
- Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167–188.
- Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.
- Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an english achievement test in china. *Journal of Classification, 30*.
- McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808–821.
- Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81–91.
- Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661–679.
- Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38*, 376–390.

- Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied Psychological Measurement, 42*, 376-392.
- Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78-96.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics, 10*, 55-73.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237-255.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.
- Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

## Chapter 2

# The Accordion Procedure: A Method for Accommodating a Large Number of Attributes in Cognitive Diagnosis Modeling

### Abstract

Cognitive diagnosis models (CDMs) refer to a family of psychometric models that aim to classify examinees' proficiency profiles on a set of fine-grained skills based on their item responses. Theoretically, the number of attributes that can be estimated by a CDM is unlimited; however, in practice, this number may not be large due to model identifiability and computational issues. These issues constrain the use of CDMs in scenarios where a comprehensive diagnosis of a complete knowledge space, such as large-scale diagnostic assessments or retrofitting summative assessments using CDMs, is of interest. In this study, the accordion procedure is proposed to address the estimation issue in scenarios where attributes can be partitioned into non-overlapping subsets based on the higher-order knowledge structure. Simulation and real-data studies are conducted to examine the performance of the proposed method. Results show that the procedure can yield comparatively accurate classifications with considerably low computation time compared to complete-profile estimation procedure.

**Keywords:** accordion procedure, cognitive diagnosis model, high-dimensional data



Cognitive diagnosis models (CDMs; de la Torre, 2011; DiBello, Roussos, & Stout, 2007; Hartz & Roussos, 2008; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008) have sparked interest among educational measurement researchers and practitioners for its capability to provide formative information on student mastery or non-mastery of a set of fine-grained skills. One of the advantages of CDMs is that, by treating latent variables to be discrete, usually binary, CDMs can accommodate a higher-dimensional latent space than multidimensional latent trait (e.g., multidimensional item response theory) models. Because each binary attribute vector forms a unique latent class, CDMs are also deemed as restricted latent class models with pre-specified label for each latent class. Over a dozen CDMs can be found in the literature, each of which assumes certain underlying cognitive processes on how examinees use skills to solve problems. For example, the deterministic inputs, noisy “AND” gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model assumes an examinee needs to possess all required skills to answer an item correctly, whereas the deterministic inputs, noisy “OR” gate (DINO; Templin & Henson, 2006) model assumes an examinee only needs at least one of the required attributes to answer an item correctly. Aside from these specific CDMs with stringent underlying assumptions, three general CDMs have been proposed in the past decade, namely, the generalized-DINA (G-DINA; de la Torre & Chiu, 2016) model, the general diagnostic model (GDM; von Davier, 2008), and the log-linear CDM (LCDM; Henson et al., 2009) to accommodate more complex data. General CDMs do not impose constraints on item parameters so that the probability of success of each distinguishable latent group can be freely estimated.

Parameters in CDMs are often estimated by the marginal maximum likelihood estimation with expectation maximization (MMLE-EM; de la Torre, 2009, 2011; von Davier, 2008) algorithm, joint maximum likelihood estimation algorithm (JMLE; Chiu, Köhn, Zheng, & Henson, 2016; de la Torre, 2009), and Markov chain Monte Carlo (MCMC; de la Torre, 2009; Hartz, 2002; Henson et al., 2009; Templin & Henson, 2006) algorithm. Due to practical concerns with estimating high-dimensional data, researchers need to balance the grain size and the number of attributes involved in a test: the finer the grain size is, the more attributes the test has to include, and consequently, the more difficult to fit CDMs. This limits the use of CDMs in scenarios where a comprehensive diagnosis over skills from multiple content domains is of interest, such as large-scale diagnostic assessments or retrofitting summative assessments using CDMs.

Theoretically, the number of attributes that can be estimated by a CDM is unlimited; however, in practice, high dimensionality is likely to cause identifiability issues or unreliable parameter estimates (Huebner, 2010; von Davier, 2008). Moreover, because the number of latent classes grows exponentially with the number of attributes, when MMLE-EM algorithm is applied to estimate item parameters, marginalization over all possible latent classes can be computationally intensive and time consuming. For example, widely used estimation packages, such as the GDINA (Ma & de la Torre, 2016) and the CDM (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016) cannot handle  $K = 20$  attributes, which involves more than one million possible latent classes, due to lack of memory.

To address the estimation issue in high-dimensional data, this study proposes a procedure that focuses only on the attributes of one particular subset of attributes at a time, while the attributes of each of the remaining subsets are collapsed to create composite nuisance

attributes. The proposed procedure can be used in situations where the large number of attributes can be partitioned into non-overlapping subsets and mastery of attributes in each subset is governed by a higher-order latent trait. The proposed procedure is applicable in both cross-sectional and longitudinal contexts.

The rest of this article is organized as follows. An introduction to CDMs will be given first, followed by the proposed procedure. The designs and results of two simulation studies and a real-data analysis are provided next. Finally, limitations and future research directions are discussed to conclude the study.

## 2.1 Theoretical Framework

### 2.1.1 CDMs

CDMs are restricted latent class models that are used in conjunction with diagnostic assessments to provide detailed feedback on students' strengths and weaknesses on a set of fine-grained attributes. As one of the general CDMs (see; de la Torre, 2011; Henson et al., 2009; von Davier, 2008), the G-DINA model (de la Torre, 2011) provides a unified framework to describe various CDMs. Let  $K_j$  be the total number of required attributes for item  $j$ . Item  $j$  can distinguish examinees into  $2^{K_j}$  latent groups, each of which represents a reduced attribute profile denoted by  $\alpha_{lj}$ , and  $l_j = 1, \dots, 2^{K_j}$ . For example, if the  $q$ -vector for item  $j$  is  $\mathbf{q}_j = (1, 1, 0)$ , it can distinguish four latent groups with reduced attribute profiles of  $\alpha_{1j} = (0, 0, \cdot)$ ,  $\alpha_{2j} = (1, 0, \cdot)$ ,  $\alpha_{3j} = (0, 1, \cdot)$ , and  $\alpha_{4j} = (1, 1, \cdot)$ , where  $\cdot$  represents a free entry. In this regard, latent classes that are distinct only at the third attribute cannot be further distinguished, and thus can be combined. The G-DINA model in the identity link

can be written as

$$P(Y_{ij} = 1 | \alpha_{lj}) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (2.1)$$

where  $\delta_{j0}$  is the intercept for item  $j$ ,  $\delta_{jk}$  is the  $k$ th attribute's main effect,  $\delta_{jkk'}$  is the two-way interaction between attribute  $k$  and  $k'$ , and  $\delta_{j12\dots K_j}$  is the  $K_j$ -way interaction.

With different link functions, the G-DINA model is equivalent to other general CDMs. For example, using the logistic link, the G-DINA model is equivalent to the LCDM (Henson et al., 2009). In addition, by setting proper constraints to the item parameters in the G-DINA model, several specific CDMs can be reparameterized using the G-DINA model. For example, the DINA (Haertel, 1989; Junker & Sijtsma, 2001) model can be derived from the G-DINA model by constraining all effects except the  $\delta_0$  and  $\delta_{j12\dots K_j}$  to zero. Its probability of success can be then written as

$$P(Y_{ij} = 1 | \alpha_{lj}) = \delta_{j0} + \delta_{j12\dots K_j} I(\alpha_{lj} = \mathbf{1}), \quad (2.2)$$

where  $I(\alpha_{lj} = \mathbf{1})$  is the indicator function determining whether  $\alpha_{lj}$  contains all of the required attributes. The intercept,  $\delta_{j0}$ , is equivalent to the guessing parameter,  $g_j$ , in the original parameterization of the DINA model, and  $1 - (\delta_{j0} + \delta_{j12\dots K_j})$  is equivalent to the slip parameter,  $s_j$ . Other specific CDMs, such as the DINO (Templin & Henson, 2006) model, additive-CDM (A-CDM; de la Torre, 2011), or reduced reparameterized unified model (R-RUM, Hartz, 2002; Hartz & Roussos, 2008) can also be derived from the G-DINA model.

### 2.1.2 Item Parameter Estimation of the G-DINA model

The MMLE-EM algorithm is commonly used to estimate item parameters for the G-DINA model (de la Torre, 2011). To implement this algorithm, the conditional likelihood,  $L(\mathbf{Y}_i|\boldsymbol{\alpha}_l)$ , is marginalized across all possible latent classes. Therefore, the resulting marginalized likelihood only depends on the item parameters. In practice, the logarithm of the marginalized likelihood, denoted by  $l(\mathbf{Y})$ , is usually taken, and can be written as

$$l(\mathbf{Y}) = \log \prod_{i=1}^N \sum_{l=1}^{2^K} L(\mathbf{Y}_i|\boldsymbol{\alpha}_l)\pi(\boldsymbol{\alpha}_l), \quad (2.3)$$

where  $\pi(\boldsymbol{\alpha}_l)$  is the prior probability of latent class  $l$ .

Equation 2.3 is maximized to find the estimates of item parameters. It has been shown in de la Torre (2011) that there exists the closed-form solution for  $\hat{P}(Y_{ij} = 1|\boldsymbol{\alpha}_{lj})$  which is given by

$$\hat{P}(Y_{ij} = 1|\boldsymbol{\alpha}_{lj}) = \frac{\sum_{i=1}^N P(\boldsymbol{\alpha}_{lj}|\mathbf{Y}_i) \times Y_{ij}}{\sum_{i=1}^N P(\boldsymbol{\alpha}_{lj}|\mathbf{Y}_i)}, \quad (2.4)$$

where  $P(\boldsymbol{\alpha}_{lj}|\mathbf{Y}_i)$  is the posterior probability of assigning examinee  $i$  into latent group  $lj$  given the examinee's response pattern. The numerator of Equation 2.4 represents the expected number of correct answers in latent group  $lj$ , and the denominator represents the expected size of latent group  $lj$ .

As discussed by de la Torre (2011), the joint attribute distribution in Equation 2.3,  $\pi(\boldsymbol{\alpha}_l)$ , can be formulated in different ways. First, a fixed prior distribution can be used if

researchers have knowledge about attribute distribution a priori. Another way is to use a saturated formulation of the joint attribute distribution, where the proportion of each latent class is estimated. Moreover, the joint distribution can also be simplified by imposing hierarchical structures on the attributes to reduce the number of admissible latent classes (Akbay, 2017; Leighton, Gierl, & Hunka, 2004). Last, as suggested by de la Torre (2011), when the number of attributes is large, a saturated joint distribution may not be feasible, therefore, a higher-order (HO; de la Torre & Douglas, 2004, 2008) attribute structure can be imposed to formulate  $\pi(\alpha_l)$ . Specifically, for examinee  $i$ , a latent trait  $\theta_i$  is posited such that the components of  $\alpha$  are assumed to be independent conditional on  $\theta_i$ . The linear logistic model can be used for this purpose, and is given by

$$\text{logit}[P(\alpha_{ik}|\theta_i)] = a_k(\theta_i - b_k), \quad (2.5)$$

where  $a_k$  and  $b_k$  are the slope and difficulty parameters, respectively, associated with attribute  $k$ . The joint attribute distribution conditional on  $\theta$  can be written as

$$P(\alpha_i|\theta_i) = \prod_{k=1}^K P(\alpha_{ik}|\theta_i)^{\alpha_{ik}} [1 - P(\alpha_{ik}|\theta_i)]^{(1-\alpha_{ik})}. \quad (2.6)$$

## 2.2 The Proposed Procedure

As discussed in de la Torre and Minchen (2014), a skill can be viewed as a rudimentary component that co-exists with other components in a similar larger domain. The hierarchy of skills reflects different granularity levels of a content domain. When the grain size of a skill becomes finer, the details embedded in a skill becomes richer. Take the assessment framework for fourth-grade mathematics of the TIMSS 2007 as an example. TIMSS is an international educational assessment program that is designed to compare mathematics and science performance across participant countries, and contains the core skills taught in the fourth and eighth-grade classrooms (Mullis et al., 2005). Three key content domains were identified in TIMSS 2007 for the fourth-grade mathematics, namely, number, geometric shapes and measures, and data display. The number domain consists of four key topics, namely, whole numbers, fractions and decimals, number sentences, and patterns and relationships. As the easiest introduction to operations with whole numbers in the primary school at the fourth-grade level, students should be able to compute whole numbers with reasonable size, estimate sums, differences, products and quotients, and solve problems using computations. TIMSS 2007 framework reflects these academic expectations of four-grade students, and considered 8 specific skills, such as *representing whole numbers using words, diagrams, or symbols*, and *demonstrating knowledge of place value, including recognizing and writing numbers in expanded form*. This example demonstrates a similar knowledge structure discussed in de la Torre and Song (2009) that different levels of abilities and skills can be identified for a test. Also, as shown in this example, skills from the same domain are usually more homogeneous than skills from different domains.

A testing framework such as this provides insights to develop the proposed procedure to reduce dimensionality when estimating CDMs for high-dimensional data by collapsing multiple skills into coarser-grained attributes for domains that are not the targets.

To better demonstrate an assessment involves skills from multiple domains, several notations need to be modified. Suppose there are  $D$  content domains involved in a test, and each domain consists of  $K(d)$  skills, where  $d = 1, \dots, D$ . Then the total number of attributes is  $K = \sum_d K(d)$ . The attribute profile for domain  $d$  is denoted by  $\alpha_{l(d)} = (\alpha_{l(d)1}, \dots, \alpha_{l(d)K(d)})$ , where  $l(d) = 1, \dots, 2^{K(d)}$ . We propose the accordion procedure (AP), to solve the estimation issue for high-dimensional data by estimating the entire knowledge space by parts. Specifically, when focusing on one domain, which is called the target domain, attributes from nontarget domains are combined or collapsed as domain-level attributes. Because the granularity of the collapsed attributes is coarser than the size of interest, the collapsed attributes only serve as statistical controls of the possible effects from nontarget domains when fitting CDMs, hence, are called nuisance attributes. AP can be either applied once to the target domain or  $D$  times to obtain the complete attribute profile depending on the research purpose.




AP can be particularly useful when items measure multiple skills in different domains in a test. For example, a released item of TIMSS 2007 fourth-grade mathematics, namely, item *M031172* and its  $q$ -vector are given in Figure 2.1. The  $q$ -vector for this item is developed by Lee, Park, and Taylan (2011). Although this item is categorized in the data display domain in the assessment framework, it involves two skills from another domain as well. The four skills that are required to answer this item are: *representing*, *comparing*,




*and ordering whole numbers as well as demonstrating knowledge of place value (number), recognize multiples, computing with whole numbers using the four operations, and estimating computations (number), read data from tables, pictographs, bar graphs, and pie charts (data display) and comparing and understanding how to use information from data (data display).* When the focus of an analysis is on the data display domain, the skills from number domain can be collapsed to a coarser composite attribute.


In general, AP can be applied to two possible scenarios. First, when retrofitting extant large-scale assessments associated with a large set of skills in CDMs, AP can be used to address the estimation problem associated with fitting a CDM to the entire attribute vectors all at once. Another possible scenario is tracking students learning progress over a period of time when skills are taught sequentially based on their difficulties. For instance, suppose there is an educational program where the “easy” set of skills are taught at the beginning of the program, followed by the “medium” difficulty set of skills, and lastly the “hard” skills. After each set of skills are taught, students’ attribute profiles are estimated for tracking their learning progress. Then at each check point, the grain size of previously learned skills and skills to be learned in the future can be adjusted to a coarser level to reduce dimensionality.

## Item 25: M031172 (Data Display: Organizing &amp; Representing)

Street	Number of houses
Main	
Center	
First	
Hill	

Mary is making a chart to show the number of houses on some streets.

Every  stands for 5 houses. There are 20 houses on Hill Street.

How many  should Mary put in the chart beside Hill Street?

- (A) 4  
 (B) 5  
 (C) 15  
 (D) 20

Copyright © 2008 International Association for the Evaluation of Educational Achievement (IEA). All rights reserved.

Item	Number								Geometric Shapes & Measures				Data Display		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
25	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1

Figure 2.1: Item M031172 from TIMSS 2007 Fourth-Grade Mathematics

An example of collapsing previously taught skills in a diagnostic assessment can be found in Tjoe and de la Torre (2014), where the focus of this assessment is the eighth-grade proportional reasoning. The researchers identified the attribute *prerequisite skills and concepts required in proportional reasoning* which includes necessary mathematics knowledge (e.g., addition, subtraction, multiplication, and division) taught before the eighth-grade. This attribute can be viewed as a composite attribute created by combining

nontarget skills.

### 2.2.1 Collapsing a Q-matrix

To apply AP to a target domain, the grain size of skills in nontarget domains needs to be adjusted. As a necessary input for CDMs, the original Q-matrix needs to be modified correspondingly by collapsing attributes in nontarget domains to form a new Q-matrix for AP analyses. Different methods can be used to collapse a Q-matrix given a target domain. For example, the disjunctive condensation rule can be applied, such that if an item measures at least one of the attributes of a domain, the collapsed  $q$ -entry is coded as 1, otherwise 0. This rule can be expressed as

$$q_{j(d')} = I\left(\sum_{k(d')=1}^{K(d')} q_{jk(d')} \geq 1\right), \quad (2.7)$$

where  $q_{j(d')}$  is the collapsed attribute for nontarget domain  $d'$ , and  $q_{jk(d')}$  is the original  $q$ -entry of the  $k$ th entry in domain  $d'$ . To obtain the  $q$ -vector for item  $j$  for AP, the target attributes and collapsed nuisance attributes need to be concatenated. For example, assume domain  $d$  is the target domain. By collapsing the attributes in other domains, the  $q$ -vector of item  $j$  for AP can be written as  $\mathbf{q}_j^{(AP)} = (q_{j1(d)}, \dots, q_{jK(d)}, q_{j(1)}, \dots, q_{j(d-1)}, q_{j(d+1)}, \dots, q_{j(D)})$ . In doing so, the number of dimensions is reduced from  $K = \sum_d K(d)$  to  $K' = K(d) + D - 1$ .

Table 2.1 shows an example of a Q-matrix for three items involving six skills from two domains. Item 1 requires the first attribute from domain 1 and the first attribute from domain 2. Item 2 requires the first and second skills only from domain 1. Item 3 requires

the first and second skills from domain 2, and the third skill from domain 1. The total number of possible latent classes in this example is  $2^6 = 64$ . With domain 1 as the target, attributes in domain 2 can be collapsed. The collapsed Q-matrix for domain 2 is given in Table 2.2 using the disjunctive condensation rule. The total number of latent classes for the collapsed Q-matrix is reduced to  $2^4 = 16$ .

Table 2.1: Q-matrix for  $D = 2, K(d) = 3$

item	Domain 1			Domain 2		
	$\alpha_{1(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{1(2)}$	$\alpha_{2(2)}$	$\alpha_{3(2)}$
1	1	0	0	1	0	0
2	1	1	0	0	0	0
3	0	0	1	1	0	1

Table 2.2: The Collapsed Q-matrix for  $D = 2, K(d) = 3$  with  $d = 1$  as the Target

item	Domain 1			Domain 2
	$\alpha_{1(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{(2)}$
1	1	0	0	1
2	1	1	0	0
3	0	0	1	1

Other condensation rules can be also applied in AP. For example, the  $q$ -entry for a nontarget domain is coded as 1 if a certain number of attributes are required by an item, otherwise 0. Furthermore, experts' judgment should also be considered when developing the collapsed Q-matrix. Because skills within a domain may not be equally important to represent the domain, collapsing a Q-matrix based on the disjunctive rule may not well summarize the domain. Therefore, with the input of experts, a more comprehensive decision can be made in developing a collapsed Q-matrix.

## 2.2.2 Item and Person Parameter Estimation of AP

The MMLE-EM algorithm can be utilized to estimate the item parameters of AP when fitting the G-DINA model using the collapsed Q-matrix without additional modifications. However, because the collapsed Q-matrix contains nuisance attributes, the estimated posterior distribution needs to be modified to obtain attribute profile estimates for the target domain. Let  $l(d)$  denote the latent class for domain  $d$ , where  $l(d) = 1, \dots, 2^{K(d)}$ . The posterior probability for examinee  $i$  with attribute profile  $\alpha_{l(d)}$  can be written as

$$\hat{\pi}_i(\alpha_{l(d)}) = \sum_{\substack{\alpha_{(d')}=0 \\ d' \neq d}}^1 \dots \sum_{\substack{\alpha_{(d'')}=0 \\ d'' \neq d}}^1 \hat{\pi}_i(\alpha_{l'}), \quad (2.8)$$

where  $\hat{\pi}_i(\alpha_{l(d)})$  is the posterior probability for examinee  $i$  in latent class  $l(d)$ ,  $\alpha_{d'}$  and  $\alpha_{d''}$  are two nuisance attributes,  $\hat{\pi}_i(\alpha_{l'})$  is the estimated posterior probability for examinee  $i$  in latent class  $l'$  (i.e., based on both the target and nuisance domains), and  $l' = 1, \dots, 2^{K'}$ .

The estimates of examinees' domain-specific attribute profiles,  $\hat{\alpha}_{i(d)}$ , can be obtained by expected a posteriori (EAP) or maximum a posteriori (MAP) based on  $\hat{\pi}_i(\alpha_{l(d)})$ . In the same manner, the EAP or MAP of a single attribute,  $\hat{\alpha}_{ik(d)}$ , can be obtained by further marginalizing  $\hat{\pi}_i(\alpha_{l(d)})$ .

## 2.3 Simulation Studies

Two simulation studies were conducted to examine the performance of AP in two different scenarios. The first study resembled a high-dimensional cross-sectional testing scenario,

where attributes from multiple domains were involved. The second study focused on examining the feasibility of utilizing AP to analyze longitudinal data, where skills can be grouped based on their levels of difficulty.

### 2.3.1 Simulation Study 1

#### *Design*

Six factors were manipulated in simulation study 1, namely, the number of domains ( $D = 2$  and 4), number of attributes per domain ( $K(d) = 5$  and 8), test length ( $3 \times K$  and  $5 \times K$ ), sample size ( $N = 500, 1000$  and 5000), item quality (low, medium, and high), and generating model (G-DINA and DINA). By multiplying  $D$  by  $K(d)$ , the total number of attributes considered in this study were  $K = 10, 16, 20$ , and 32. The item quality was manipulated by setting the lowest and highest probabilities of success, denoted by  $P_0$  and  $P_1$ , respectively, for each item. Specifically, when item quality is low,  $P_0 \sim U(.25, .35)$  and  $P_1 \sim U(.65, .75)$ ; when item quality is medium,  $P_0 \sim U(.15, .25)$  and  $P_1 \sim U(.75, .85)$ ; and when item quality is high,  $P_0 \sim U(.05, .15)$  and  $P_1 \sim U(.85, .95)$ . Probabilities for other latent classes were generated with monotonicity constraints assuming that mastering additional required attributes resulted in non-descending probability of success. Attribute profiles were estimated using EAP.

The true attribute profiles were generated using the HO model. Specifically, the HO  $\theta = \{\theta_d\}$  were generated from  $\mathcal{N}_D(\mathbf{0}, \Sigma)$ , where the diagonal elements of  $\Sigma$  were set to 1 and the off-diagonal elements were set to .5, which implied the HO  $\theta$ s were moderately correlated. The slope parameter  $a_{k(d)}$  for the HO model was fixed to 1 and the

difficulty parameter  $b_{k(d)}$  was sampled from  $N(0, 1)$ . The probability of mastering individual attribute,  $P(\alpha_{k(d)}|\theta_d)$ , were generated by Equation 2.6, and subsequently converted to binary attributes.

Q-matrices were created according to  $D$ ,  $K(d)$  and  $J$ . The general rules of constructing the Q-matrices can be summarized as follows. First, each Q-matrix was required to contain at least one identity matrix to guarantee the property of Q-completeness (Köhn & Chiu, 2017). Second, each attribute was measured by the same number of items. Third, there existed 40% to 60% items measuring attributes across different domains. As an example, the Q-matrix for  $D = 2$ ,  $K(d) = 5$  and  $J = 50$  is given in Table 2.3. As shown in this Q-matrix, the first 20 items are single-attribute items, and the rest are two-attribute and three-attribute items that measure skills across two domains.

Two CDMs were used to generate item responses, namely, the G-DINA and DINA model. For both generating models, probabilities of success for latent classes were generated according to the item quality. One hundred replications were generated for each condition.

Table 2.3: Q-matrix for  $D = 2, K(d) = 5, J = 50$ 

Item 1 - 25										Item 26-50									
$\alpha_{1(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{4(1)}$	$\alpha_{5(1)}$	$\alpha_{1(2)}$	$\alpha_{2(2)}$	$\alpha_{3(2)}$	$\alpha_{4(2)}$	$\alpha_{5(2)}$	$\alpha_{1(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{4(1)}$	$\alpha_{5(1)}$	$\alpha_{1(2)}$	$\alpha_{2(2)}$	$\alpha_{3(2)}$	$\alpha_{4(2)}$	$\alpha_{5(2)}$
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0
0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0
0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1



The domain-level vector-wise classification accuracy ( $CVC_d$ ) and the correct attribute-wise classification accuracy (CAC) were calculated to compare the agreement of classification results with the true attribute profiles and elements. The CAC and  $CVC_d$  were given by

$$CAC = \sum_{i=1}^N \sum_{d=1}^D \sum_{k(d)=1}^{K(d)} \frac{I[\alpha_{ik(d)} = \hat{\alpha}_{ik(d)}]}{NDK(d)}, \quad (2.9)$$

and,

$$CVC_d = \sum_{i=1}^N \sum_{d=1}^D \frac{I[\alpha_{i(d)} = \hat{\alpha}_{i(d)}]}{ND}. \quad (2.10)$$

In addition to the classification accuracy, the computation time (in seconds), denoted by CT(sec.), was also recorded for AP. The performance of AP was compared with the complete-profile estimation (CPE) approach via regular MMLE-EM algorithm. Due to the dramatic increase in CT and memory usage when  $K$  was above 10, CPE became intractable for  $K = 16, 20$  and  $32$  conditions. Hence, CPE was only applied to the  $D = 2$ , and  $K(d) = 5$  conditions. To better compare the results, the relative efficiency (RE) of AP against the CPE was calculated. The RE of an evaluation index was defined as the ratio between the classification accuracy or computation times for AP and CPE.

### *Results*

The results of classification accuracy and CT for simulation study 1 are given in Tables 2.4 to 2.7. The results for the largest sample size,  $N = 5000$ , were very similar to those for  $N = 1000$ , thus were omitted from the tables. When  $D = 2$  and  $K(d) = 5$ , as shown in Table 2.4, item quality had a huge impact on the classification accuracy. For example, when  $N =$

500,  $CVC_d$  increased from .17 and .20 to .75 and .64, for fitting the G-DINA and DINA model in AP, respectively. Similar patterns were found for CAC as well. Furthermore, when item quality increased, the CT dramatically decreased for fitting both models to the data. For example, when  $N = 500$ , fitting the G-DINA model to data with low item quality required 4.6 seconds of computation time, whereas only required 1.1 seconds to data with high item quality.

Table 2.4:  $CVC_d$ , CAC, CT(sec.) for  $D = 2$ ,  $K(d) = 5$

$N$	Item Quality	Model	$J = 30$			$J = 50$		
			$CVC_d$	CAC	CT(sec.)	$CVC_d$	CAC	CT(sec.)
500	Low	G-DINA	.17	.70	4.58	.22	.73	5.92
		DINA	.20	.72	3.39	.27	.76	2.76
	Medium	G-DINA	.43	.84	2.43	.53	.88	2.80
		DINA	.45	.85	1.73	.48	.86	1.22
	High	G-DINA	.75	.94	1.09	.80	.96	1.54
		DINA	.64	.91	.90	.74	.94	1.19
1000	Low	G-DINA	.20	.72	8.29	.26	.76	11.30
		DINA	.24	.74	4.51	.30	.77	3.49
	Medium	G-DINA	.49	.86	3.90	.57	.89	4.15
		DINA	.48	.86	2.52	.50	.86	2.02
	High	G-DINA	.77	.95	1.68	.81	.96	2.30
		DINA	.64	.92	1.39	.76	.94	1.72

In addition, sample size and test length, as expected, had a positive impact on the classification accuracy. As the sample size doubled from 500 to 1000,  $CVC_d$  increased slightly from .01 to .06 when  $D = 2$  and  $K(d) = 5$ . The slight improvement was also found in CAC when sample size was doubled. When the text length was increased from 30 to 50, there was .02 to .12 increase in  $CVC_d$ , and 0 to .04 increase in CAC. For the computational efficiency, however, mixed results were found. For instance, although the CT for fitting

the G-DINA model increased with the increase in test length, the CT for fitting the DINA model decreased with longer test when the item quality was not high. Furthermore, doubling the sample size resulted in longer CT, especially for fitting the G-DINA model to data with low item quality. For example, as shown in Table 2.4, fitting the G-DINA model required 4.6 seconds when  $N = 500$  and  $J = 30$ , and 8.3 seconds when sample size was doubled. Comparing the different models fitted to the data, although the resulting CAC and  $CVC_d$  was similar to each other, fitting the DINA model was computationally faster than the G-DINA model, particularly when the item quality was low.

Table 2.5:  $CVC_d$ , CAC, CT(sec.) for  $D = 2$ ,  $K(d) = 8$

$N$	Item	Model	$J = 48$			$J = 60$		
			$CVC_d$	CAC	CT(sec.)	$CVC_d$	CAC	CT(sec.)
500	Low	G-DINA	.06	.70	38.02	.09	.73	65.33
		DINA	.07	.71	33.86	.12	.76	32.43
	Medium	G-DINA	.24	.83	23.69	.35	.88	34.33
		DINA	.28	.85	17.82	.41	.89	13.59
	High	G-DINA	.61	.94	1.44	.75	.96	14.65
		DINA	.65	.95	7.60	.76	.97	9.97
1000	Low	G-DINA	.07	.71	88.92	.10	.75	152.68
		DINA	.08	.72	75.02	.16	.78	43.92
	Medium	G-DINA	.28	.85	38.16	.40	.89	59.99
		DINA	.32	.86	26.60	.44	.90	20.29
	High	G-DINA	.65	.95	17.19	.77	.97	24.82
		DINA	.67	.95	12.31	.77	.97	14.44

The aforementioned impacts of item quality, sample size, and test length on classification accuracy still held true for results shown in Tables 2.5, 2.6 and 2.7 for the conditions for  $D = 2$ ,  $K(d) = 8$ ,  $D = 4$ ,  $K(d) = 5$ , and  $D = 4$ ,  $K(d) = 8$ , respectively. Note that, as the dimensionality increased (i.e., larger  $D$  and  $K(d)$ ), the CT also dramatically increased

due to increased amount of data. For example, it took, on average, 1.7 seconds to estimate data using the DINA model for the total number of attributes  $K = 10$ ,  $N = 1000$ ,  $J = 50$ , and high item quality, and 4.3 minutes for the corresponding conditions when  $K = 32$ .

Table 2.6:  $CVC_d$ , CAC, CT(sec.) for  $D = 4$ ,  $K(d) = 5$

$N$	Item	Model	$J = 60$			$J = 100$		
			$CVC_d$	CAC	CT(sec.)	$CVC_d$	CAC	CT(sec.)
500	Low	G-DINA	.15	.68	47.88	.20	.72	64.06
		DINA	.18	.70	33.98	.27	.76	33.96
	Medium	G-DINA	.35	.81	32.86	.51	.87	32.31
		DINA	.43	.83	18.41	.58	.89	18.88
	High	G-DINA	.66	.92	17.04	.83	.96	18.87
		DINA	.73	.93	11.45	.85	.97	13.20
1000	Low	G-DINA	.16	.69	101.37	.23	.74	135.62
		DINA	.21	.72	56.18	.32	.78	46.54
	Medium	G-DINA	.39	.83	55.27	.57	.89	51.92
		DINA	.46	.84	27.31	.60	.90	26.00
	High	G-DINA	.69	.93	26.44	.85	.97	29.11
		DINA	.74	.94	17.23	.85	.97	19.25

The relative efficiencies, namely,  $RE_{CVC_d}$ ,  $RE_{CAC}$  and  $RE_{CT}$  of AP compared with CPE for  $D = 2$  and  $K(d) = 5$  are given in Table 2.8. The results show that AP maintained comparable classification accuracy compared with CPE, as the  $RE_{CVC_d}$  and  $RE_{CAC}$  ranged from .75 to 1.25. Surprisingly, under many conditions where item quality was unfavorable, the classification accuracy of AP was even higher than CPE. For instance, the  $RE_{CVC_d}$  for  $N = 1000$ ,  $J = 30$ , low item quality was 1.25 when fitting the G-DINA model and 1.41 when fitting the DINA model. As item quality and test length increased, the classification results obtained by CPE became more and more accurate, and consequently, the  $RE_{CVC_d}$  dropped, especially when the DINA model was fitted to the data. For example, when fitting

the DINA model for data with high item quality and long test length, AP was .75 and .76 accurate compared with CPE, although AP still remained a comparably high CAC (i.e., .94 and .95, respectively). In contrast, using AP with the G-DINA model still resulted in close to 1  $RE_{CVC_d}$ s for the above conditions. However, although CACs were comparable, AP was shown to be substantially more computationally efficient than CPE. The  $RE_{CT}$  ranged from .02 to .09 indicating that the AP was approximately 10 to 50 times faster than CPE. In summary, AP maintained relatively high classification accuracy particularly at the individual attribute level, but was computationally much faster compared with CPE.

Table 2.7:  $CVC_d$ , CAC, CT(sec.) for  $D = 4$ ,  $K(d) = 8$

$N$	Item		$J = 96$			$J = 160$		
	Quality	Model	$CVC_d$	CAC	CT(sec.)	$CVC_d$	CAC	CT(sec.)
500	Low	G-DINA	.06	.70	295.40	.09	.74	609.42
		DINA	.06	.70	259.09	.14	.76	388.90
	Medium	G-DINA	.18	.81	252.51	.39	.89	280.76
		DINA	.21	.81	222.51	.46	.90	186.87
	High	G-DINA	.49	.91	170.32	.79	.97	177.29
		DINA	.57	.93	129.36	.80	.97	151.42
1000	Low	G-DINA	.06	.70	681.91	.11	.76	1306.19
		DINA	.06	.70	649.45	.17	.79	580.72
	Medium	G-DINA	.20	.81	535.51	.42	.90	474.20
		DINA	.25	.83	408.65	.48	.90	276.13
	High	G-DINA	.52	.92	314.45	.80	.97	280.73
		DINA	.59	.93	212.50	.81	.97	257.78

Table 2.8:  $RE_{CVC_d}$ ,  $RE_{CAC}$ , and  $RE_{CT}$  for  $D = 2$ ,  $K(d) = 5$ 

$N$	Item	Quality	Model	$J = 30$			$J = 50$		
				$RE_{CVC_d}$	$RE_{CAC}$	$RE_{CT}$	$RE_{CVC_d}$	$RE_{CAC}$	$RE_{CT}$
500	Low	G-DINA	1.06	1.01	.05	1.22	1.03	.04	
		DINA	1.18	1.03	.08	1.04	1.01	.02	
	Medium	G-DINA	1.13	1.02	.04	1.04	1.01	.04	
		DINA	1.15	1.04	.03	.83	.97	.04	
	High	G-DINA	1.06	1.01	.04	.94	.99	.08	
		DINA	1.00	1.00	.05	.76	.94	.08	
1000	Low	G-DINA	1.25	1.04	.03	1.24	1.04	.02	
		DINA	1.41	1.06	.03	1.00	.99	.02	
	Medium	G-DINA	1.20	1.04	.02	1.02	1.00	.03	
		DINA	1.07	1.01	.02	.82	.96	.04	
	High	G-DINA	1.04	1.01	.04	.93	.99	.06	
		DINA	1.00	.99	.04	.75	.95	.09	

### 2.3.2 Simulation Study 2

#### *Design*

In simulation study 2, the performance of AP for tracking student progress in longitudinal data were examined. Several factors were manipulated in the same manner as simulation study 1, namely, the number of attributes per domain, test length, sample size, item quality, and data generating model (G-DINA and DINA). The number of time points  $D$  was fixed to three, representing three sets of skills associated with three different difficulty levels.

Different from simulation study 1, three waves of response data and proficiency profiles were generated. To properly simulate the growth of student ability over time, the HO  $\theta$  were generated from  $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\mu} = (-1, 0, 1)$ , and the diagonal elements of  $\boldsymbol{\Sigma}$  were set to .25 and the off-diagonal elements were set to .225, which was equivalent to

$\rho = .9$  between abilities at two different time points. The slope parameter  $a_{k(d)}$  for the HO model was fixed to 1.5 and the difficulty parameters for the first, second and third domain were sampled from  $N(-1, .25)$ ,  $N(0, .25)$  and  $N(1, .25)$ , respectively. Specifying the parameters as stated above allowed for student abilities to grow over time, also allowing for previously mastered skills to be forgotten.

Q-matrices were created following the general rules discussed in simulation study 1, and the item responses were generated using the complete Q-matrix and attribute profiles. Again, one hundred random samples were generated for each condition.

AP was used to evaluate student performance at three time points: the beginning, the middle, and the end of an educational program. At each time point, the target domain was corresponded to the current level of instruction, and skills from other levels were collapsed. AP was then applied three times to different target domains at the three different time points. The  $CVC_d$ , CAC, and CT were provided to evaluate the performance of AP. Because the total number of attributes at each time point was 15, applying CPE at three time points is computationally difficult. Hence, in study 2 the CPE was not applied to the data.

### *Results*

Similar patterns were found in the longitudinal data as those in simulation study 1. For comparison purposes, only the results for  $K(d) = 5$  are presented in Table 2.9. The  $CVC_d$  and CAC ranged from .72 to .99 and .19 to .88, respectively. The results suggested satisfactory classification accuracy at each time point, when item quality was at least medium. Again, item quality and test length had a positive impact on classification accuracy. A

slight improvement was still observed when sample size was doubled.

The CT ranged from 3.9 to 34.1 seconds when  $K(d) = 5$  indicating a high computational efficiency. Similar to the findings in the previous study, fitting the G-DINA model to low item quality data resulted in longer CT than fitting the DINA model or to data with higher item quality. The maximum CT appeared when fitting the G-DINA model to data with low item quality and long test length, which nonetheless took only more than half a minute to finish.

Table 2.9:  $CVC_d$ , CAC, CT(sec.) for  $D = 3$ ,  $K(d) = 5$

$N$	Item	Model	$J = 45$			$J = 75$		
			$CVC_d$	CAC	CT(sec.)	$CVC_d$	CAC	CT(sec.)
500	Low	G-DINA	.19	.72	11.84	.23	.75	17.63
		DINA	.22	.73	10.80	.32	.78	8.65
	Medium	G-DINA	.44	.85	6.06	.57	.89	8.19
		DINA	.50	.86	5.43	.62	.90	5.23
	High	G-DINA	.76	.95	3.94	.87	.97	6.09
		DINA	.81	.96	3.88	.87	.97	4.93
1000	Low	G-DINA	.22	.73	20.46	.28	.77	34.14
		DINA	.25	.75	14.25	.34	.79	14.43
	Medium	G-DINA	.47	.86	9.56	.59	.90	14.50
		DINA	.52	.87	9.87	.63	.91	8.74
	High	G-DINA	.78	.95	7.54	.87	.97	10.48
		DINA	.81	.96	8.41	.88	.97	7.69

## 2.4 Real-Data Illustration

To demonstrate the practical feasibility of AP, the TIMSS 2007 fourth-grade mathematics dataset was analyzed. TIMSS is an international educational comparison program for mathematics and science performance, which contained core skills that are taught in



the fourth and eighth-grade classrooms (IEA; Mullis, Martin, & Foy, 2008). As briefly mentioned before, the three knowledge domains that TIMSS 2007 covered were number, geometric shapes and measures, and data display, which accounted for 40%, 40% and 20% of the assessment (Mullis et al., 2008). A total of 38 skills were identified by TIMSS under three content domains, which were re-specified and combined to a set of 15 attributes by Lee et al. (2011). A detailed list of the 15 attributes are provided in Table 2.10.

Lee et al. (2011) fitted the TIMSS 2007 data using the DINA model to provide possible diagnostic information that could have a direct impact to classroom teaching and learning. To be able to conduct this analysis, a Q-matrix, as shown in Table 2.11, was defined for the two fourth-grade mathematics blocks, M03 and M04, which appeared in booklets 4 and 5. A total of 15 multiple-choice and 10 constructed-response items were included in these two blocks. The Q-matrix was developed by five mathematics educators via a process where five Q-matrices were developed independently, and then discussed to form a final Q-matrix. All items in the Q-matrix are dichotomous, except two items: *M041275* and *M031247*, which were polytomous response items with a maximum score of 2. These two items were dichotomized such that a score of 1 is recorded only if the highest score was achieved, and 0 otherwise.

Nine participant countries or regional entities, namely, 1) England, 2) US, 3) Australia, 4) Ontario, Canada, 5) Alberta, Canada, 6) British Columbia, Canada, 7) New Zealand, 8) Massachusetts, US, and 9) Minnesota, US were chosen to form the analysis sample. The selected countries or regional entities were primarily English-speaking countries that participated in TIMSS 2007, and the majority of the students in the sample used English as their test language. Due to the matrix sampling technique that TIMSS 2007 utilized

Table 2.10: Attributes Identified for the TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5

Domain	Attribute
N	$\alpha_{1(1)}$ : representing, comparing and ordering whole number as well as demonstrating knowledge of place value
	$\alpha_{1(2)}$ : recognizing multiples, computing with whole numbers using the four operations, and estimating computations
	$\alpha_{1(3)}$ : solve problems, including those set in real life contexts
	$\alpha_{1(4)}$ : solve problems involving proportions
	$\alpha_{1(5)}$ : recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents
	$\alpha_{1(6)}$ : solve problems involving simple fractions and decimals including their addition and subtraction
	$\alpha_{1(7)}$ : find the missing number or operation and model simple situations involving unknowns in number sentence or expressions
	$\alpha_{1(8)}$ : describe relationships in patterns and their extensions
GM	$\alpha_{2(1)}$ : measure, estimate and understand properties of lines and angles and be able to draw them
	$\alpha_{2(2)}$ : classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties
	$\alpha_{2(3)}$ : calculate and estimate perimeters, area and volume
	$\alpha_{2(4)}$ : locate points in an informal coordinate to recognize and draw figures and their movement
DD	$\alpha_{3(1)}$ : read data from tables, pictographs, bar graphs, and pie charts
	$\alpha_{3(2)}$ : compare and understand how to use information from data
	$\alpha_{3(3)}$ : understand different representations and organizing data using tables, pictographs, and bar graphs

*Note.* N = number; GM = geometric shapes and measures; DD = data display

Table 2.11: Q-matrix for the TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5

Item	Domain: N								Domain: GM				Domain: DD		
	$\alpha_{1(1)}$	$\alpha_{1(2)}$	$\alpha_{1(3)}$	$\alpha_{1(4)}$	$\alpha_{1(5)}$	$\alpha_{1(6)}$	$\alpha_{1(7)}$	$\alpha_{1(8)}$	$\alpha_{2(1)}$	$\alpha_{2(2)}$	$\alpha_{2(3)}$	$\alpha_{2(4)}$	$\alpha_{3(1)}$	$\alpha_{3(2)}$	$\alpha_{3(3)}$
M041052	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M041056	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M041069	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
M041076	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
M041281	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
M041164	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
M041146	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
M041152	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
M041258A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M041258B	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
M041131	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0
M041275	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
M041186	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
M041336	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0
M031303	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031309	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031245	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
M031242A	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
M031242B	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
M031242C	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
M031247	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
M031219	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
M031173	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031085	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M031172	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1

Note. N = number; GM = geometric shapes and measures; DD = data display

(Olson, Martin, & Mullis, 2008), where students only responded to one of the test booklets, the selected booklets were only exposed to a portion of the entire sample for each country. There were a total of 14 test booklets in TIMSS 2007, and approximately 14% of examinees for each country responded to booklets 4 or 5. More detailed information about the analysis sample can be found in Table 2.12.

Table 2.12: Sample Information

Country	Sample Size	Sample Size of Booklets 4 & 5	% Test Language in English
ENG	4316	623	100%
USA	7896	1131	100%
AUS	4108	602	100%
COT	3496	510	66%
CAB	4037	588	96%
CBC	4153	582	98%
NZL	4940	691	100%
UMA	1747	255	100%
UMN	1846	263	100%
Overall	36539	5245	

*Note.* ENG = England; USA = United States; AUS = Australia; COT = Ontario, Canada; CAB = Alberta, Canada; CBC = British Columbia, Canada; NZL = New Zealand; UMA = Massachusetts, US; UMN = Minnesota, US

The DINA and G-DINA model were fitted to the data using both AP and CPE. To compare the classification results, the domain level vector-wise and attribute-wise agreement rate, denoted by  $VAR_d$  and AAR, between AP and CPE were calculated. To examine the proximity of the disagreement between AP and CPE, for each domain,  $VAR_d$  and AAR were also calculated allowing disagreement on one attribute. In addition, CT was also recorded to evaluate the computational efficiency, and  $RE_{CT}$ , was also computed. The

results of the real data are given in Table 2.13.

Table 2.13:  $VAR_d$ , AAR,  $RE_{CT}$  for TIMSS 2007 Fourth-Grade Mathematics Data

Model	$VAR_N$	$VAR_N(K \geq 7)$	$VAR_{GM}$	$VAR_{GM}(K \geq 3)$	$VAR_{DD}$	$VAR_{DD}(K \geq 2)$	AAR	$RE_{CT}$
DINA	.18	.23	.52	.65	.05	.75	.89	.02
G-DINA	.10	.17	.13	.33	.18	.27	.81	.01

*Note.*  $VAR_N$  = VAR for Number;  $VAR_N(K \geq 7)$  = VAR with 1 Disagreed Attribute for Number;  $VAR_{GM}$  = VAR for Geometric Shapes and Measurement;  $VAR_{GM}(K \geq 3)$  = VAR with 1 Disagreed Attribute for Geometric Shapes and Measurement;  $VAR_{DD}$  = VAR for Data Display;  $VAR_{DD}(K \geq 2)$  = VAR with 1 Disagreed Attribute for Data Display

As shown in Table 2.13, the AAR of AP and CPE for both fitting the DINA and G-DINA model was above .8. The  $VAR_d$ , in contrast, did not demonstrate uniformly high agreement across domains due to the different  $K(d)$ . For instance, the VARs for domain N were not ideal due to the large number of attributes involved, whereas the VARs for domain GM were much higher. Compared to the G-DINA model, the DINA model resulted in higher classification agreement between AP and CPE possibly because it is a simpler model. Finally, the results show that AP was highly efficient, and only took 1% - 2% of the CT CPE required to analyze the data.

## 2.5 Conclusions and Discussions

This study tackled the issue in utilizing CDMs to estimate diagnostic assessment data with high dimensionality. As such, it extended the use of CDMs in situations where diagnosing a large set of attributes from multiple content domains is of interest. The proposed approach, AP, utilizes the attribute structure and adjusts the grain size of nontarget attributes to form composite attributes to simplify the estimation issue. The performance of AP was examined in two simulation studies, the first of which demonstrated that AP obtained high attribute-level and domain-level classification accuracy when the item quality was at least

medium for cross-sectional data. The second simulation study showed the usefulness of AP in tracking student progress over time where the nontarget attributes did not provide much information. In the real data illustration, AP was used to fit the TIMSS 2007 fourth-grade mathematics data and showed high computational efficiency.

Throughout the simulation studies and real data illustration, AP demonstrated great feasibility in simplifying the problem of estimating one large model with high dimensionality into estimating several models with a smaller set of attributes. However, for this approach to be effective, a knowledge structure which allows partitioning attributes is necessary. In both simulation studies, the attributes within a domain were set to be more homogeneous than attributes from different domains. Thus, the collapsed composite attributes from a domain still represented the shared meaning of the domain. In contrast, ignoring the information in other domains may affect the attribute classification accuracy. To evaluate the impact, a small set of data were re-fitted using the test data split by each domain, and the resulting CACs were about 8% - 10% lower than those obtained using AP.

Another implication of this study is that the grain size of attributes does not have to remain static throughout different testing stages. As shown in the simulation study 2, when the target domain changed over time, it was admissible to combine attributes that were not informative any more, and the resulting classification accuracy still remained satisfactory.

Despite the promising results, this study had several limitations. First of all, when the item quality was low or test length was short, AP did not provide accurate vector-wise classifications of examinees. One possible research direction is to utilize other sources

of information to enhance the classification accuracy when the test is not sufficiently informative. When long test length is not practicable, one way to maintain a relatively high classification accuracy is to administer the test using computerized adaptive testing (CAT). However, few research addressed the issue in implementing CAT for high-dimensional cognitive diagnosis.

In addition, this study did not consider the situation where the attributes cannot be partitioned into mutually exclusive subsets. As an example, a skill in TIMSS 2007 can be associated with a content domain and a cognitive domain ((i.e., recall, recognize, compute, retrieve, measure, and classify/order; Mullis et al., 2008). Further research is warranted to accommodate this type of assessment structure. As another practical issue, this study did not provide practical solutions for modeling a even larger set of attributes (i.e.,  $K \gg 32$ ). One possibility is to further reduce the granularity of attributes, however the performance is remained to be examined.

Last, when a test involves a large number of attributes, the identifiability of all latent classes could become an issue. In the simulation studies, the Q-matrices were developed with the completeness property (Köhn & Chiu, 2017), so all latent classes were distinguishable. However, in the real-data illustration, the 25-item Q-matrix for the TIMSS 2007 was shown to be incomplete (Yamaguchi & Okada, 2018), which potentially could negatively influence the classification accuracy. Hence, systematic investigations of the impact of the incompleteness of the Q-matrix on attribute classification accuracy need to be undertaken.

## 2.6 References

- Akbay, L. (2017). *Identification, estimation, and Q-matrix validation of hierarchically structured attributes in cognitive diagnosis* (Doctoral dissertation). Retrieved from <https://doi.org/doi:10.7282/T3RR21JV>.
- Chiu, C.-Y., Köhn, H.-F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika*, *81*, 1069–1092.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.
- de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.
- de la Torre, J., & Minchen, N. D. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89–97.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.
- DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 970–1030). Amsterdam: North-Holland Publications.
- George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*, 1–24. doi: 10.18637/jss.v074.i02
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.
- Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). Retrieved from <http://hdl.handle.net/2142/87393>.
- Hartz, S. M., & Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series*, 2008.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive



- diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.
- Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment*, *15*, 1–7.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
- Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*, 112–132.
- Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.
- Ma, W., & de la Torre, J. (2016). GDINA: The generalized DINA model framework. *R package version 0.13.0*. Available online at: <http://CRAN.R-project.org/package=GDINA>.
- Mullis, I. V., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Olson, J., Martin, M., & Mullis, I. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*, 237–255.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.
- Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the timss 2007 fourth grade mathematics assessment. *PloS one*, *13*, e0188691.

## Chapter 3

# Improving Attribute Classification Accuracy in High Dimensional Data: A Four-Step Latent Regression Approach

### Abstract

Cognitive diagnosis modeling aims to provide detailed and actionable feedback on a set of finer-grained attributes. For feedback to be informative, the skills involved in a test must be . However, current computational constraints limit the attribute size to about 20. The accordion procedure (AP) has been proposed to handle much larger attribute sizes by focusing on one subset of attributes at a time, and creating nuisance attributes by collapsing the attributes of the remaining subsets. In this study, covariates are incorporated to supplement information obtained from AP. A four-step latent regression approach, which is both computationally manageable when high-dimensional data are involved and flexible when specifications at each step need to be adjusted, is proposed. A simulation study is conducted to examine the performance of the proposed approach. Results demonstrate that incorporating covariates can improve the AP correct classification rates particularly when the test alone is not sufficiently informative. A real-data example is also provided to demonstrate the feasibility of the proposed approach.

**Keywords:** cognitive diagnosis model, four-step latent regression, high-dimensional data

Cognitive diagnosis models (CDMs; for some examples, see de la Torre, 2011; Di-Bello, Roussos, & Stout, 2007; Hartz & Roussos, 2008; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008) have sparked interest among educational measurement researchers and practitioners because of its capability to provide formative information about the strengths and weaknesses of students on a set of fine grained skills or attributes. Theoretically, the number of attributes that can be estimated by a CDM is unlimited; however, in practice, this number may not exceed 15 due to a number of computational constraints (de la Torre, 2017). This limits the use of CDMs in scenarios where a comprehensive diagnosis in a knowledge domain is of interest, such as large-scale diagnostic assessments or retrofitting summative assessments using CDMs.

The accordion procedure (AP; de la Torre, 2017) has been proposed to address the high dimensionality estimation issue by focusing only on the attributes of one particular subset at a time, while the attributes of each of the remaining subsets are collapsed to create composite nuisance attributes. However, when the test is not sufficiently informative due to short test length or poor item quality, other sources of information might be needed to increase the classification accuracy. In this study, a four-step latent regression approach, which utilizes relevant covariates as ancillary information to make classification of CDMs more reliable, is proposed.

The rest of this paper is organized as follows. First, an introduction to CDMs and existing latent regression methods for incorporating covariates is given, followed by the proposed four-step approach. Simulation real-data studies of the Trends in International Mathematics and Science Study (TIMSS) 2007 are presented then. Finally, remarks and thoughts on possible future research are discussed to conclude this study.

### 3.1 CDMs

CDMs refer to a family of psychometric models that aim to classify examinees' skills (also called attributes) based on their item responses to tests designed for diagnostic purposes (for some examples, see de la Torre, 2011; DiBello et al., 2007; Hartz & Roussos, 2008; Henson et al., 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008). For the purposes of this paper, we used the following notations. Suppose we are interested in a test that contains  $J$  items measuring  $K$  attributes, and is responded to by  $N$  examinees. Let  $i$  index the examinees, where  $i = 1, \dots, N$ ;  $j$  the items, where  $j = 1, \dots, J$ ; and  $k$  the attributes, where  $k = 1, \dots, K$ . One of the major goals of CDMs is to classify each examinee into one of the  $2^K$  latent classes. Let  $l$  denote the latent class with the attribute vector  $\alpha_l = \{\alpha_{lk}\}$ , where  $l = 1, \dots, 2^K$ . The  $k$ th entry is  $\alpha_{lk} = 1$  if attribute  $k$  is mastered, and is  $\alpha_{lk} = 0$  otherwise. The item response matrix is denoted by  $\mathbf{Y} = \{Y_{ij}\}$ , where  $Y_{ij} = 1$  indicates a correct response; otherwise,  $Y_{ij} = 0$ .

A Q-matrix which is a  $J \times K$  matrix specifies the item-attribute association (Tatsuoka, 1983). The  $j$ th row vector of a Q-matrix, denoted by  $\mathbf{q}_j = \{q_{jk}\}$ , is called the  $q$ -vector for item  $j$ . When solving item  $j$  requires attribute  $k$ ,  $q_{jk} = 1$ , otherwise,  $q_{jk} = 0$ .

The generalized-deterministic inputs, noisy "and" gate (G-DINA; de la Torre, 2011) model is one of the general CDMs (see also, Henson et al., 2009; von Davier, 2008) that subsume a wide class of specific models, such as the DINA model (Haertel, 1989; Junker & Sijtsma, 2001), the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006), and the additive-CDM (A-CDM; de la Torre, 2011). Let  $K_j$  be the total number of required attributes for item  $j$ , which allows the  $2^K$  attribute profiles be collapsed into

$2^{K_j}$  reduced attribute profiles or latent groups. For instance, if  $\mathbf{q}_j = (1, 1, 0)$ , it can only distinguish four collapsed latent groups with profiles of:  $(0, 0, -)$ ,  $(1, 0, -)$ ,  $(0, 1, -)$  and  $(1, 1, -)$ . In this example, latent classes that are distinct only at the third attribute cannot be further distinguished, and thus can be combined. Let  $\boldsymbol{\alpha}_{lj} = (\alpha_{l1}, \dots, \alpha_{lK_j})$  denote the reduced attribute profile of the collapsed latent group for item  $j$ , where  $lj = 1, \dots, 2^{K_j}$ . The G-DINA model in the identity link can be written as

$$P(Y_{ij} = 1 | \boldsymbol{\alpha}_{lj}) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (3.1)$$

where  $\delta_{j0}$  is the intercept for item  $j$ ,  $\delta_{jk}$  is the  $k$ th attribute's main effect,  $\delta_{jkk'}$  is the two-way interaction between attribute  $k$  and  $k'$ , and  $\delta_{j12\dots K_j}$  is the  $K_j$ -way interaction.

With different link functions, the G-DINA model is equivalent to other general CDMs. For example, if the logistic link is used, the G-DINA model is equivalent to the LCDM (Henson et al., 2009). Moreover, the main effect G-DINA model with the logistic link is identical to the GDM. By setting proper constraints to the item parameters in the G-DINA model, several specific CDMs can be reparameterized using the G-DINA model.

Item parameters, either parameterized as  $P(Y_{ij} = 1 | \boldsymbol{\alpha}_{lj})$  or  $\boldsymbol{\delta}$ , can be estimated using marginal maximum likelihood estimation (MMLE) algorithms with the closed-form solution (de la Torre, 2011). After the item parameter estimates are obtained, inferences on each examinee's attribute profile can be drawn from individual posterior distribution. Specifically, the expected a posteriori (EAP) and maximum a posteriori (MAP) are two commonly used methods (Huebner & Wang, 2011). The EAP finds the expected value of a posterior distribution, or equivalently, the marginal posterior probability of mastering an

attribute. For example, for attribute  $k$ , its EAP is given by

$$\tilde{\alpha}_{ik}^{(EAP)} = P(\alpha_{ik} = 1 | \mathbf{Y}_i) = \sum_{l=1}^{2^K} P(\alpha_l | \mathbf{Y}_i) I(\alpha_{lk} = 1), \quad (3.2)$$

where  $I(\alpha_{lk} = 1)$  is an indicator function that equals to 1 when  $\alpha_{lk}$  is 1, and 0 otherwise. To obtain a binary estimate of  $\alpha_k$ ,  $\tilde{\alpha}_{ik}^{(EAP)}$  is usually dichotomized at .5, that is,  $\hat{\alpha}_{ik}^{(EAP)} = 1$ , if  $\tilde{\alpha}_{ik}^{(EAP)} \geq .5$ ; otherwise,  $\hat{\alpha}_{ik}^{(EAP)} = 0$ . The MAP, on the other hand, finds the maximum or the mode of the posterior distribution  $P(\alpha_l | \mathbf{Y}_i)$ , and it can be written as

$$\hat{\alpha}_i^{(MAP)} = \arg \max_{\alpha_l} \{P(\alpha_l | \mathbf{Y}_i)\}. \quad (3.3)$$

### 3.1.1 AP

Regardless of which model is used to classify examinees, a common problem persists: the curse of dimensionality (Bellman, 1957). That is, the more attributes a test involves, the more difficult, if not impossible, to classify examinees accurately because the number of latent classes grows exponentially with the number of attributes. AP attempts to solve the estimation issue in high-dimensional data by estimating the entire knowledge space by parts. Specifically, when focusing on one domain (i.e., the target domain), attributes from nontarget domains can be combined or collapsed as domain-level attributes. Because the granularity of the collapsed attributes is coarser than the size of interest, the collapsed attributes only serve as a statistical control for the possible effects from nontarget domains when fitting CDMs, hence, are called nuisance attributes. AP can be either applied once to the target domain or multiple times to obtain the complete attribute profile depending

on the research purposes.

Suppose there are  $D$  domains involved in a test, and each domain consists of  $K(d)$  skills, where  $d = 1, \dots, D$ . Then the total number of attributes is  $K = \sum_d K(d)$ . The attribute profile for domain  $d$  is denoted by  $\alpha_{l(d)} = (\alpha_{l(d)1}, \dots, \alpha_{l(d)K(d)})$ , where  $l(d) = 1, \dots, 2^{K(d)}$ . To apply AP to a target domain, the original Q-matrix needs to be modified correspondingly by collapsing attributes in nontarget domains to form a new Q-matrix for AP analyses. One possible collapsing rule is the disjunctive condensation rule that the collapsed  $q$ -entry  $q_{j(d')}$  is equal to 1, if at least one attribute in domain  $d'$  is required by item  $j$ ; otherwise,  $q_{j(d')} = 0$ . This rule can be expressed as

$$q_{j(d')} = \begin{cases} 1 & \text{if } \sum_{k(d')=1}^{K(d')} q_{jk(d')} \geq 1, \\ 0 & \text{otherwise,} \end{cases} \quad (3.4)$$

where  $q_{j(d')}$  is the collapsed attribute for nontarget domain  $d'$ , and  $q_{jk(d')}$  is the original  $q$ -entry of the  $k$ th entry in domain  $d'$ . Results demonstrated that the classification of AP was equally accurate compared with the complete-profile estimation, but the computation time of AP was much shorter (de la Torre, 2017).

### 3.2 Incorporating Covariates in CDMs

Ancillary information or covariates can be incorporated to improve the precision, not only of the person parameter estimates but also of item parameter estimates for latent variable models (see; Ackerman & Davey, 1991; Ayers, Rabe-Hesketh, & Nugent, 2013; de la Torre & Song, 2009; Kahraman & Kamata, 2004; Mislevy, 1987; Mislevy & Sheehan,

1989; Vermunt, 2010). These variables may include examinees' performance on the overall test or subtests, the structures of the underlying abilities, and examinees' background variables such as demographic information or educational standings (de la Torre & Song, 2009).

Covariates of examinees are readily available in several educational testing programs. For example, the TIMSS and the Program for International Student Assessment (PISA) both develop questionnaires for students, teachers and school principals that contain numerous background variables. Despite the availability of covariates in testing data, the techniques to incorporate covariates to classify examinees' proficiency classes in CDMs have not been extensively explored. Nonetheless, several previous works that have investigated incorporating covariates into CDMs are worth discussing.

Ayers et al. (2013) incorporated observed covariates in the DINA model through a one-step latent regression approach. Suppose  $\mathbf{Z} = \{\mathbf{Z}_i\}$  be the matrix of covariates, and  $\mathbf{Z}_i$  is the covariate vector of examinee  $i$ . Then,  $P(\alpha_{ik}|\mathbf{Z}_i)$  is modeled as a logistic function which can be expressed as

$$\text{logit}[P(\alpha_{ik}|\mathbf{Z}_i)] = \beta_0 + \mathbf{Z}_i' \boldsymbol{\beta}_k + \xi_i, \quad (3.5)$$

where  $\beta_0$  and  $\beta_k$  are the coefficients, and  $\xi_i$  is the random intercept associated with examinee  $i$ . Assuming local independence, the conditional probability of examinee  $i$  in latent class  $l$  given  $\mathbf{Z}_i$  can be obtained by

$$\pi(\alpha_l|\mathbf{Z}_i) = \prod_{k=1}^K P(\alpha_{lk}|\mathbf{Z}_i)^{\alpha_{lk}} [1 - P(\alpha_{lk}|\mathbf{Z}_i)]^{1-\alpha_{lk}}. \quad (3.6)$$



Park and Lee (2014) extended this approach such that, in addition to modeling the effects of covariates on individual attributes, the effects of covariates on responses can also be modeled as

$$\text{logit}[P(Y_{ij}|\boldsymbol{\alpha}_i, \mathbf{Z}_i)] = f_j + d_j\eta_{ij} + \boldsymbol{\gamma}'_j\mathbf{Z}_i, \quad (3.7)$$

where  $f_j$  and  $d_j$  are the reparameterized guessing and slip parameters, and  $\boldsymbol{\gamma}_j$  are the effects of covariates.

Because the one-step approach involves simultaneous estimation of the CDM and the latent regression coefficients, it produces unbiased estimates of both item parameters and latent regression coefficients. However, Vermunt (2010) argued that the one-step approach in the context of latent class analysis (LCA) has several disadvantages, and a number of which apply to the CDM context as well. First, sometimes the one-step approach is impractical because any modifications to CDMs or latent regressions require refitting the entire model, especially in the case of exploratory researches where adding or dropping covariates is often needed. Second, the one-step approach does not fit with the purposes of secondary analyses where the major interest is to study the relationship between latent classes and covariates. Third, the one-step approach forces classifying examinees and regressing latent classes on covariates at the same stage of a study, which may not always be feasible — the CDM analyses and latent regression analyses may be conducted in separate stages. Thus, a multistep approach to incorporate covariates, especially for conducting secondary analyses (Iaconangelo, 2017), is still needed.

Iaconangelo (2017) extended the three-step approach (see; Bolck, Croon, & Hage-naars, 2004; Vermunt, 2010) and proposed either using logistic model or multinomial logistic model to regress individual attributes or latent classes on the covariates, respectively.

To take a closer look, in step 1, CDMs are fitted to the data; in step 2, examinees are assigned to latent classes; in step 3, either attribute vector estimate  $\hat{\alpha}_l$  or individual attribute estimate  $\hat{\alpha}_k$  is regressed on the covariates with the necessary correction weights. To rectify the potential bias due to classification errors, sample-level and posterior-distribution level correction weights are derived from the classification error probabilities (see below), and used in the third step when fitting the regression models.

The three-step approach offers a great deal of flexibility in customizing the CDM, as well as the latent regression independently. For example, regular CDM analyses such as Q-matrix validation, model selection, or model-data fit can be conducted, and modifications to CDMs can be made to avoid the potential impact of the Q-matrix or model misspecification on the classification results. Also, in step 3, adding or dropping covariates in the regression analyses can be easily done without redoing the previous steps (Iaconangelo, 2017; Vermunt, 2010). Last but not least, common variable selection techniques, such as the stepwise selection, least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) or principal component regression (Jolliffe, 1982), can be easily applied when there are a large number of available covariates.

### **3.3 Incorporating Covariates for the Accordion Procedure**

As introduced previously, covariates can be incorporated into CDMs through either a one-step or three-step latent regression approach (Dayton & Macready, 1988; Iaconangelo, 2017; Park & Lee, 2014). In the context of this work which involves diagnosing a large set of attributes, the three-step approach is deemed more suitable because AP needs to be applied in the first step to solve the estimation issue. The latent class assignments can then

be obtained in the second step, followed by the regression analyses. However, the three-step approach was not designed to update the classification results utilizing covariates as ancillary information, but only to obtain unbiased coefficient estimates. Hence, the three-step approach is modified to be applicable in AP, and extended to a four-step approach, where, in the fourth step, the information obtained from CDMs and latent regressions can be combined to better estimate examinees' latent class memberships.

### 3.3.1 Classification Error Probabilities and Correction Weights

Classification error probabilities (CEP) quantify the amount of error in the latent class classification, and can be evaluated at either the attribute-vector or individual attribute level (Iaconangelo, 2017). In this study, the approach of regressing individual attribute to covariates is of greater interest, because when the number of attributes is large, regressing latent classes on the covariates can be intractable, whereas regressing individual attributes on the covariates remains manageable. To account for the potential bias due to regressing the incorrect classification results on covariates, correction weights need to be extracted from the matrix of attribute-level classification error probabilities (ACEPs; Iaconangelo, 2017; Vermunt, 2010). The ACEPs for attribute  $k$ , which is a  $2 \times 2$  matrix, can be computed at either the sample-level ( $SL_k$ ) or individual posterior-distribution level ( $PDL_{ik}$ ).  $SL_k$  differs from  $PDL_{ik}$  in that the former is computed by averaging the ACEPs across the whole sample, and is the same for each examinee, whereas the latter is computed using the individual marginal posterior probability, hence is unique for each examinee. That is, for attribute  $k$ , the same correction weights are assigned to examinees with the same mastery status when using  $SL_k$ , whereas the correction weights are unique to each examinee's

marginal posterior probability when using  $PDL_{ik}$ .

To compute the  $SL_k$  and  $PDL_{ik}$ , the marginal posterior probability of mastering each attribute needs to be computed by aggregating the posterior probability. In the context of AP, the marginal posterior probabilities of attributes only from the target domain are computed using Equation 3.2. Define  $\mathbf{P}_{ik(d)}$  as the vector of marginal posterior probabilities of mastering and nonmastering attribute  $k(d)$  for examinee  $i$ , which is given by

$$\mathbf{P}_{ik(d)} = (P(\alpha_{k(d)} = 0 | \mathbf{Y}_i), P(\alpha_{k(d)} = 1 | \mathbf{Y}_i)). \quad (3.8)$$

Define  $\mathbf{A}_{ik(d)}$  as the vector of attribute assignments, which is written by

$$\mathbf{A}_{ik(d)} = (I(\hat{\alpha}_{k(d)} = 0), I(\hat{\alpha}_{k(d)} = 1)). \quad (3.9)$$

When  $\hat{\alpha}_{k(d)} = 1$ ,  $\mathbf{A}_{ik(d)} = [0 \ 1]$ ; otherwise,  $\mathbf{A}_{ik(d)} = [1 \ 0]$ . In the context of AP,  $SL_{k(d)}$  and  $PDL_{ik(d)}$  are calculated using the elements in  $\mathbf{P}_{ik(d)}$  and  $\mathbf{A}_{ik(d)}$ . The  $SL_{k(d)}$  is given by

$$SL_{k(d)} = \begin{bmatrix} \frac{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=0)}{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i)} & \frac{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=1)}{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i)} \\ \frac{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=0)}{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i)} & \frac{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=1)}{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i)} \end{bmatrix}, \quad (3.10)$$

and the  $PDL_{ik(d)}$  as

$$PDL_{ik(d)} = \begin{bmatrix} \frac{NP(\alpha_{k(d)}=0 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=0)}{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i)} & \frac{NP(\alpha_{k(d)}=0 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=1)}{\sum_{i=1}^N P(\alpha_{k(d)}=0 | \mathbf{Y}_i)} \\ \frac{NP(\alpha_{k(d)}=1 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=0)}{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i)} & \frac{NP(\alpha_{k(d)}=1 | \mathbf{Y}_i) I(\hat{\alpha}_{k(d)}=1)}{\sum_{i=1}^N P(\alpha_{k(d)}=1 | \mathbf{Y}_i)} \end{bmatrix}. \quad (3.11)$$

Below is an example of  $SL_{k(d)}$

$$SL_{k(d)} = \begin{bmatrix} .78 & .22 \\ .23 & .77 \end{bmatrix}.$$

Each row represents a true proficiency of  $\alpha_{k(d)}$  and each column represents the attribute assignment  $\hat{\alpha}_{k(d)}$ . In this example, if the true  $\alpha_{k(d)} = 0$ , 78% of examinees with the true proficiency of nonmaster will be classified as nonmaster, and 22% of these examinees will be classified as master. Similarly, if the true  $\alpha_{k(d)} = 1$ , 23% of the examinee whose true proficiency is master, will be classified as nonmaster and 77% as master. Ideally, when a test is informative, the diagonal elements will approach to 1 indicating a high proportion of examinees are classified correctly.

The vector of correction weights for attribute  $k(d)$ , denoted by  $\mathbf{w}_{ik(d)} = [w_{ik(d)0} \ w_{ik(d)1}]$ , can then be extracted as the  $k$ th column vector of the matrix of either  $SL_{k(d)}$  or  $PDL_{ik(d)}$ . Using the above example again, if an examinee is classified as nonmaster of attribute  $k(d)$ , the first column,  $\mathbf{w}_{ik(d)} = [.78 \ .23]$ , is used as the correction weights to reduce the bias in the latent regression.

### 3.3.2 Four-Step Approach

Because the three-step approach does not update the latent class assignments after the regression analyses, the classification of the three-step approach only relies on the response data through fitting a CDM. However, when covariates are related to latent classes, they provide ancillary information that can be used to augment information obtained from

CDMs. Therefore, an additional step is proposed to combine information obtained from latent regressions and CDMs to update the classification results. This modified approach is called the four-step approach.

The four-step approach in AP can be described as follows. In the first step, CDM is fitted to the data focusing on the domain  $d$  using AP. The marginal posterior probabilities of mastering attributes in domain  $d$  are obtained.

In the second step, EAP or MAP is used to estimate attribute profile for each examinee. In addition, the correction weights are obtained. The third step fits a logistic regression for each attribute. For example, for attribute  $\alpha_{k(d)}$ , the logistic regression model can be expressed as

$$P(\alpha_{k(d)} | \mathbf{Z}_i) = \frac{\exp(\beta_{k(d)0} + \mathbf{Z}_i' \boldsymbol{\beta}_{k(d)})}{1 + \exp(\beta_{k(d)0} + \mathbf{Z}_i' \boldsymbol{\beta}_{k(d)})}, \quad (3.12)$$

where  $\beta_{k(d)0}$  and  $\boldsymbol{\beta}_{k(d)}$  are coefficients associated with attribute  $k(d)$ . To incorporate the correction weights, the log likelihood can be written as

$$l_{k(d)} = \sum_{i=1}^N \log \sum_{\alpha_{k(d)}=0}^1 P(\alpha_{k(d)} | \mathbf{Z}_i) w_{ik(d)} \alpha_{k(d)}, \quad (3.13)$$

where  $w_{ik\alpha_{k(d)}}$  is the corresponding element in  $\mathbf{w}_{ik(d)}$ . As discussed earlier, the correction weight vector  $\mathbf{w}_{ik(d)}$  can be estimated by either  $SL_{k(d)}$  or  $PDL_{ik(d)}$ . Note that when substituting  $\mathbf{A}_{ik(d)}$  for  $w_{ik\alpha_{k(d)}}$  in Equation 3.13, the third step simplifies to an uncorrected approach where the potential classification error is ignored (Iaconangelo, 2017; Vermunt, 2010).

Quasi-Newton optimization methods, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) and the limited-memory BFGS with boundaries (L-BFGS-B), can be utilized to

maximize the log likelihood given in Equation 3.13. The L-BFGS-B is a popular optimizer, developed based on the BFGS algorithm, which is computationally efficient and allows users to set boundaries for estimated parameters. The L-BFGS-B algorithm is available in the R base function “optim” (R Core Team, 2013).

In the last step, Bayes’ theorem can be applied to combine information obtained from the CDM in the first step,  $P(\alpha_k|\mathbf{Y}_i)$ , and information from the logistic regression,  $L(\mathbf{Z}_i|\alpha_k)$ . This is given by

$$P(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{L(\mathbf{Z}_i|\alpha_{k(d)})P(\alpha_{k(d)}|\mathbf{Y}_i)}{\sum_{\alpha_{k(d)}=0}^1 L(\mathbf{Z}_i|\alpha_{k(d)})P(\alpha_{k(d)}|\mathbf{Y}_i)}, \quad (3.14)$$

where  $P(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i)$  is the the updated posterior probability of mastering attribute  $k(d)$  for examinee  $i$ , given the item response vector  $\mathbf{Y}_i$  and the covariate vector  $\mathbf{Z}_i$ . Based on the updated posterior probability, the classification of attribute  $k(d)$  can be updated using EAP or MAP.

## 3.4 Simulation Study

### 3.4.1 Design

A simulation study was conducted to evaluate the extent to which incorporating covariates can improve classification accuracy by comparing the four-step approach to other methods. Specifically, five methods were considered in this study, namely, 1) regressing the true  $\alpha_{k(d)}$  on the covariates in the third step, denoted by “True”; 2) regressing  $\hat{\alpha}_{k(d)}$  on the

covariates with *PDL* correction weights in the third step, denoted by “PDL”; 3) regressing  $\hat{\alpha}_{k(d)}$  on the covariates with *SL* correction weights in the third step, denoted by SL; 4) regressing  $\hat{\alpha}_{k(d)}$  on the covariates without correction weights, denoted by UC; 5) classifying examinees based on AP without incorporating covariates, denoted by AP. Note that in real data, true  $\alpha_{k(d)}$  will always be unknown, hence the results of “True” only serve as the “gold standard” - the best results the four-step approach can possibly achieve.

Six factors were considered in the simulation study: the number of domains ( $D = 2$  and 4), number of attributes per domain ( $K(d) = 5$  and 8), test length ( $2 \times$  and  $4 \times K$ ), sample size ( $N = 500$  and 2000), item quality (low, medium, and high) and association between  $\alpha_{k(d)}$  and covariates (strong and weak). The total number of conditions was  $2 \times 2 \times 2 \times 2 \times 3 \times 2 = 96$ .

Multiplying  $D$  by  $K(d)$ ,  $K = 10, 16, 20$ , and 32 number of attributes were considered in the study. Because of the wide range of  $K$ , fixing the test lengths across all possible combinations of  $D$  and  $K(d)$  could be problematic. For example,  $J = 60$  may be sufficiently long for conditions with  $D = 2$  and  $K(d) = 5$ , but not necessarily for conditions with  $D = 4$  and  $K(d) = 8$ . Hence, the test lengths were designed to depend on  $D$  and  $K(d)$ . Specifically, it was defined as  $J = 2 \times D \times K(d)$  and  $J = 4 \times D \times K(d)$ . For example, for  $D = 2$  and  $K(d) = 5$ , the two test length conditions are 20 and 40 items.

The item quality was manipulated by setting the lowest and highest probabilities of success, denoted by  $P_0$  and  $P_1$  for each item. Specifically, when item quality was low,  $P_0 \sim U(.25, .35)$  and  $P_1 \sim U(.65, .75)$ ; when item quality was medium,  $P_0 \sim U(.15, .25)$  and  $P_1 \sim U(.75, .85)$ ; and when item quality was high,  $P_0 \sim U(.05, .15)$  and  $P_1 \sim U(.85, .95)$ .

The higher-order (HO; de la Torre & Douglas, 2004) model was utilized to manipulate



the association between attributes and covariates,  $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$ , and to generate attribute profiles. In this study, three covariates were used throughout the simulation study. The HO  $\boldsymbol{\theta}$  and covariates  $\mathbf{Z}$  were generated from the multivariate normal distribution (MVN):

$$(\boldsymbol{\theta}, \mathbf{Z}) = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3) \sim \mathcal{N}_{D+3}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (3.15)$$

The  $\boldsymbol{\Sigma}$  were manipulated with respect to the association between attributes and covariates. For example, when  $D = 2$  and the association is strong, the  $\boldsymbol{\Sigma}$  is set to

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & .40 & .90 & .50 & .20 \\ .40 & 1.0 & .30 & .60 & .80 \\ .90 & .30 & 1.0 & .25 & .25 \\ .50 & .60 & .25 & 1.0 & .25 \\ .20 & .80 & .25 & .25 & 1.0 \end{bmatrix},$$

in which the top-left  $2 \times 2$  block-diagonal matrix is the variance-covariance matrix of  $\boldsymbol{\theta}$ , the bottom right  $3 \times 3$  block-diagonal matrix is the variance-covariance matrix of  $\mathbf{Z}$ , and the off-diagonal  $3 \times 2$  matrix indicates the covariance between  $\boldsymbol{\theta}$  and  $\mathbf{Z}$ . As shown in the matrix, the correlation between HO  $\boldsymbol{\theta}$ s is set to .4, the correlations between  $\mathbf{Z}$ s are set to .25 to avoid the multicollinearity, and the correlations between  $\boldsymbol{\theta}$ s and  $\mathbf{Z}$ s vary from .2 to .9 assuming. In contrast, when the association is weak, the  $\boldsymbol{\Sigma}$  is set to

$$\Sigma = \begin{bmatrix} 1.0 & .40 & .20 & .20 & .20 \\ .40 & 1.0 & .20 & .20 & .20 \\ \hline .20 & .20 & 1.0 & .20 & .20 \\ .20 & .20 & .20 & 1.0 & .20 \\ .20 & .20 & .20 & .20 & 1.0 \end{bmatrix},$$

in which all correlations, except those between  $\theta$ s, are set to .20.

The  $\Sigma$ s for other conditions were specified similarly and modified to maintain the positive definiteness of  $\Sigma$ . By setting  $\Sigma$ s as described above, the McFadden's pseudo  $R^2$  of the logistic regressions of true attributes on covariates is around .45 when the association was strong, and around .05 when the association was weak. The attribute profiles were then generated using the HO model, defined in Equation 3.16, where  $a_{k(d)} = 3.5$  and  $b_{k(d)} \sim N(0, .5)$ .

$$\text{logit}[P(\alpha_{ik(d)} | \theta_{id})] = a_{k(d)}(\theta_{id} - b_{k(d)}) \quad (3.16)$$

The slope parameter of the HO model was set to 3.5 for the association specified in  $\Sigma$  between HO  $\theta$  and  $Z$  to have a considerably large impact on the association between  $\alpha_{k(d)}$  and  $Z$ .

Eight Q-matrices were created according to  $D$ ,  $K(d)$  and the test length. The general rules for creating Q-matrices are summarized as follows. First, each Q-matrix contained at least one identity matrix to guarantee the property of Q-completeness as defined by Köhn and Chiu (2017). Second, each attribute was measured by the same number of items. Third, there were equal number of items measuring different domains. And fourth, the Q-matrices for  $J = 4 \times D \times K(d)$  were obtained by doubling the Q-matrices for  $J =$

$2 \times D \times K(d)$ . As an example, the Q-matrix for  $D = 2$ ,  $K(d) = 5$  and  $J = 20$  is provided in Table 3.1. Item responses were generated based on the complete attribute profile using the G-DINA model. One hundred replications for each condition were generated.

Table 3.1: Q-matrix for  $D = 2$ ,  $K(d) = 5$ ,  $J = 20$

item	$\alpha_{1(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{4(1)}$	$\alpha_{5(1)}$	$\alpha_{1(2)}$	$\alpha_{2(2)}$	$\alpha_{3(2)}$	$\alpha_{4(2)}$	$\alpha_{5(2)}$
1	1	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	1
11	1	1	0	0	0	1	0	0	0	0
12	0	1	1	0	0	0	1	0	0	0
13	0	0	1	1	0	0	0	1	0	0
14	0	0	0	1	1	0	0	0	1	0
15	1	0	0	0	1	0	0	0	0	1
16	1	0	0	0	0	0	0	0	1	1
17	0	1	0	0	0	0	0	1	1	0
18	0	0	1	0	0	0	1	1	0	0
19	0	0	0	1	0	1	1	0	0	0
20	0	0	0	0	1	1	0	0	0	1

The correct attribute-wise classification accuracy (CAC) and the domain-level vector-wise classification accuracy ( $CVC_d$ ) were calculated to compare the agreement of classification results with the true attribute elements and profiles. The CAC is calculated as

$$CAC = \sum_{i=1}^N \sum_{d=1}^D \sum_{k(d)=1}^{K(d)} \frac{I[\alpha_{ik(d)} = \hat{\alpha}_{ik(d)}]}{NDK(d)}, \quad (3.17)$$

and the  $CVC_d$  as

$$CVC_d = \sum_{i=1}^N \sum_{d=1}^D \frac{I[\alpha_{i(d)} = \hat{\alpha}_{i(d)}]}{ND}. \quad (3.18)$$

In addition, to evaluate the information added by the covariates,  $P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i)$  is defined as

$$P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i) = \max[P(\alpha_{k(d)} = 0|\mathbf{Y}_i, \mathbf{Z}_i), P(\alpha_{k(d)} = 1|\mathbf{Y}_i, \mathbf{Z}_i)]. \quad (3.19)$$

$P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i) \in (.5, 1)$  can be used to represent the classification certainty, in that, when the classification of  $\alpha_{k(d)}$  is more certain,  $P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i)$  is closer to 1. For AP, which does not incorporate covariates,  $P^*(\alpha_{k(d)}|\mathbf{Y}_i)$  was calculated.

### 3.4.2 Results

The results for  $D = 2$  were similar to those for  $D = 4$ . In addition, the CAC performed similarly across the different methods. Hence, only  $\text{CVC}_d$  for  $D = 4$  are presented in Tables 3.2 and 3.3 for  $K(d) = 5$  and 8, respectively.

To begin with, as expected, the test length had a positive impact on classification accuracy. As shown in Table 3.2, when  $D = 4$ ,  $K(d) = 5$ , and  $N = 500$ ,  $\text{CVC}_d$  ranged from .19 to .68 when  $J = 40$ , and .37 to .90 when  $J = 80$ . When the association was strong, for  $N = 2000$ , the  $\text{CVC}_d$  ranged from .26 to .75 when  $J = 40$ , and .49 to .90 when  $J = 80$ . Comparing  $\text{CVC}_d$  for different sample sizes, the results for  $N = 500$  showed lower classification accuracy compared with those for  $N = 2000$ , especially when the test length was short or item quality was low. For example, for  $J = 40$  and low item quality conditions, the differences of  $\text{CVC}_d$  between  $N = 500$  and 2000 ranged from .07 to .17, whereas for conditions with doubled test length or high item quality, the differences decreased to less than .01. The positive impact of the test length and sample size on  $\text{CVC}_d$  remained true for  $K(d) = 8$ , although the improvements in  $\text{CVC}_d$  from  $N = 500$  to  $N = 2000$  for  $J = 64$

and low item quality were small.

When the association between the attributes and the covariates was strong, comparing the four-step methods (i.e., True, PDL, SL and UC) with AP showed .01 to .11 improvement in  $CVC_d$  for  $K(d) = 5$  and  $N = 500$ . When the sample size increased to  $N = 2000$ , the improvement ranged from .03 to .13, where the largest improvement was observed when  $J = 40$  and the item quality was low. It can be noted that when the association was weak, incorporating irrelevant covariates still performed equally well or slightly better than AP (i.e., 0 to .03 improvement in  $CVC_d$ ). A close inspection of the results show that the improvement to  $CVC_d$  of incorporating covariates depended on how much information the test by itself can provide. That is, if the test was sufficiently informative, incorporating covariates did not provide much on improving the classification accuracy. For instance, when  $J = 80$  and item quality was high, the increase on  $CVC_d$  was at most .01. Similar patterns can be found for  $K(d) = 8$ , when the association was strong, the improvements in  $CVC_d$  if the four-step methods were used ranged from 0 to .10, whereas when the association was weak, the improvements were only 0 to .02.

The comparison among the four-step methods demonstrated True performed the best, as expected, whereas UC performed the worst. The discrepancies among the four methods were reduced as the test became more informative. For example, the  $CVC_d$  of the four methods for  $J = 80$ , high item quality, and strong association conditions were almost identical (i.e., .89 to .90). In contrast, the  $CVC_d$  of True were .08 to .10 higher than the other three methods in the  $J = 40$ , low item quality, and strong association condition. It can also be observed that, the performance of PDL was similar to that of SL. Last, the same pattern of results can be observed by  $K(d) = 5$  and 8. In summary, when the test

was not sufficiently informative (i.e., short test lengths or poor item qualities), covariates can provide ancillary information that made classifications more accurate, whereas when the test was already informative, the benefit of incorporating covariates was negligible.

Table 3.2:  $CVC_d$  for  $D = 4, K(d) = 5$

Item		$N = 500$						$N = 2000$				
$J$	Quality	Assoc.	True	PDL	SL	UC	AP	True	PDL	SL	UC	AP
40	Low	Weak	.20	.19	.19	.19	.19	.28	.27	.27	.26	.26
		Strong	.22	.20	.20	.20	.19	.39	.31	.31	.29	.26
	Medium	Weak	.42	.41	.41	.41	.41	.50	.49	.49	.49	.48
		Strong	.46	.43	.43	.43	.40	.56	.53	.52	.52	.47
	High	Weak	.68	.68	.68	.68	.67	.73	.73	.73	.73	.72
		Strong	.70	.69	.69	.69	.67	.75	.74	.74	.74	.72
80	Low	Weak	.40	.39	.38	.38	.37	.50	.49	.49	.49	.49
		Strong	.49	.45	.45	.43	.38	.56	.54	.54	.54	.49
	Medium	Weak	.67	.67	.67	.67	.66	.71	.71	.71	.71	.71
		Strong	.71	.69	.69	.69	.66	.72	.72	.72	.72	.71
	High	Weak	.89	.89	.89	.89	.89	.90	.90	.90	.90	.90
		Strong	.90	.89	.89	.89	.89	.90	.90	.90	.90	.90

*Note.* True = four-step approach with the true attributes; PDL = four-step approach with the posterior-distribution level correction weights; SL = four-step approach with the sample-level correction weights; UC = four-step approach without correction; AP = accordion procedure

To evaluate the extent inclusion of the covariates improved the classification certainty, the distributions of  $P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i)$  for the different methods were examined. For illustration purposes, the results for  $D = 4, K(d) = 5, N = 2000$  and short test length are given in Figures 3.1 to 3.3. The x-axis of the figures represents five bins of  $P^*(\alpha_{k(d)}|\mathbf{Y}_i, \mathbf{Z}_i)$ , with an bin width of .1, and the y-axis represents the proportion of examinees. The bin labeled "ge 0.9" contained the examinees with the highest classification certainty, and the bin labeled "le 0.6" the lowest.

Table 3.3:  $CVC_d$  for  $D = 4, K(d) = 8$ 

$J$	Item	Quality	Assoc.	$N = 500$				$N = 2000$				
				True	PDL	SL	UC	AP	True	PDL	SL	UC
64	Low	Weak	.06	.06	.06	.06	.06	.06	.06	.06	.06	.06
		Strong	.06	.06	.06	.06	.06	.07	.06	.06	.06	.06
	Medium	Weak	.18	.17	.17	.17	.17	.24	.24	.24	.24	.23
		Strong	.19	.18	.18	.18	.18	.26	.25	.25	.25	.23
	High	Weak	.47	.46	.46	.46	.46	.55	.55	.55	.55	.54
		Strong	.49	.48	.47	.47	.46	.57	.56	.56	.56	.53
128	Low	Weak	.16	.15	.15	.15	.15	.31	.30	.30	.30	.29
		Strong	.24	.19	.18	.18	.15	.38	.36	.36	.34	.28
	Medium	Weak	.45	.44	.44	.44	.43	.56	.55	.55	.55	.55
		Strong	.49	.47	.47	.46	.43	.58	.57	.57	.57	.54
	High	Weak	.76	.76	.76	.76	.76	.81	.81	.81	.81	.81
		Strong	.78	.77	.78	.77	.76	.82	.81	.81	.81	.81

*Note.* True = four-step approach with the true attributes; PDL = four-step approach with the posterior-distribution level correction weights; SL = four-step approach with the sample-level correction weights; UC = four-step approach without correction; AP = accordion procedure

When the association was weak, incorporating covariates did not strongly influence the classification certainty. For instance, as shown in the first row of Figure 3.1, the proportion of examinees who were classified with the highest certainty increased less than 5% when the four-step approach was used. Moreover, the proportion of examinees at the lowest certainty level did not decrease notably. However, when the association was strong, incorporating covariates resulted in more examinees classified as  $P^*(\alpha_{k(d)}|\mathbf{Y}_i) > .9$ . To take a closer look, when PDL or SL was used, 15% more examinees were classified with the highest certainty level, as the proportions in the other certainty levels all decreased.

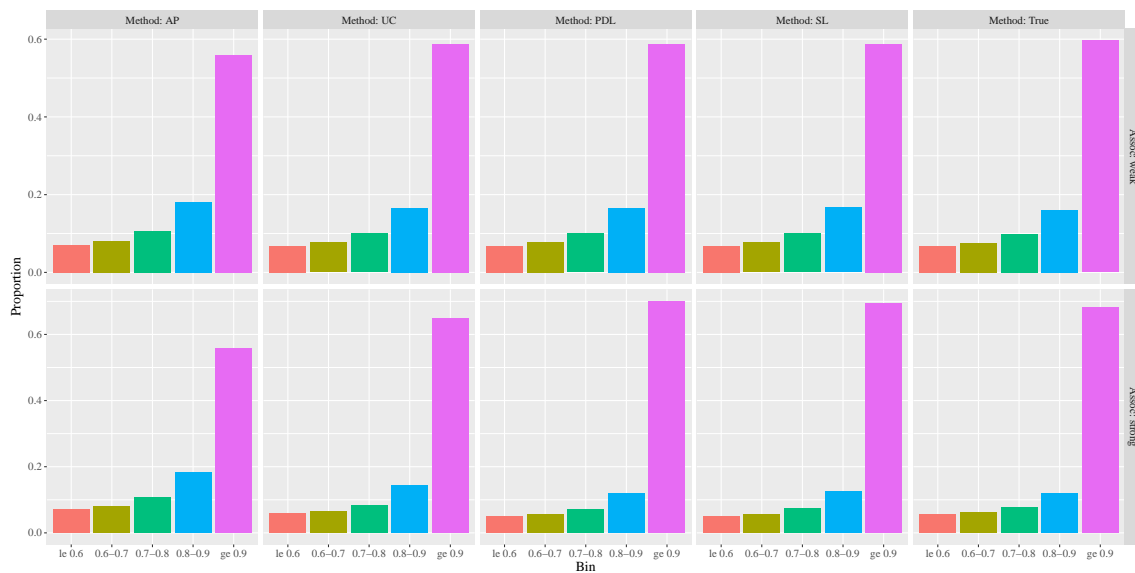


Figure 3.1: Classification Certainty for  $D = 4$ ,  $K(d) = 5$ ,  $N = 2000$ , Short Test, Low Item Quality

Comparing the classification certainty across different item qualities, incorporating covariates performed well when the test was not informative or the test length was not favorable. As the item quality increased from low to high, the improvement in classification certainty became smaller. For example, as can be seen in Figure 3.3, when the item quality was high, the improvement in classification certainty was trivial. These patterns held true when the test length was doubled.



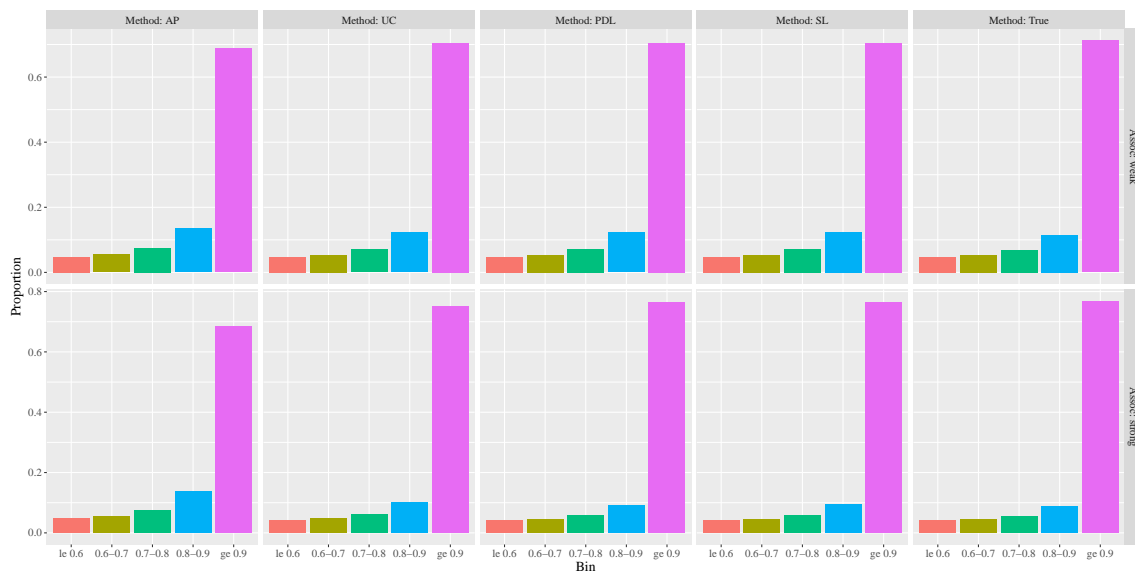


Figure 3.2: Classification Certainty for  $D = 4$ ,  $K(d) = 5$ ,  $N = 2000$ , Short Test, Medium Item Quality

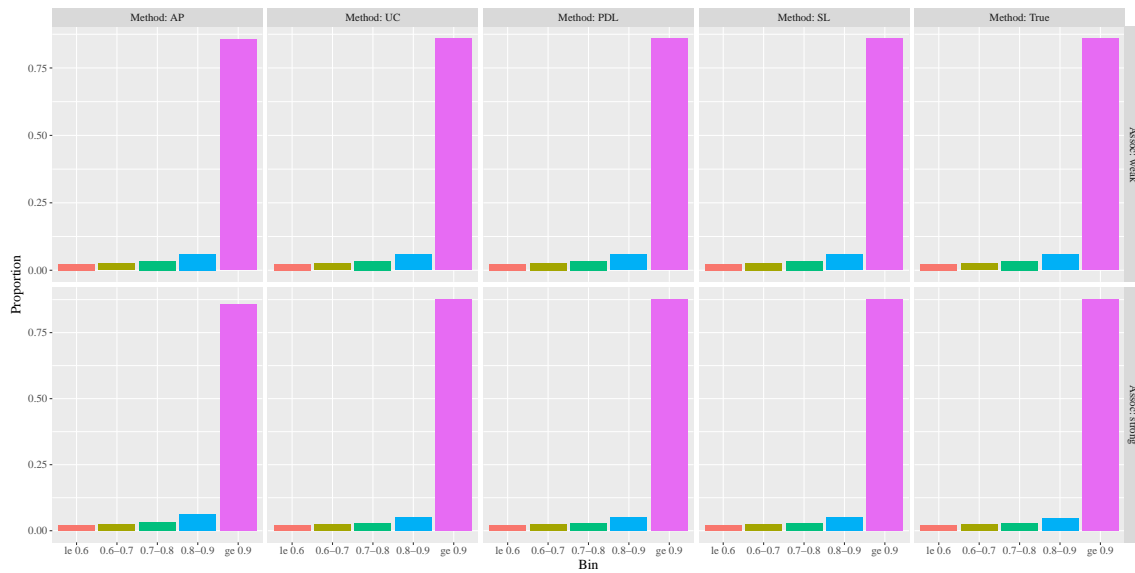


Figure 3.3: Classification Certainty for  $D = 4$ ,  $K(d) = 5$ ,  $N = 2000$ , Short Test, High Item Quality

### 3.5 Real-Data Illustration

The TIMSS 2007 fourth-grade mathematics data were analyzed to illustrate how the four-step method performed in real data. TIMSS is an international educational comparison program for mathematics and science performance designed and conducted by the International Association for the Evaluation of Educational Achievement (IEA; Mullis, Martin, & Foy, 2008). TIMSS utilizes a matrix sampling approach, where each student only responds to one booklet containing both mathematics and science testing items (Mullis et al., 2005). Not only does the TIMSS database contain students' responses to the items in the testing booklets, but it also contains rich background information about the educational context of each student. This makes TIMSS an ideal database to demonstrate how the four-step approach can be used in real-world.

Lee, Park, and Taylan (2011) developed a Q-matrix, which is given in Table 3.5, for the two fourth-grade mathematics blocks, M03 and M04, found in booklets 4 and 5. More details about the test design can be found in the TIMSS 2007 technical report (Olson, Martin, & Mullis, 2008). There are a total of 15 multiple-choice and 10 constructed-response items included in this Q-matrix. The Q-matrix was developed by first defining 15 attributes under three content domains, namely, Number (N), Geometric Shapes and Measures (GM), and Data Display (DD), based on the TIMSS 2007 mathematics framework (Lee et al., 2011). The definitions of the attributes are itemized in Table 3.4.

The final Q-matrix of the 25 selected items was formed independently by mathematics educators and combined through discussions. Most of the items in the Q-matrix are

dichotomous response items, except two items: M041275 and M031247 which were polytomous response items with a maximum score of 2. These two items were dichotomized such that examinees scored a 1 if the highest score was achieved, and 0 otherwise.

After reviewing possible student background variables that may provide ancillary information, five variables were selected as covariates, namely, gender, language of testing, plausible values (PVs) of earth science, PVs of life science, and PVs of physics. Student's gender and language of testing are dichotomous variables. Gender was coded as 1 for male and 0 for female. For language of testing, because some participant countries use more than one language for daily classroom activities, TIMSS provided multiple forms of tests in different languages. To track different test languages students used in TIMSS 2007, language of testing was coded as 1 if the primary test language was used, and 0 if the country-specific language was used. The PVs for three science content domains were also selected as covariates. TIMSS provided five PVs, which can be deemed as five ability estimates with certain amount of uncertainty for each student, for each content domain as well as for the overall performance. More details about PVs can be found in TIMSS 2007 technical report Olson et al. (2008). All covariates were standardized prior to the analysis.

Nine countries or regional entities, namely, 1) England, 2) US, 3) Australia, 4) Ontario, Canada, 5) Alberta, Canada, 6) British Columbia, Canada, 7) New Zealand, 8) Massachusetts, US, and 9) Minnesota, US were selected from the TIMSS 2007 database as the analysis sample. Note that the US sample is different from the Massachusetts and Minnesota sample which served as benchmark states in TIMSS 2007 (Olson et al., 2008). These countries or regional entities were selected because they were all primarily English-speaking countries that participated in TIMSS 2007, and the majority of the students in

Table 3.4: Attributes Identified for TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5

Domain	Attribute
N	$\alpha_{1(1)}$ : representing, comparing and ordering whole number as well as demonstrating knowledge of place value
	$\alpha_{1(2)}$ : recognizing multiples, computing with whole numbers using the four operations, and estimating computations
	$\alpha_{1(3)}$ : solve problems, including those set in real life contexts
	$\alpha_{1(4)}$ : solve problems involving proportions
	$\alpha_{1(5)}$ : recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents
	$\alpha_{1(6)}$ : solve problems involving simple fractions and decimals including their addition and subtraction
	$\alpha_{1(7)}$ : find the missing number or operation and model simple situations involving unknowns in number sentence or expressions
	$\alpha_{1(8)}$ : describe relationships in patterns and their extensions
GM	$\alpha_{2(1)}$ : measure, estimate and understand properties of lines and angles and be able to draw them
	$\alpha_{2(2)}$ : classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties
	$\alpha_{2(3)}$ : calculate and estimate perimeters, area and volume
	$\alpha_{2(4)}$ : locate points in an informal coordinate to recognize and draw figures and their movement
DD	$\alpha_{3(1)}$ : read data from tables, pictographs, bar graphs, and pie charts
	$\alpha_{3(2)}$ : compare and understand how to use information from data
	$\alpha_{3(3)}$ : understand different representations and organizing data using tables, pictographs, and bar graphs

*Note.* N = number; GM = geometric shapes and measures; DD = data display

Table 3.5: Q-matrix for TIMSS 2007 Fourth-Grade Mathematics Booklets 4 and 5

Item	Domain: N								Domain: GM				Domain: DD		
	$\alpha_{1(1)}$	$\alpha_{1(2)}$	$\alpha_{1(3)}$	$\alpha_{1(4)}$	$\alpha_{1(5)}$	$\alpha_{1(6)}$	$\alpha_{1(7)}$	$\alpha_{1(8)}$	$\alpha_{2(1)}$	$\alpha_{2(2)}$	$\alpha_{2(3)}$	$\alpha_{2(4)}$	$\alpha_{3(1)}$	$\alpha_{3(2)}$	$\alpha_{3(3)}$
M041052	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M041056	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M041069	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
M041076	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
M041281	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
M041164	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
M041146	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
M041152	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
M041258A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M041258B	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
M041131	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0
M041275	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
M041186	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
M041336	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0
M031303	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031309	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031245	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
M031242A	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
M031242B	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
M031242C	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
M031247	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
M031219	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
M031173	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M031085	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M031172	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1

Note. N = number; GM = geometric shapes and measures; DD = data display

the sample used English as their test language. Students who were not assigned to booklet 4 or 5, or did not have any valid responses on the two booklets were dropped from the analysis sample. The final sample size was 5236.

Five different approaches were used to analyze the dataset, namely, the complete-profile estimation (COM), AP, PDL, SL, and UC. The results were compared in terms of the classification certainty and the point-biserial correlations, denoted by  $r_{pb}$ , between attribute classification and overall math performance. For the latter index, a high attribute-total correlation can be expected if students are accurately classified, and vice versa. The five PVs were treated as five random samples and results were averaged across PVs. The G-DINA model was used to estimate the data, and EAP was used to estimate attribute profiles.

The  $R^2_{McFadden}$  of the latent regressions, as shown in Table 3.6, ranged from .01 to .18 for domain N, .01 to .20 for domain GM, and .14 to .19 for domain DD, indicating that the covariates were not equally correlated with each attribute. To better demonstrate the classification certainty results, three attributes, namely,  $\alpha_{(2)3}$ ,  $\alpha_{(3)3}$ , and  $\alpha_{(1)3}$ , were selected to represent three distinct scenarios. Specifically,  $\alpha_{(2)3}$  had a low initial certainty using AP, with an average  $P^*(\alpha_{k(d)}|\mathbf{Y}_i)$  of .62, and a low  $R^2_{McFadden}$  of .01;  $\alpha_{(3)3}$  had a low initial classification certainty of .55 but a high  $R^2_{McFadden}$  of .19; and  $\alpha_{(1)3}$  had a high initial classification certainty of .92 and a high  $R^2_{McFadden}$  of .18. Note that attributes with a high initial classification certainty and low  $R^2_{McFadden}$  cannot be found in this dataset. As shown in Figure 3.4, similar patterns observed in the simulation study can be found in the real-data analysis. Specifically, when the initial classification certainty was low (i.e.,  $\alpha_{(2)3}$  and  $\alpha_{(3)3}$ ), incorporating covariates increased the number of examinees with more certain

classifications. AP and COM resulted in the worst classification accuracy, whereas the four-step approaches classified more examinees with higher certainty. In contrast, when the initial classification certainty was high, incorporating the covariates did not improve the classification results remarkably.

Last, the  $r_{pb}$  between attribute estimates and the overall mathematics performance are demonstrated in Table 3.6. As can be seen from the table, the correlations obtained by PDL and SL were always higher than those obtained by AP and COM. Also, although UC resulted in higher correlations for several attributes, when  $R^2_{McFadden}$  was close to 0, incorporating covariates without correction weights may be problematic, as the correlation coefficients decreased, even reversed. Examining the  $r_{pb}$  for attributes involved in Figure 3.4, the attributes with the lowest certainty (i.e.,  $\alpha_{(2)3}$  and  $\alpha_{(3)3}$ ) resulted in the highest changes in the correlations when covariates were incorporated. In contrast,  $\alpha_{(1)3}$ , which was well estimated without covariates, had only a small increase in the correlation after incorporating covariates.

Table 3.6: McFadden's  $R^2$  and  $r_{pb}$  of Attribute Classification and Overall Mathematics Performance

Domain	Attribute	$R^2_{McFadden}$	$r_{pb}$				
			PDL	SL	UC	AP	COM
N	$\alpha_{1(1)}$	.15	.63	.64	.61	.54	.59
	$\alpha_{1(2)}$	.17	.66	.66	.63	.61	.61
	$\alpha_{1(3)}$	.18	.69	.69	.68	.65	.65
	$\alpha_{1(4)}$	.04	.63	.63	.10	.27	.22
	$\alpha_{1(5)}$	.14	.61	.62	.58	.57	.56
	$\alpha_{1(6)}$	.13	.58	.58	.57	.57	.57
	$\alpha_{1(7)}$	.01	.52	.48	.00	.16	.16
	$\alpha_{1(8)}$	.11	.62	.63	.62	.52	.50
GM	$\alpha_{2(1)}$	.11	.61	.61	.60	.53	.56
	$\alpha_{2(2)}$	.20	.69	.69	.68	.66	.60
	$\alpha_{2(3)}$	.01	.55	.41	-.24	.07	.05
	$\alpha_{2(4)}$	.05	.60	.60	.58	.27	.53
DD	$\alpha_{3(1)}$	.18	.60	.60	.60	.56	.42
	$\alpha_{3(2)}$	.14	.63	.61	.60	.55	.03
	$\alpha_{3(3)}$	.19	.59	.59	.53	.55	.23

*Note.* N = number; GM = geometric shapes and measures; DD = data display

### 3.6 Discussions and Conclusions

This study explored incorporating covariates to improve CDM classification for high-dimensional testing data. By extending the three-step approach introduced by Iaconangelo (2017) and Vermunt (2010) to the four-step approach, the classification accuracy can be improved by borrowing information from relevant covariates. This study also tackled a particular situation where the number of attributes in a test is large, thus making the



Figure 3.4: Classification Certainty for  $\alpha_{(2)3}$ ,  $\alpha_{(3)3}$ , and  $\alpha_{(1)3}$ 

one-step latent regression approach intractable. The performance of the proposed four-step approach was evaluated in simulation and real-data studies. In the simulation study, where a relatively high-dimensional setting was created, AP with the four-step approach increased the classification accuracy, provided there was a strong association between the attributes and the covariates, and the test was not sufficiently long or the item quality was not high. In addition, the TIMSS 2007 fourth-grade mathematics data were fitted using the four-step approach with gender, test language, and science domain abilities being the covariates. The results demonstrate that, when covariates are added, the correlations between attribute classification and overall mathematics performance are increased.

Several advantages of the four-step approach are worth mentioning. To begin with, the four-step approach never harms the classification even if the covariates are not correlated with attributes. As demonstrated in the simulation study, the classification accuracy of

the four-step approach is always greater or equal to that of AP. When covariates are not associated with attributes,  $P(\alpha_{k(d)}|\mathbf{Z}_i)$  will be close to  $P(\alpha_{k(d)})$ , which is the proportion of examinees with  $\alpha_{k(d)}$ , then the posterior probability  $P(\alpha_{k(d)}|\mathbf{Y}_i)$  will be weighted by  $P(\alpha_{k(d)})$  to compute the combined posterior probability in Equation 3.14. Second, because the four-step approach has a great deal of flexibility, researchers have the options to regress attributes on different sets of covariates, or not to use covariates if they are irrelevant to some of the attributes. In practice, as shown in the TIMSS 2007 example, the covariates may not be equally correlated with the attributes, hence, researchers have the freedom to decide when to incorporate covariates and when not to.

Third, the proposed procedure was developed in conjunction with AP, which is computationally much faster than the complete-profile estimation as demonstrated by (de la Torre, 2017), the CPU time of AP is only at most 37% of that of the complete-profile estimation. By taking the advantage of AP's the computation efficiency, the four-step approach is also computationally efficient. Additionally, because each domain is analyzed independently, the four-step program can be easily programmed to be paralleled on multiple threads to gain extra computational efficiency. In the real data example, the four-step approach took only 3.56 minutes with parallel computing in R.

Last, because the four-step approach can dramatically improve the classification certainty, it can be integrated with computerized adaptive testing (CAT) with variable length stopping rules. If some of relevant covariates are available prior to the test administration, they can be used to provide information about examinees' profiles. Hence, incorporating covariates in CAT may be used to shorten the test length.

One limitation of this study is that only a relatively small set of covariates without severe multicollinearity were used in both the simulation and real-data analyses. In practice, when the number of covariates increases, multicollinearity could be an issue. Hence certain variable selection and regularization methods are needed. Developing these methods to work in conjunction with the four-step approach can be a future research direction.

### 3.7 References

- Ackerman, T. A., & Davey, T. C. (1991). *Concurrent adaptive measurement of multiple abilities*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification*, *30*, 195–224.
- Bellman, R. (1957). *Dynamic programming*. Princeton: Princeton University Press.
- Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3–27.
- Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173–178.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.
- de la Torre, J. (2017). *The accordion approach: A method for accommodating a large number of attributes in cognitive diagnosis modeling*. Paper presented at the Global Chinese Conference on Educational Information and Assessment - Chinese Association of Psychological Testing Annual Conference, Taichung, Taiwan.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.
- DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 970–1030). Amsterdam: North-Holland Publications.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.
- Hartz, S. M., & Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series*, 2008.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407–419.

- Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models* (Doctoral dissertation). Retrieved from <https://doi.org/doi:10.7282/T3W95D95>.
- Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 300–303.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, 28, 407–426.
- Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, 82, 112–132.
- Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11, 144–177.
- Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11, 81–91.
- Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54, 661–679.
- Mullis, I. V., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Olson, J., Martin, M., & Mullis, I. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, 38, 376–390.
- R Core Team. (2013). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287–305.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the*

- Royal Statistical Society. Series B (Methodological)*, 267–288.
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450–469.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287–307.

## Chapter 4

# Strategies for Implementing CD-CAT in High-Dimensional Testing Situations

### Abstract

Cognitive diagnosis computerized adaptive testing (CD-CAT) has been proposed to make testing more efficient by selecting the optimal set of items for each examinee. However, when the number of attributes of interest is large, implementing CD-CAT becomes infeasible. This study proposes a series of strategies to implement CD-CAT in high-dimensional testing situations. Specifically, a novel procedure is proposed as item pool calibration method. An item selection method is modified to work in conjunction with the proposed calibration method. Moreover, two procedures to incorporate covariates in CD-CAT are explored to shrink the prior distributions of examinees. Simulation studies are conducted to evaluate the performance of the proposed approaches. The results show the proposed methods can be applied to high-dimensional situations with improved testing efficiency.

**Keywords:** accordion procedure, cognitive diagnosis model, computerized adaptive testing, high-dimensional data

Cognitively diagnostic assessments (CDAs) are formative assessments designed to provide information regarding students' cognitive strengths and weaknesses on a set of fine-grained skills (de la Torre & Minchen, 2014; DiBello, Roussos, & Stout, 2007). CDAs are particularly useful when it can provide immediate diagnostic feedback, based on which adjustments to classroom instructions or remedial measures can be made. One of the recent developments in delivering a CDA is to utilize the computerized adaptive testing (CAT) technique, and thus is referred to as cognitive diagnosis computerized adaptive testing (CD-CAT; Cheng, 2009; Hsu, Wang, & Chen, 2013; Kaplan, de la Torre, & Barrada, 2015; McGlohen & Chang, 2008; X. Xu, Chang, & Douglas, 2003). Originally developed for item response theory (IRT) models, CAT provides a test tailored to each examinee, where items are selected based on certain psychometric properties (Lord, 1971; Meijer & Nering, 1999; van der Linden & Glas, 2000). Compared to traditional paper-and-pencil test, CAT can produce more accurate ability estimates with shorter test length (H.-H. Chang & Ying, 1996; Lord, 1980). Some pioneering work of real-world CD-CAT applications can be found in the literature. For example, Liu, You, Wang, Ding, and Chang (2013) developed a Browser/Server based CD-CAT delivery system for a large-scale English test with eight attributes for 5th and 6th grade students in China. In addition, Wu, Kuo, and Yang (2012) implemented the knowledge-structure based adaptive testing algorithm to diagnose 5th grade students' mathematics skills in Taiwan.

Despite the potential benefits of CD-CAT, few research has explored its application in high-dimensional testing situations. Several studies have used a relatively low number of attributes (i.e.,  $K \leq 8$ ) for the simulation or real-data studies (e.g., Cheng, 2009; Hsu et



al., 2013; Kaplan et al., 2015; Liu et al., 2013). When applying existing CD-CAT techniques to high-dimensional settings, such as learning map modeling (Dynamic Learning Maps Science Consortium, 2015), several challenges emerge. First, similar to the problem encountered in fitting cognitive diagnosis models (CDMs) to high-dimensional data, calibration of item pools with  $K > 20$  attributes becomes nearly impossible. Second, computing item selection indices which require marginalizing across all  $2^K$  latent classes can be both time and computer memory consuming. Third, to terminate CD-CAT using a high minimum of the maximum (minimax) posterior stopping rule may require impractically long tests than in situations where only a small set of skills are of interest. Therefore, larger item pools or higher quality items are needed to shorten the test length, which can substantially increase the cost of the test development.

To address the above issues, this study proposes: 1) a novel procedure to calibrate item pools that involve a considerably much larger set of attributes; 2) a modified item selection method to better fit the current context; and 3) a way to incorporate ancillary information to better estimate examinees' prior probability distributions to further increase the efficiency of implementing CD-CAT in high-dimensional settings.

The rest of this paper is organized as follows. A brief introduction to the technical background of CDMs and CD-CAT is given next, followed by the proposed strategies. Simulation studies are provided to evaluate the performance of the proposed strategies. Finally, the limitations and possible future research are discussed to conclude the paper.

## 4.1 Cognitive Diagnosis Models

CDMs (see, for a range of examples, de la Torre, 2009, 2011; Haertel, 1989; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Rupp & Templin, 2008; Templin & Henson, 2006; von Davier, 2008) are essentially restricted latent class models that can be used to analyze CDA data to estimate examinees' profiles of a set of fine-grained attributes. Let  $\alpha_i = \{\alpha_{ik}\}$  be the attribute profile of examinee  $i$ , where  $i = 1, \dots, N$ . The  $k$ th element, where  $k = 1, \dots, K$ , equals to 1 if the examinee masters attribute  $k$ ; and 0 otherwise. Let  $Y_{ij}$  denote the observed response of examinee  $i$  on item  $j$ , where  $j = 1, \dots, J$ . The item-attribute association is a key component of CDMs, and is referred to as the Q-matrix (Tatsuoka, 1983, 1985). The  $q$ -vector for item  $j$  is denoted by  $\mathbf{q}_j = \{q_{jk}\}$ , where the  $k$ th entry of 1 indicates answering item  $j$  correctly requires attribute  $k$ ; and 0 otherwise.

Various CDMs model the probability of success,  $P(Y_{ij} = 1 | \alpha_i)$ , as different functional forms. For example, as one of the general CDMs, the generalized deterministic inputs, noisy "and" gate (G-DINA; de la Torre, 2011) model allows freely estimate the probability of success for each distinguishable latent group based on the  $q$ -vector of an item. Suppose item  $j$  measures  $K_j$  attributes. Item  $j$  can distinguish examinees into  $2^{K_j}$  unique latent groups, each of which represents a reduced attribute profile denoted by  $\alpha_{lj}$ ,  $l = 1, \dots, 2^{K_j}$ . Examinees in the same latent group are assumed to have the same probability of success. For example, if the  $q$ -vector for item  $j$  is  $\mathbf{q}_j = (1, 1, 0)$ , it can distinguish four latent groups with reduced attribute profiles of  $\alpha_{1j} = (0, 0, \cdot)$ ,  $\alpha_{2j} = (1, 0, \cdot)$ ,  $\alpha_{3j} = (0, 1, \cdot)$  and  $\alpha_{4j} = (1, 1, \cdot)$ , where  $\cdot$  represents a free entry. The G-DINA model in the identity link can be

written as

$$P(Y_{ij} = 1 | \alpha_{lj}) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (4.1)$$

where  $k'$  indicates another attribute other than  $k$ ,  $\delta_{j0}$  is the intercept for item  $j$ ,  $\delta_{jk}$  is the  $k$ th attribute's main effect,  $\delta_{jkk'}$  is the two-way interaction between attribute  $k$  and  $k'$ , and  $\delta_{j12\dots K_j}$  is the  $K_j$ -way interaction.

## 4.2 CD-CAT

Weiss and Kingsbury (1984) summarized the necessary components for an IRT-based CAT system, namely, 1) a measurement model, 2) an item pool, 3) an entry level, 4) an item selection rule, 5) a scoring method, and 6) a termination rule. In the realm of CD-CAT, all of the above components can be still adopted with certain modifications. For instance, IRT models are used to calibrate item pools in CAT system, whereas, CDMs are used correspondingly in CD-CAT as calibration models. In this section, three major components, namely, the entry level, item selection methods and termination rules are discussed in the context of CD-CAT.

### 4.2.1 Entry Level

The entry level in IRT-based CAT is referred to the difficulty levels for the items selected at the beginning stage of CAT. In CD-CAT, however, due to the multidimensional nature of the latent variables in CDMs, previous research have proposed random selection (Kaplan et al., 2015), using item selection method with a flat prior (Cheng, 2009; Hsu et al., 2013),

or the Q-optimal procedure for the selection of the initial items (Y.-P. Chang, Chiu, & Tsai, in press; G. Xu, Wang, & Shang, 2016), when no information regarding distributions of examinees' latent classes is available.

## 4.2.2 Item Selection Method

Several item selection methods can be found in the literature, including the Shannon entropy-based method (Shannon, 1948; X. Xu et al., 2003), the Kullback-leibler (KL) divergence-based method and its modifications (see, Cheng, 2009; Cover & Thomas, 1991; Kaplan et al., 2015; X. Xu et al., 2003), and the G-DINA model discrimination index (GDI; Kaplan et al., 2015) method. The KL-based methods and GDI are briefly reviewed below.

The KL based item selection methods, such as the KL (Cover & Thomas, 1991; X. Xu et al., 2003), the posterior-weighted KL (PWKL; Cheng, 2009), and the modified posterior-weighted KL (MPWKL; Kaplan et al., 2015), have been proposed in the CD-CAT context. The KL divergence measures the distance between two probability density functions: the true and the model predicted distribution (Cover & Thomas, 1991). X. Xu et al. (2003) applied it as an item selection method in CD-CAT. At stage  $s$ , the KL is given by

$$KL_{ij}^{(s)} = \sum_{c=1}^{2^K} \sum_{y=1}^1 \log \left( \frac{P(Y_{ij} = y | \hat{\alpha}_i^{(s)})}{P(Y_{ij} = y | \alpha_c)} \right) P(Y_{ij} = y | \hat{\alpha}_i^{(s)}), \quad (4.2)$$

where  $\hat{\alpha}_i^{(s)}$  is the estimate of the attribute profile of examinee  $i$  at stage  $s$ ,  $\alpha_c$  is another possible attribute profile, and  $c = 1, \dots, 2^K$ . The KL divergence compares the estimated attribute profile at current stage with all possible attribute profiles, and aggregates the

distances using unweighted sum. The item that maximizes the KL divergence is selected at stage  $s + 1$ .

One disadvantage of using the KL as an item selection method is that it assumes, at any stage of CD-CAT, an examinee is equally likely to belong to each latent class, which is inefficient because it ignores the information provided by item responses in previous stages (Cheng, 2009). In addition, the KL assumes the point estimate of attribute profile is accurate at a given stage. This is not necessarily the case especially at the early stages of CD-CAT (Kaplan et al., 2015). Kaplan et al. (2015) proposed the MPWKL to address these issues. The MPWKL is formulated as

$$MPWKL_{ij}^{(s)} = \sum_{d=1}^{2^K} \left\{ \sum_{c=1}^{2^K} \left[ \sum_{y=0}^1 \log \left( \frac{P(Y_{ij} = y | \alpha_d)}{P(Y_{ij} = y | \alpha_c)} \right) P(X_{ij} = y | \alpha_d) \pi^{(s)}(\alpha_c) \right] \pi^{(s)}(\alpha_d) \right\}, \quad (4.3)$$

where  $\alpha_c$  and  $\alpha_d$  are two attribute profiles, and  $\pi^{(s)}(\alpha_c)$  and  $\pi^{(s)}(\alpha_d)$  represent their posterior probabilities. The MPWKL considers all possible  $2^K$  attribute profiles in the numerator rather than a single point estimate of an examinee's attribute profile. The item with the maximum MPWKL is selected to be administered at stage  $s + 1$ .

Another item selection method is the GDI (Kaplan et al., 2015), which is denoted by  $\zeta_{ij}^{2(s)}$ . The GDI was originally developed as a method for validating Q-matrix empirically by de la Torre and Chiu (2016). It measures the weighted variance of probabilities of success among collapsed latent groups for a given item. In the CD-CAT context, the GDI is formulated as

$$\zeta_{ij}^{2(s)} = \sum_{l_j=1}^{2^{K_j}} \pi^{(s)}(\alpha_{l_j}) \left[ P(Y_{ij} = 1 | \alpha_{l_j}) - \bar{P}_j \right]^2, \quad (4.4)$$

where  $\pi^{(s)}(\alpha_{l_j})$  is the posterior probability of the collapsed latent group  $\alpha_{l_j}$  at stage  $s$ ,

and  $\bar{P}_j = \sum_{l_j=1}^{2^{K_j}} \pi^{(s)}(\alpha_{lj})P(Y_{ij} = 1|\alpha_{lj})$ . In CD-CAT, the GDI is an indicator of how well an item can differentiate among collapsed latent groups at a given stage. Note that unlike the GDI proposed for validating Q-matrix, the GDI used for item selection is not static for a given item. That is, when the posterior distribution changes as CD-CAT proceeds, the GDI for an item will change as well. The item with the highest GDI at stage  $s$  is the most discriminating item for an examinee, and will be selected at stage  $s + 1$ . A comparison among the PWKL, MPWKL and GDI showed that the MPWKL and the GDI outperformed the PWKL (Kaplan et al., 2015); however, the GDI was shown to be more computationally efficient than the MPWKL. For these reasons, the GDI will be used in this study.

### 4.2.3 Termination Rule

Another important aspect of CD-CAT is how to terminate the test. A fixed-length termination rule can be adopted in CD-CAT such that each examinee needs to complete the same number of items. This termination rule seems to be fair in terms of the number of items responded by each examinee, but not so in terms of measurement accuracy (Hsu et al., 2013). Because examinees with different proficiency profiles may be assigned to respond to different sets of items, when the length is fixed, the information provided by one set may be different from that provided by another.

The posterior distribution reflects the precision of the classification (Huebner, 2010): the more peaked the distribution is, the more reliable estimate of attribute profile can be obtained. Hsu et al. (2013) proposed the variable-length termination rule based on the

largest posterior probability of latent classes. This is also referred to as the minimax posterior probability (Kaplan et al., 2015). The minimax posterior termination rule stops a test when the maximum posterior probability reaches a prespecified cutoff so that examinees are expected to achieve a minimum precision when CD-CAT stops.

### 4.3 Calibrating Item Pools with High Dimensionality

When a test involves a large set of attributes, regular calibration methods, such as marginal maximum likelihood estimation with expectation-maximization (MMLE-EM; de la Torre, 2009, 2011; von Davier, 2008), become intractable due to computational constraints. The accordion procedure (AP; de la Torre, 2017) offers a solution to cope with high-dimensional scenarios, where attributes can be partitioned into non-overlapping subsets, such as different content domains, based, say, on a higher-order (de la Torre & Douglas, 2004) knowledge structure. AP simplifies the issue in estimating one large model into estimating several models with a smaller set of attributes. Specifically, AP reduces the test dimensionality by collapsing nontarget attributes into coarser-grained attributes, and repeating the process until all attributes are estimated.

Suppose there are  $D$  domains involved in a test, and domain  $d$  consists of  $K(d)$  attributes, where  $d = 1, \dots, D$ . The attribute profile for domain  $d$  is denoted by  $\alpha_{l(d)} = (\alpha_{l(d)1}, \dots, \alpha_{l(d)K(d)})$ , where  $l(d) = 1, \dots, 2^{K(d)}$ . To apply AP to a target domain, the original Q-matrix needs to be modified accordingly by collapsing attributes in the nontarget domains such that the collapsed domain-level  $q$ -entry  $q_{j(d')} = 1$ , if at least one attribute in nontarget domain  $d'$  is required by item  $j$ ; otherwise,  $q_{j(d')} = 0$ . Results based on simulation studies show that the attribute-level classification accuracy of AP is comparable with

the complete-profile estimation, and the computation time of AP is much shorter (de la Torre, 2017). More details of AP can be found in Chapter 2.

There are several benefits of using AP to calibrate item pools. First, AP can work with items measuring skills in multiple domains so that when developing new items, they are not constrained to be written to be associated only with skills in one specific domain. In addition, for a given target domain of AP, the dimensionality is reduced so that existing item selection methods can be calculated based on the collapsed attribute profiles, which makes computation more feasible.

Once item pools are calibrated using AP, one issue emerges, as in, the parameters of an item are stored separately by different target domains. For example, for item  $j$ , the item parameters obtained by AP has potentially  $D$  sets of parameters. Thus, CD-CAT needs to be carried out for each domain until a fixed-length or variable-length termination rule is satisfied. Additionally, the existing item selection methods need further modifications to be better suited for the context of AP.

#### **4.4 The Modified-GDI**

Because calibrated item pools contain coarse-grained nuisance attributes, which are not of interest, the information regarding nuisance attributes needs to be marginalized to guarantee that items selected by CD-CAT have a clear focus on the target domain. The GDI item selection method (Kaplan et al., 2015) is modified so that the most informative items focusing on the skills of the target domain are selected. The modified-GDI (M-GDI) for



examinee  $i$ , item  $j$  on the domain  $d$  is given by

$$\text{M-GDI}_{ij(d)}^{(s)} = \sum_{l(d)=1}^{2^{K(d)}} \hat{\pi}_i^{(s)}(\alpha_{l(d)}) \left[ \hat{P}(Y_{ij} = 1 | \alpha_{l(d)}) - \bar{P}_j \right]^2, \quad (4.5)$$

where

$$\hat{P}(Y_{ij} = 1 | \alpha_{l(d)}) = \frac{\sum_{d' \neq d}^1 \alpha_{(d')} \dots \sum_{d'' \neq d}^1 \alpha_{(d'')} \hat{P}(Y_{ij} = 1 | \alpha_{l'}) \hat{\pi}_i^{(s)}(\alpha_{l'})}{\sum_{d' \neq d}^1 \alpha_{(d')} \dots \sum_{d'' \neq d}^1 \alpha_{(d'')} \hat{\pi}_i^{(s)}(\alpha_{l'})}, \quad (4.6)$$

$d'$  and  $d''$  are two non-target domains, and  $\bar{P}_j$  is the weighted average of probabilities of success for item  $j$ . Equation 4.5 can be calculated using either the direct output of AP, or the marginalized probability of success and posterior probability for the target domain.

Table 4.1 gives an example of the success probabilities estimated by AP, and the estimated posterior probabilities for latent class  $l'$  containing nuisance attributes and  $l(d)$  without nuisance attributes. In this example, the fourth attribute is a nuisance attribute and the latent classes that are distinct at the fourth attribute are collapsed to obtain the success probabilities associated with target domains,  $\hat{P}(Y_{ij} = 1 | \alpha_{l(d)})$ , and the marginalized posterior probabilities  $\pi_i^{(s)}(\alpha_{l(d)})$ . The M-GDI of this example is calculated as

$$\text{M-GDI} = .12 \times (.28 - .42)^2 + \dots + .18 \times (.63 - .42)^2 = .02,$$

where .42 is the weighted average of the probabilities of success across latent classes. Note that the M-GDI is domain specific; therefore, an informative item for one domain may not be informative for another. Same as the GDI item selection method, the item with the largest M-GDI is selected.

Table 4.1: Posterior Probabilities and Success Probabilities for Latent Class  $l'$  and  $l(d)$ 

$\alpha_{l'}$	(0,0,0,0)	(1,0,0,0)	(0,1,0,0)	(0,0,1,0)	(0,0,0,1)	(1,1,0,0)	(1,0,1,0)	(1,0,0,1)
$\hat{\pi}_i(\alpha_{l'})$	.05	.10	.06	.04	.07	.03	.10	.05
$\hat{P}(Y_{ij} \alpha_{l'})$	.10	.30	.10	.10	.40	.30	.30	.80
	(0,1,1,0)	(0,1,0,1)	(0,0,1,1)	(1,1,1,0)	(1,1,0,1)	(1,0,1,1)	(0,1,1,1)	(1,1,1,1)
$\hat{\pi}_i(\alpha_{l'})$	.04	.02	.09	.06	.04	.06	.07	.12
$\hat{P}(Y_{ij} \alpha_{l'})$	.10	.40	.40	.30	.80	.80	.40	.80
$\alpha_{l(d)}$	(0,0,0,-)	(1,0,0,-)	(0,1,0,-)	(0,0,1,-)	(1,1,0,-)	(1,0,1,-)	(0,1,1,-)	(1,1,1,-)
$\hat{\pi}_i(\alpha_{l(d)})$	.12	.15	.08	.13	.07	.16	.11	.18
$\hat{P}(Y_{ij} \alpha_{l(d)})$	.28	.47	.18	.31	.59	.49	.29	.63

#### 4.4.1 Estimating Prior Distributions using Covariates

Multi-step latent regression approaches which incorporate covariates into CDMs can be found in the literature (see, for example, Iaconangelo, 2017). In a four-step latent regression approach, CDMs are fitted to the data in the first step, followed by the assignment of examinees' latent classes. Individual attributes or attribute profiles are then regressed onto the covariates with the appropriate correction weights to account for the uncertainty in the classifications. Last, the examinees' posterior probabilities are updated combining the information obtained by CDMs and latent regressions. Inferences on each examinee's attribute profile can be drawn from the updated individual posterior distribution. See Chapter 3 for more details of the four-step latent regression procedure.

In CD-CAT, when covariates, such as examinees' performance on other subject matters or previous item responses, are available, they can be used as ancillary information to estimate the probability that an examinee belongs to a given latent class. Therefore, a relatively peaked rather than flat prior distribution can be used to optimize the item selection at the early stage of CD-CAT.

Two possible cases where ancillary information can be used to estimate the prior distribution of an examinee are considered. For case 1, suppose previous item responses,  $\mathbf{Y}^{(0)} = \{\mathbf{Y}_i^{(0)}\}$ , and covariates,  $\mathbf{Z}^{(0)} = \{\mathbf{Z}_i^{(0)}\}$ , are available prior to CD-CAT. When CD-CAT starts, the first item selected for examinee  $i$  is based on the M-GDI with  $\pi_i^{(1)}(\alpha_{l(d)})$  estimated by  $P(\alpha_{l(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)})$ . Assuming local independence among attributes,  $P(\alpha_{l(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)})$  can be obtained by computing the joint probability:

$$P(\alpha_{l(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)}) = \prod_{k(d)=1}^{K(d)} P(\alpha_{k(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)})^{\alpha_{k(d)}} [1 - P(\alpha_{k(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)})]^{[1-\alpha_{k(d)}]}, \quad (4.7)$$

where  $P(\alpha_{k(d)}|\mathbf{Y}_i^{(0)}, \mathbf{Z}_i^{(0)})$ , obtained by the four-step latent regression approach, is the probability of the mastery of  $\alpha_{k(d)}$  conditional on  $\mathbf{Y}_i^{(0)}$  and  $\mathbf{Z}_i^{(0)}$ . Note that, if no informative covariates is available in this case, the posterior probability  $P(\alpha_{l(d)}|\mathbf{Y}_i^{(0)})$  estimated using previous response data can serve as the prior probability for CD-CAT.

Case 2 considers a situation where there is no available previous response data for a group of examinees to be tested using CD-CAT. In this situation, examinees' prior distributions can be estimated using relationships between the attributes and covariates that have been previously estimated, provided the past testing data are available. Let  $\hat{\alpha}^{(A)}$  and  $\mathbf{Z}^{(A)}$  be the attribute profile matrix and covariate matrix of a previous testing sample  $A$ , respectively. Let  $\hat{\Lambda} = \{\hat{\lambda}_{k(d)}\}$  denote the estimated regression coefficients using the three-step procedure by regressing  $\hat{\alpha}^{(A)}$  onto  $\mathbf{Z}^{(A)}$ . For a different sample  $B$ , let  $\mathbf{Z}^{(B)} = \{\mathbf{Z}_i^{(B)}\}$  denote the covariates of examinees in sample  $B$ .  $\hat{P}(\alpha_{k(d)}|\mathbf{Z}_i^{(B)}, \hat{\lambda}_{k(d)})$  can be obtained by

$$\hat{P}(\alpha_{k(d)}|\mathbf{Z}_i^{(B)}, \hat{\lambda}_{k(d)}) = \frac{\exp(\hat{\lambda}_{k(d)}^T \mathbf{Z}_i)}{1 + \exp(\hat{\lambda}_{k(d)}^T \mathbf{Z}_i)}. \quad (4.8)$$

Again, assuming conditional independence, the prior probability of examinee  $i$  in a specific latent class  $\pi_i^{(1)}(\alpha_{I(d)})$  can be calculated using Equation 4.7.

## 4.5 Simulation Studies

Two simulation studies were conducted to examine the feasibility of the proposed approaches to implement CD-CAT for high-dimensional scenarios, and the extent to which incorporating covariates in CD-CAT to estimate examinees' prior distributions can further improve its efficiency. The two studies corresponded to the aforementioned cases 1 and 2: simulation study 1 was designed to compare the different ways of using previous response data and covariates to estimate the examinees' prior distributions, whereas simulation study 2 aimed to compare the use of flat prior distributions and distributions estimated from the covariates when no previous response data exist. In both studies, AP was used as the calibration method, and the M-GDI was used as the item selection method.

### 4.5.1 Simulation Study 1

#### *Design*

Four approaches to estimate the examinees' prior distributions were compared in simulation study 1, namely, 1) estimation based on regressing the true  $\alpha_{k(d)}$  onto the covariates, denoted by True; 2) estimation based on regressing the estimated  $\hat{\alpha}_{k(d)}$  with correction weights, denoted by C; 3) estimation based on regressing the estimated  $\hat{\alpha}_{k(d)}$  without correction weights, denoted by UC; and 4) estimation based on previous response data only, denoted by RO. Three approaches, True, C and UC, incorporated both  $\mathbf{Y}^{(0)}$  and  $\mathbf{Z}^{(0)}$ ,

whereas, RO depended on  $\mathbf{Y}^{(0)}$  only. Despite different information incorporated, none of the methods started CD-CAT with a flat prior distribution.

Six factors considered in simulation study 1 were the number of domains ( $D = 2$  and 4), number of attributes per domain ( $K(d) = 5$  and 8), item pool size (100 and 400), item quality (low, medium, and high), generating model (G-DINA and DINA), and minimax posterior termination rule (.4, .5, .6, .7 and .8). The total number of attributes considered in this study were equal to  $K = 10, 16, 20,$  and 32.

The item quality was manipulated by setting the lowest and highest success probabilities, denoted by  $P_0$  and  $P_1$ , respectively. When item quality was low,  $P_0 \sim U(.25, .35)$  and  $P_1 \sim U(.65, .75)$ ; when item quality was medium,  $P_0 \sim U(.15, .25)$  and  $P_1 \sim U(.75, .85)$ ; and when item quality was high,  $P_0 \sim U(.05, .15)$  and  $P_1 \sim U(.85, .95)$ . When the generating model was the G-DINA model, the success probabilities for latent classes other than  $P_0$  and  $P_1$  were generated with monotonicity constraints.

The higher-order (HO; de la Torre & Douglas, 2004) model was utilized to generate correlated attributes and covariates. In this study, the number of covariates was fixed to three. The HO  $\boldsymbol{\theta}$  and covariates  $\mathbf{Z}$  were first drawn from the multivariate normal distribution (MVN):

$$(\boldsymbol{\theta}, \mathbf{Z}) = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3) \sim \mathcal{N}_{D+3}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (4.9)$$

The  $\boldsymbol{\Sigma}$  were designed to reflect a strong association between individual attributes and covariates. For example, the  $\boldsymbol{\Sigma}$  for  $D = 2$  was set to

$$\Sigma = \begin{bmatrix} 1.0 & .40 & .90 & .50 & .20 \\ .40 & 1.0 & .30 & .60 & .80 \\ .90 & .30 & 1.0 & .25 & .25 \\ .50 & .60 & .25 & 1.0 & .25 \\ .20 & .80 & .25 & .25 & 1.0 \end{bmatrix},$$

where the upper-left block-diagonal matrix was the variance-covariance matrix of  $\theta$ ; the bottom-right block-diagonal matrix was the variance-covariance matrix of  $\mathbf{Z}$ ; and the off-diagonal matrix indicated the covariance between  $\theta$  and  $\mathbf{Z}$ . The  $\Sigma$ s for other  $D$  and  $K(d)$  conditions were specified in a similar manner. The above specifications resulted in a McFadden's pseudo  $R^2 \approx .45$  in regressing the true attributes on the covariates. The attribute profiles were generated using the HO model, with the slope parameter fixed to  $a_{k(d)} = 3.5$  and  $b_{k(d)} \sim N(0, .5)$ . The sample size for calibrating item pools using AP were fixed to 10,000 to acquire reliable item parameter estimates. The sample size for examinees to be tested by CD-CAT was set to 2,000.

For case 1, a fixed-length test with  $J = 2 \times K$  items was administered to the examinees in advance of the CD-CAT. The item quality of the fixed-length test was set to be the same as that of items in CD-CAT. Examinees' attribute profiles were assumed to be the same in both fixed-length test and CD-CAT. Examinees who did not satisfy a minimax value were tested using the CD-CAT with their prior distributions estimated using the fixed-length test response data.

The Q-matrices for the fixed-length test and the item pools were constructed satisfying the following requirements. First, all Q-matrices contained at least one identity matrix. Second, each attribute was measured by the same number of items. Third, there were equal number of items measuring different domains. Fourth, the Q-matrices for the item pools with 400 items were obtained by quadrupling the Q-matrices for the item pools with 100 items. Last, the maximum number of attributes measured by an item was three.

To evaluate how informative the prior distributions were estimated by the proposed methods, the proportion of examinees who met a given minimax value by responding to the fixed-length test was recorded. This proportion represented the number of examinees who were accurately classified using the previous tests before using CD-CAT. Examinees who met the the minimax value were exempted from taking CD-CAT. The efficiency of CD-CAT was evaluated by the averaged total test length of examinees when CD-CAT was terminated by a given minimax value.

### *Results*

Figures 4.1 and 4.2 show the proportions of examinees who met the prespecified minimax value in advance of CD-CAT under the two generating models. A few similar patterns were found for both figures. For instance, as shown in Figure 4.1, True and C always determined the largest proportions of examinees met the minimax value, except for  $K(d) = 8$  and low minimax values (i.e., .4) conditions, where RO resulted in the largest proportions. As item quality of the fixed-length test increased, as expected, the proportions increased as well. When a high minimax value was used (i.e., .8), approximately 75% of the sample did not reach the cutoff even when item quality was high, and consequently, CD-CAT was administered to make classification equally reliable for the whole sample.

The patterns found for the G-DINA as the generating model were similar to those found for the DINA model, except that when item quality was low and  $K(d) = 8$ , the proportions of examinees that met each minimax value decreased considerably due to the increased dimensionality.

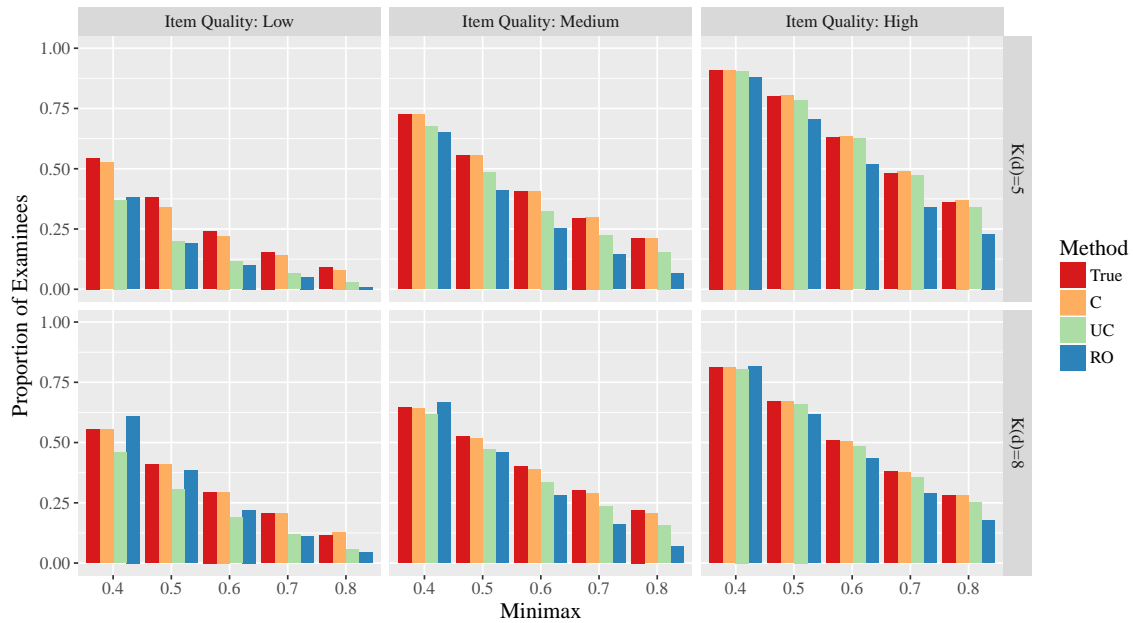


Figure 4.1: Proportion of Examinees Met Minimax Values Prior to CD-CAT,  $D = 2$ , G-DINA

For illustration purposes, the results of test lengths under various minimax values for  $D = 2$  and G-DINA as the generating model are shown in Figures 4.3 and 4.4 for  $K(d) = 5$  and 8, respectively. For  $K(d) = 5$ , the size of item pool had a large impact on the test lengths, except when item quality was high. For instance, when item quality was low, the test lengths determined by the highest minimax value (i.e., .8) were approximately 60 items for True and C, and 70 items for RO and UC. However, when item pool size was quadrupled, the test lengths stopped by the minimax value of .8 for all methods were



below 40. When item quality was medium and item pool size was 100, the test lengths for a minimax value of .8 were below 12.5 items, whereas when the pool size increased to 400, the test lengths decreased to at most 7.5 items. When item quality was high, due to the rich information provided by the fixed-length test, not many items were needed to be administered using CD-CAT. Despite the different item pool sizes, to reach a minimax value of .8, only at most 2 to 3 additional items were required.

As expected, the minimax value had a positive impact on the test lengths: the higher the prespecified precision level, the longer the test length. The increase in test lengths for a relatively small item pool was steeper than the increase found in a larger item pool. This indicated that when item quality was unfavorable, the M-GDI had difficulty in finding the optimal items to shrink the individual posterior distribution, thus an item pool size of 100 for  $K = 10$  might not be sufficient.

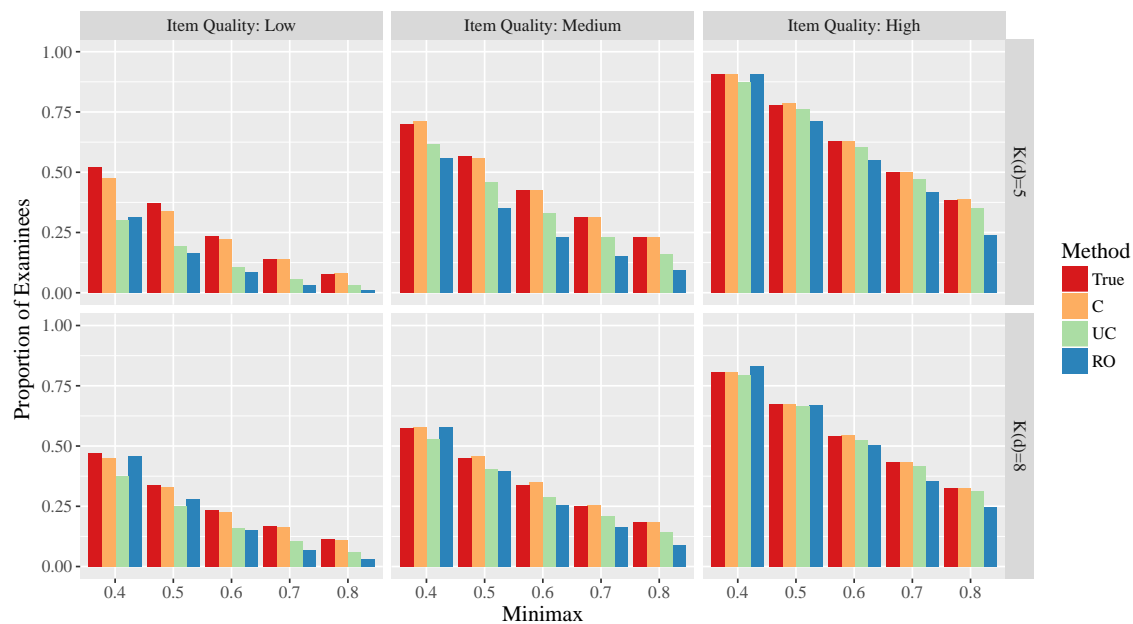


Figure 4.2: Proportion of Examinees Met Minimax Values Prior to CD-CAT,  $D = 2$ , DINA

A comparison among four different methods in Figure 4.3 show that, True and C performed similarly, and they both outperformed UC and RO in terms of test lengths. For instance, when the item pool size was 100, True and C used approximately 20 fewer items to reach a minimax value of .8, compared with UC. As the item quality increased, the differences in test lengths among the four methods became more similar. Interestingly, when item quality was low, UC resulted in a longer test length than RO, suggesting that it was not always safe to incorporate covariates, however informative they were, using a multi-step approach without necessary corrections.

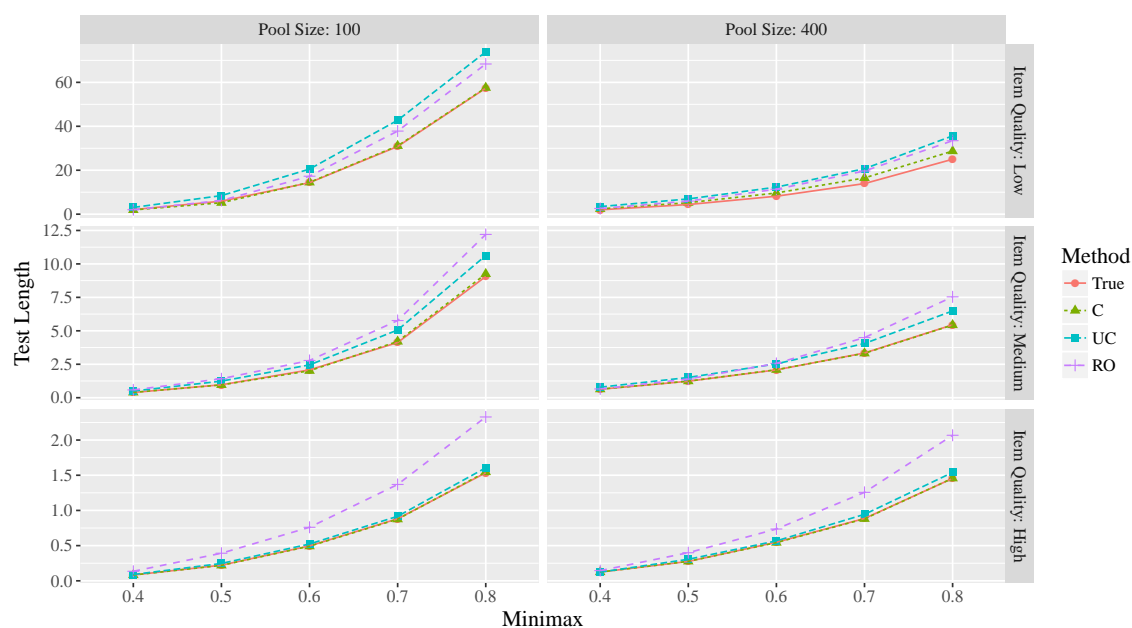


Figure 4.3: Test Length for Case 1,  $D = 2$ ,  $K(d) = 5$ , G-DINA

When  $K(d) = 8$ , the patterns found for  $K(d) = 5$  still held true, except that more items were required to reach the same level of minimax value. For example, when the size of item pool was 100 and item quality was medium, it required 10 more items for all methods

to reach a minimax value of .8, compared with  $K(d) = 5$ .

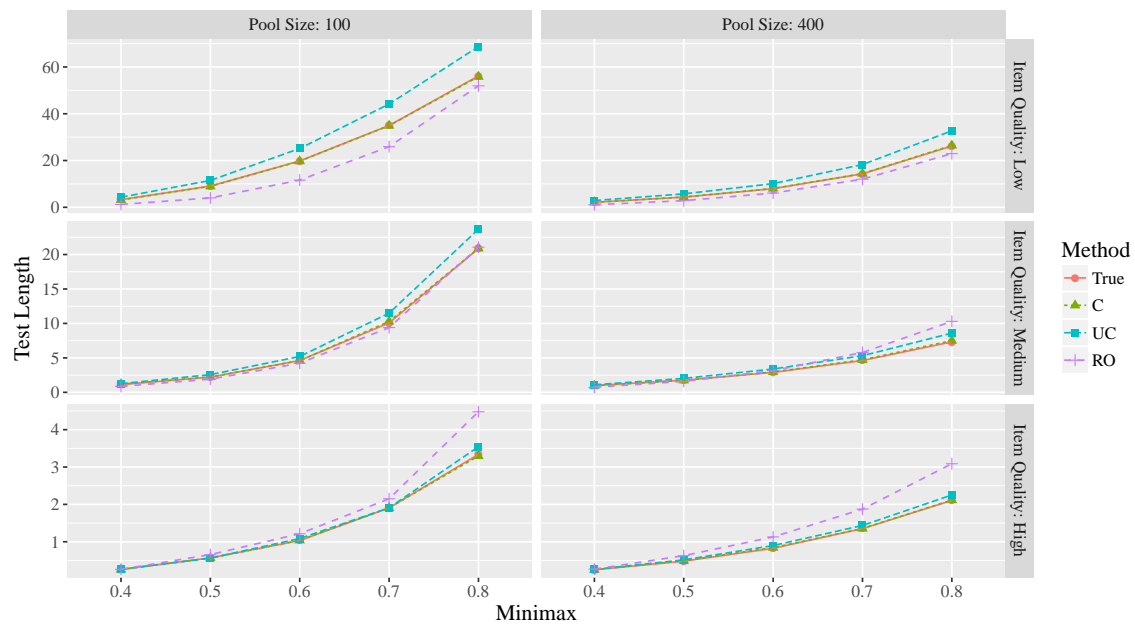


Figure 4.4: Test Length for Case 1,  $D = 2$ ,  $K(d) = 8$ , G-DINA

## 4.5.2 Simulation Study 2

### *Design*

Simulation study 2 was conducted to examine the performance of CD-CAT in high-dimensional testing situations where, unlike simulation study 1, no prior responses were available. To shrink the prior distribution, covariates and coefficients estimated using previous testing data were used to estimate examinees' prior distributions. In this study, four approaches were considered, where the first three approaches were the same as those in simulation study 1, namely, True, C, and UC. True estimated latent regression coefficients by regressing the true attributes onto covariates, whereas C and UC estimated the coefficients

using estimated attributes with or without corrections. A fourth method, denoted by Unif, which did not incorporate covariates and assumed a discrete uniform prior distribution at the beginning of CD-CAT, was added.

The same factors from simulation study 1 were also considered in simulation study 2, namely, the number of domains, number of attributes per domain, item pool size, item quality, data generating model, and minimax values (.4, .5, .6, .7, and .8). Similarly, the item pools, attribute profiles and item responses were generated in the same manner as in the previous simulation study. The test lengths by different minimax values of different methods were compared. Again, 10,000 examinees were used to calibrate the item pools using AP, and to obtain the coefficient estimates, and 2,000 examinees were tested using CD-CAT.

### *Results*

Results for  $D = 2$  and G-DINA as the generating model are presented in Figures 4.5 and 4.6 for  $K(d) = 5$  and 8, respectively. Similar to the findings from simulation study 1, item quality and item pool size had a large impact on test lengths. An extreme case was found when the item pool size was 100 and item quality was low, Unif could not be terminated when the minimax value larger than .6 even all items in a small item pool were selected. When a much larger item pool was available, Unif could be terminated at all minimax values. Moreover, the test lengths of all methods dramatically decreased as the item quality increased. For instance, the test lengths for Unif reduced from 75 to 30 items when the item quality increased from low to medium.

A comparison among four methods showed that using the estimated prior resulted in

much shorter test lengths. For example, as shown in Figure 4.5, when the item quality was low and a high minimax value (i.e., .8) was used, the test lengths for True, C and UC were approximately 30 items, whereas, using a flat prior resulted in an average test length of 75 items. To reduce the test length of Unif to 30 item under the same minimax value of .8, the item quality was required to be increased to at least medium. The test lengths of True, C and UC were close to each other when the item pool was large or item quality was high. In the most unfavorable condition (i.e., small item pool and low item quality), C produced the shortest test lengths among the three methods incorporating covariates. When  $K(d)$  increased to 8, an item pool with 100 items, was not sufficiently large if the item quality was low or a minimax value of greater than .7 was used. Specifically, when the item pool size was 100 and item quality was low, Unif never reached even the the smallest minimax value (i.e., .4). When covariates were incorporated, a 100-item item pool would suffice for minimax values lower than .6.

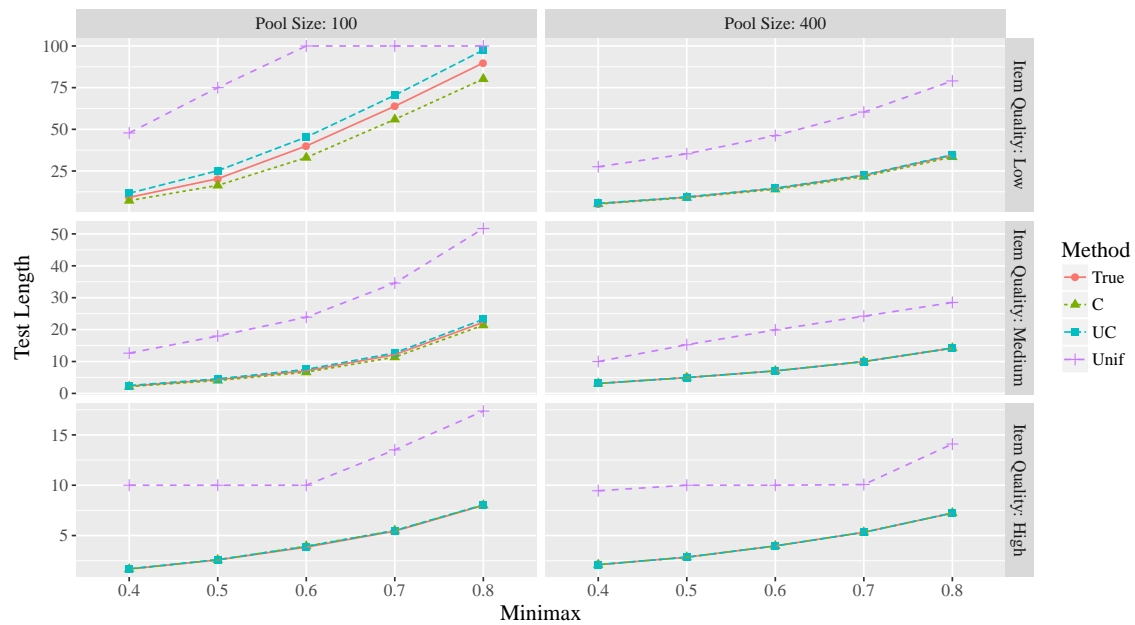


Figure 4.5: Test Length for Case 2,  $D = 2$ ,  $K(d) = 5$ , G-DINA

To better understand how the estimated prior distribution shortened the test length, the shifts of posterior distribution of an examinee with the domain-specific attribute profile  $\alpha = (1, 0, 1, 1, 1)$ , obtained by two methods Unif and C, are given in Figures 4.7 and 4.8. The X-axis represented all 32 possible domain-specific latent classes and the Y-axis the posterior probability. The posterior distribution at each stage was coded in a unique color. The warmth of the color represented the stage of CD-CAT: the warmer the color, the later the stage.

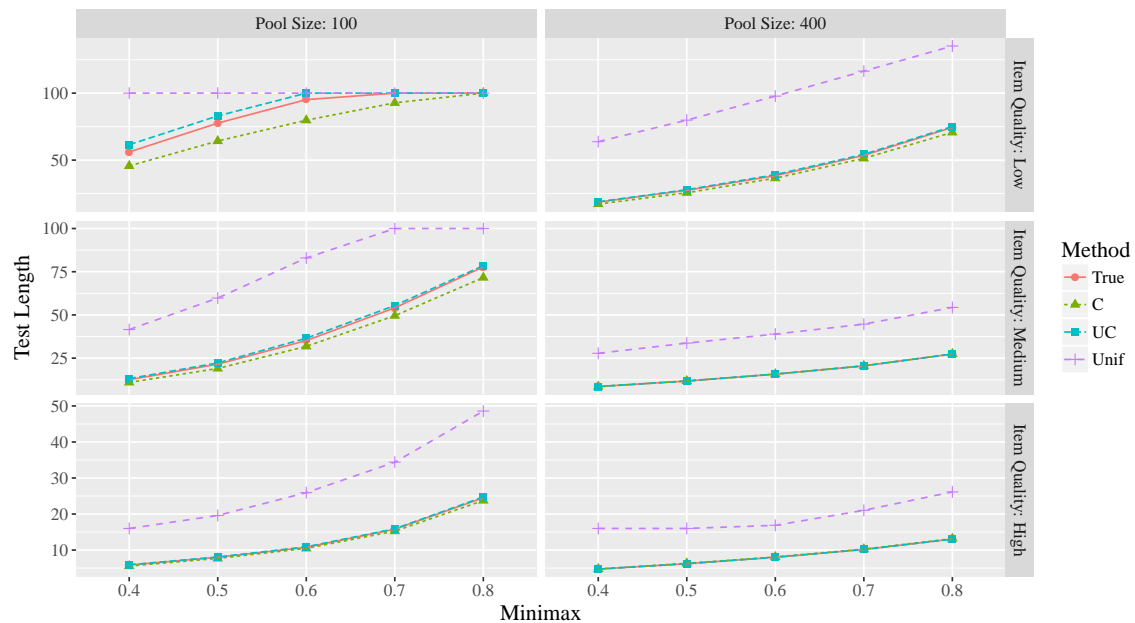


Figure 4.6: Test Length for Case 2,  $D = 2$ ,  $K(d) = 8$ , G-DINA

As can be seen from Figure 4.7, CD-CAT started from a uniform distribution represented by the light blue rectangular box located at the bottom of the figure. In the middle stage of CD-CAT, when 10 items were administered, the posterior distribution was peaking at a wrong latent class, C26 with  $\alpha = (0, 0, 1, 1, 1)$ , which misspecified the second attribute. After approximately 20 items were administered, the posterior distribution peaked at the correct latent class: C30.

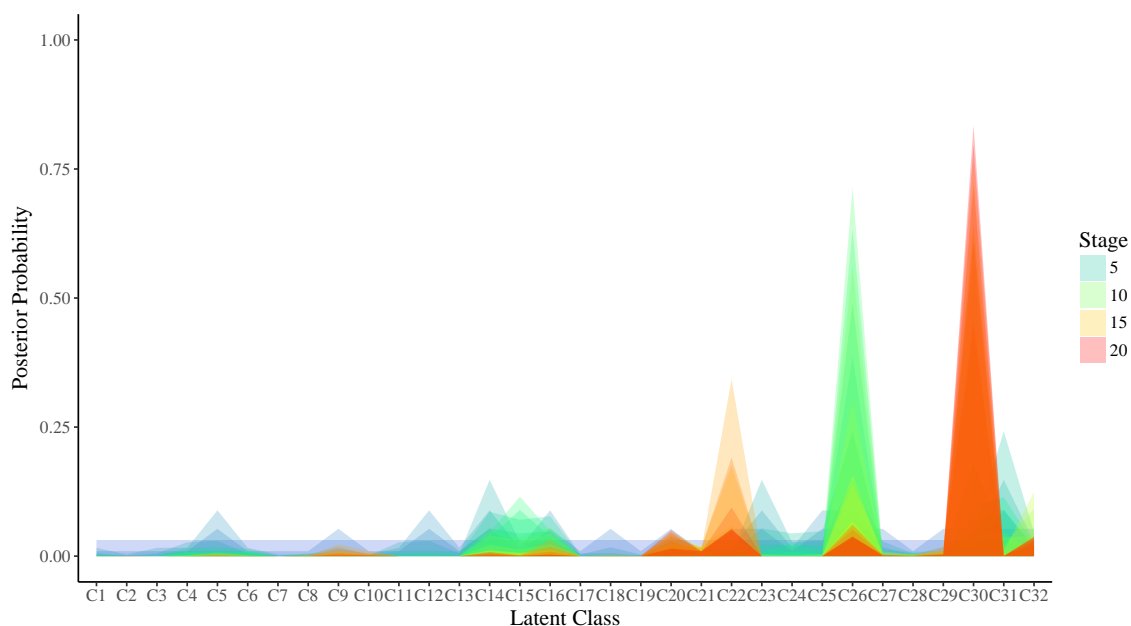


Figure 4.7: Shifts of Posterior Distribution for  $\alpha = (1, 0, 1, 1, 1)$  across Testing Stages: Uniform Prior

In comparison with starting with a uniform prior, as shown in Figure 4.8, the estimated prior started with several possible latent classes, such as, C8 with  $\alpha = (1, 0, 1, 0, 0)$ , C20 with  $\alpha = (1, 0, 1, 1, 0)$  and C30 which was the correct latent class. As the CD-CAT proceeded, the posterior probabilities on the wrong latent classes reduced quickly (shown in green), and eventually the distribution peaked at the correct latent class, C30. In addition, using the estimated prior was less frequently distracted by similar, but incorrect latent classes. For example, Figure 4.7 showed even in the middle stage of CD-CAT, the posterior distribution using a uniform prior still put latent classes: C14, C15 and C16 as possibilities, whereas, the posterior distribution with estimated prior never peaked at those three latent classes.



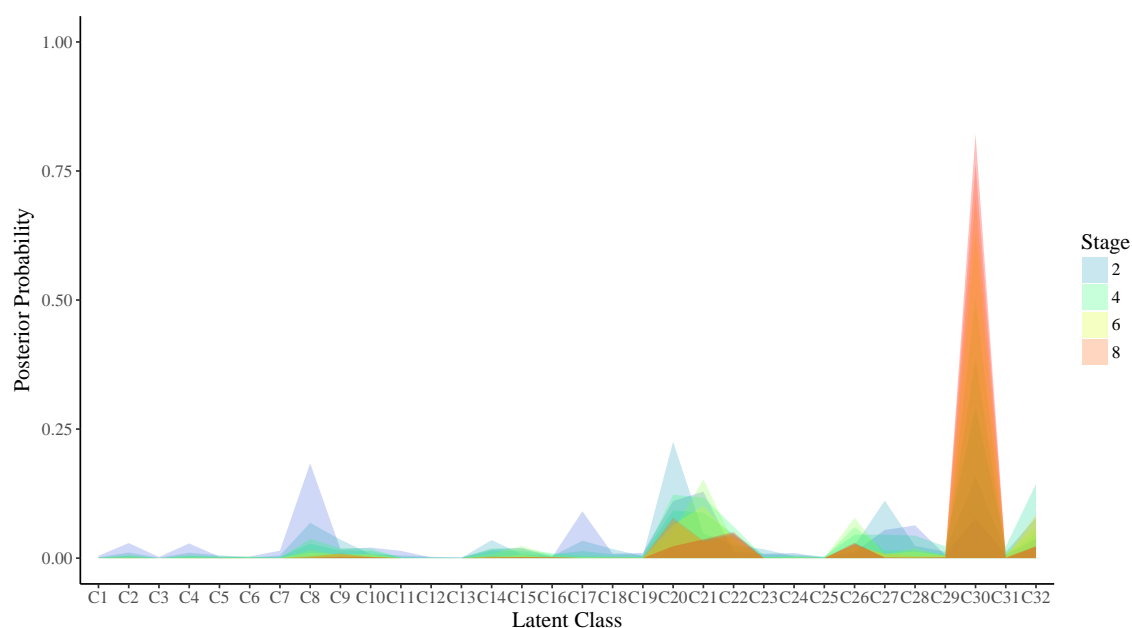


Figure 4.8: Shifts of Posterior Distribution for  $\alpha = (1, 0, 1, 1, 1)$  across Testing Stages: Informative Prior

## 4.6 Discussion

This paper proposed a series of strategies for implementing CD-CAT in high-dimensional situations, specifically, an item pool calibration method, a modified item selection method, and possible approaches to estimate examinees' prior distributions. Two simulation studies were designed to examine the performance of the proposed methods. The results show that AP and M-GDI can be used as calibration method and item selection method to make implementation of CD-CAT in high-dimensional testing situations feasible. Furthermore, when the informative covariates are incorporated to estimate the examinees' prior distributions, the test lengths can be dramatically shortened even under the most unfavorable item quality and item pool size conditions. In contrast, ignoring covariates and using a flat prior

requires larger item pools and higher quality items, which will increase the cost of the item pool development, hence, the practicability of a CD-CAT system will be diminished.

When ideal conditions, such as a sufficiently large item pool or high item quality, are not met, all source of ancillary information should be considered to make the classification accuracy acceptable. This paper considered two cases in which ancillary information can be incorporated at the beginning of CD-CAT. The first approach utilized examinees' previous item responses and covariates to estimate the prior distribution. This approach can be useful when historical data regarding examinees' performance are available. The second approach is applicable when there is a group of new users to a CD-CAT system, for whom no prior performance data is available. The approach estimates the prior distributions using models trained from past testing data. In both approaches, incorporating ancillary information can make testing much more efficient.

Although the results are promising, the current work has a number of limitations. To begin with, the M-GDI only selects the domain-level optimal items because it marginalizes over the nuisance attributes. As such, the M-GDI selects items that only measure the target domain. However, when items are informative on multiple domains, an item selection method should be able to recognize such items for it simultaneously contributes to multiple domains. Further research is needed to determine if a global-optimal item selection method in high-dimensional settings can be found. If so, additional work that look into the relationship between domain-level and global minimax indices would be of interest because it is not clear whether both indices will lead to the same conclusion. As another limitation, this study considers only one item selection method. However, modifications to other methods to work with item pools calibrated using AP are also possible.

More research comparing various item selection methods in this context is therefore warranted. Last, in both simulation studies, the covariates were assumed to be associated with the attributes so that the estimated examinees' prior distributions were guaranteed to be meaningful. This, however, may not be the case in real-world applications. Sometimes researchers may need to consider a much larger set of covariates to achieve a high McFadden's pseudo  $R^2$ . As such, additional practical issues (e.g., multicollinearity, covariate selection, model overfitting) may emerge. The impact of these issues on the estimation of examinees' prior distributions needs to be carefully studied. Furthermore, techniques such as regularization methods, heuristic algorithms for variable selections, and the K-fold cross-validation method may offer solutions to these issues, but their effectiveness in this particular context remains to be determined.

Although CD-CAT is used primarily for low-stake situations (Kaplan et al., 2015), it may also be used in high-stakes contexts. When high-stakes decisions are involved, covariates must be used with caution to ensure that this use does not lead to testing inequities. For example, examinees with poor performance in the past examinations may be unfairly estimated to have low starting proficiency profiles that could influence their subsequent classifications. It is, therefore, imperative that validity evidence supporting the use of certain covariates is collected (AERA, APA, & NCME 2014). Moreover, determining under which situations, if any, differential use of covariates may be reasonable would provide guidance on the use of ancillary information in practical testing situations.

## 4.7 References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.
- Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (in press). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*. doi: 10.1177/0146621618813113
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.
- de la Torre, J. (2017). *The accordion approach: A method for accommodating a large number of attributes in cognitive diagnosis modeling*. Paper presented at the Global Chinese Conference on Educational Information and Assessment - Chinese Association of Psychological Testing Annual Conference, Taichung, Taiwan.
- de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.
- de la Torre, J., & Minchen, N. D. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89–97.
- DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 970–1030). Amsterdam: North-Holland Publications.
- Dynamic Learning Maps Science Consortium. (2015). *Dynamic learning maps essential elements for science*. Lawrence, KS: University of Kansas.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*,

- 191–210.
- Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563–582.
- Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, 15*, 1–7.
- Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models* (Doctoral dissertation). Retrieved from <https://doi.org/doi:10.7282/T3W95D95>.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.
- Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167–188.
- Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an english achievement test in china. *Journal of Classification, 30*.
- Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3–31.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808–821.
- Meijer, R. R., & Nering, M. L. (1999). *Computerized adaptive testing: Overview and introduction*. Sage Publications Sage CA: Thousand Oaks, CA.
- Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics, 10*, 55–73.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

- van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375.
- Wu, H.-M., Kuo, B.-C., & Yang, J.-M. (2012). Evaluating knowledge structure-based adaptive testing algorithms and system development. *Educational Technology & Society*, *15*, 73–88.
- Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*, 291–315.
- Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

## Chapter 5

### Summary

Cognitive diagnosis models (CDMs) have been applied to various contexts, such as, fraction subtraction skills (de la Torre & Douglas, 2004), proportional reasoning skills (Tjoe & de la Torre, 2013, 2014), spatial rotation skills (Culpepper, 2015), and clinical diagnosis (de la Torre, van der Ark, & Rossi, 2018; Templin & Henson, 2006). Most of these CDM applications were developed focusing on one specific target domain with a relatively small set of fine-grained attributes. Nevertheless, because of the specification of the Q-matrix and possible underlying structures between attributes, CDMs have more to offer. By collapsing the attributes in the Q-matrix based upon the underlying attribute structures, the granularity of individual attributes can be adjusted, which consequently reduces the dimensionality in the latent space by collapsing latent classes. Hence, this feature can be utilized to address the CDM estimation issues for high-dimensional data.

With the aim to tackle the issues in applying CDMs to high-dimensional settings, this dissertation proposed: 1) the accordion procedure (AP) to address the estimation issue by focusing on one subset of attributes at a time, and collapsing the attributes of the remaining subsets; 2) the four-step latent regression approach with correction weights in the context of AP to improve the classification accuracy; and 3) a series of CD-CAT strategies, namely, an item pool calibration method, item selection method and prior distribution estimation

method. Proposed in Study I, AP simplified the problem of estimating one large CDM into estimating multiple smaller models. Two simulation studies were designed to evaluate the performance of AP for cross-sectional and longitudinal data. The performance of AP was compared with the complete-profile estimation (CPE) procedure in the cross-sectional data. The results showed that AP achieved comparable classification accuracy, especially at the attribute level, compared with CPE. More importantly, AP was shown to be 10 to 50 times computationally faster than CPE. The results of the longitudinal data showed that AP yielded satisfactory attribute-level classification accuracy while maintained high computational efficiency. Aside from the simulation studies, AP was also fitted to the TIMSS 2007 fourth-grade mathematics data and showed its practicability in real-data.

Study I showed that the classification accuracy suffers when tests were not informative due to short test length or poor item quality. To improve the classification accuracy under unfavorable test conditions, Study II proposed a four-step latent regression procedure to incorporate covariates to supplement the information obtained using CDMs alone in the context of AP. Simulation and real-data studies were conducted to evaluate the performance of the four-step procedure. The results of the simulation study showed that when the item quality was below medium or test length was short, the four-step procedure with informative covariates improved the domain-level classification accuracy compared to the procedure without covariates. To get a deep understanding of how the four-step procedure improved the classification accuracy, the examinees' classification certainty was examined. The results showed the four-step procedure increased the proportion of examinees who were classified with high certainty. Again, the TIMSS 2007 data were analyzed using the four-step procedure, and the classification results were compared to those obtained by



AP and CPE. The results showed that, incorporating covariates provided higher biserial correlations between individual estimated attributes and examinees' overall mathematics scaled scores.

Study III focused on addressing issues in implementation of cognitive diagnosis computerized adaptive testing (CD-CAT) in high-dimensional situations. Utilizing the techniques developed in the previous two studies, AP was proposed as the item pool calibration method, and an item selection method was modified accordingly. The four-step latent regression procedure was used to estimate examinees' prior distributions to make the item selection in the early stage more efficient. Two simulation studies were designed to evaluate the performance of proposed procedures in two scenarios. The first scenario assumed that examinees' previous response data and covariates were available, and the second assumed that latent regression models trained from past testing data were available. Results showed that, incorporating covariates to estimate the examinees' prior distributions always resulted in better testing efficiency in terms of test lengths than ignoring the covariates. A further investigation on the shifts of an examinee's posterior distributions across different testing stages showed that the estimated prior distribution shrank faster to the correct latent class than using a flat prior.

A few limitations of this dissertation are worth mentioning. First, in Study I, all simulated Q-matrices are complete (Köhn & Chiu, 2017), which may not be the case in practice. When the number of attributes is large, constructing a Q-matrix with an identity matrix may be difficult. Further research is needed to investigate how incomplete Q-matrices will affect the classification accuracy of AP. Second, a simple set of covariates that are highly

correlated with attributes are generated in the simulation studies of Study II and III. In reality, practical issues, such as, multicollinearity or model overfitting may emerge. Further research is needed to evaluate the impact of these issues on the classification results and the estimated posterior probability distributions.

As the first studies that addressed the practical issues in fitting CDMs in high-dimensional scenarios, this dissertation broadens the practicability of CDMs in testing situations where diagnosing a relatively large set of skills across multiple content domains is of interest. Aside from developing cognitively diagnostic assessments (CDAs) in a traditional way where skills from a relatively small domain are focused, AP opens the possibility of using CDMs in other testing designs. For example, one possibility is to develop large-scale CDAs measuring a full set of skills students learned in a semester or school year. The comparisons of students' mastery profiles at the state or country-level may be meaningful to shed light on the impact of instructions or educational standards on student learning. Moreover, CD-CAT with AP can be used to evaluate students learning status in a learning map where the goal for each student is to acquire all the necessary skills defined by certain academic standards (e.g., Dynamic Learning Maps Science Consortium, 2015). At different learning stages, AP can be used to collapse uninformative attributes as shown in Study I.

## 5.1 References

- Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*, 454–476.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.
- de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51*, 281–296.
- Dynamic Learning Maps Science Consortium. (2015). *Dynamic learning maps essential elements for science*. Lawrence, KS: University of Kansas.
- Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 112–132.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.
- Tjoe, H., & de la Torre, J. (2013). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education, 6*, 17–26.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237–255.