認知診断モデルにおけるモデル選択の比較

―シミュレーションによる小サンプル状況下での検証―

山口一大（東京大学）

Some Model Comparisons in Cognitive Diagnostic Models

Relatively Small Sample Situation with Simulation Study

Kazuhiro Yamaguchi

Graduate School of Education, the University of Tokyo

Author Note

Kazuhiro Yamaguchi, Graduate School of Education, the University of Tokyo, Tokyo, Japan

Correspondence concerning this article should be addressed to Kazuhiro Yamaguchi, Graduate School of Education, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan; E-mail: kazz530@p.u-tokyo.ac.jp

This research was supported by a grant, Youth Scholar Program from Center for Excellence in School Education, Graduate School of Education, The University of Tokyo.

Abstract

A lot of cognitive diagnostic models (CDMs) have been developed in several decades. The objective of this study is to check how we can detect misspesifications among data generation models and analysis models in relatively small sample size situations. We employed simulation study for the purpose. We got three results. First, that Bayesian information criterion (BIC) indicated LLM (linear logistic model) as optimal model when G-DINA (generalized deterministic noisy inputs “and” gate) model was true model. Second, when the LLM and A-CDM (additive CDM) were true models, it was difficult to distinguish these model with Akaike information criterion (AIC) and BIC. Third, AIC and BIC can select R-RUM (reparametarized reduced unified model), DINA (deterministic noisy inputs “and” gate) model and DINO (deterministic noisy inputs “or” gate) model models as correct model. We discuss these results.

Keywords: Cognitive Diagnostic Models, Model Comparison, Simulation Study

認知診断モデルにおけるモデル選択の比較

シミュレーションによる小サンプル状況下での検証

1 問題と目的

認知診断モデル（CDM, cognitive diagnostic models; Leignton & Giear, 2007; Rupp, Templin, & Henson, 2010; 山口・岡田，2017b）は学力テストから学習者の知識状態を推定し，学習上のつまづきについての情報を得ることができる有用なモデルとして近年盛んに研究されている。従来のテストの分析ツールとしては古典的テスト理論や項目反応理論（IRT, item response theory; Embretson & Raise, 2000）といった理論が整備されてきており，コンピュータアダプティブテストや，等質なテスト構成を可能にするなど，テストの品質を統計的に評価するための非常に重要な理論として発展してきた。

IRT は単一の次元上に学習者を順序づけることや，異なったテストを受験した受験者を，一定の条件のもと比較可能にすることが重要な点である。しかし，その一方でテストから得られる学習者にとっての有用な情報は限られている。つまり，他の学習者との相対的な潜在的特性の高低が中心となり，学習に有用な情報を得ることが必ずしも容易ではない点が問題としてあげられる。近年，能力を一次元的に測定する評価から学習者の診断的・形成的評価へ評価方法が変化していることが指摘されており（植野・荘島, 2010），テストそれ自体を学習に利用するための方法論の整備が必要といえよう。

CDM は IRT とは異なり，テストから学習者にとって有用な情報を抽出できる。例えば，分数の計算において「通分」がといった学習要素の習得が不十分であれば，その要素を習得するために学習時間を利用するという判断が可能である。学習内容が膨大になっている現代において，有限の学習時間をどのように配分するのかという判断は効率的な学習を行ううえで重要であろう。また，CDM を用いたテストからは学習者のみならず，教師にとっても有用な情報を得ることができると考えられる。例えば，クラス内で特定の学習要素の習得が不十分である子どもの割合がわかれば，それによってカリキュラムや教授内容の修正を計ることも可能となろう。CDM によって，複数の学習要素の習得の有無の組合せに関する客観的な情報を得ることが可能となる。また，学校で実施される小テストも学習者の状態を理解するために有用と考えられるが，CDM を利用することでそうした小テストの解答に必要な複数の能力の組合せについての情報を得ることができる。このような，能力の組み合わせを素点から理解するのは容易では無いだろう。ただし，こういった CDM の利点は IRT の有用性を否定するものではない。CDM と IRT はテストの目的に合わせて使い分けるべきものであり，どちらの理論が優れているのかという議論それ自体は意味をなさない点には注意が必要である。

CDM は個人に適した学習を可能にし，多様な学習を支援する強力なモデルであると考えられる。しかしながら，実際に CDM を利用する学校のテスト場面を想定すると，CDM の利用に際しての問題が生じる。とくに，後述するように，

CDM には数多くのモデルが内包されており，度のモデルが実際のテストに適しているのか，CDM 利用者は事前に判断するのは難しい。学習要素の診断テストにおいては，どのような学習要素をどのように組み合わせると問題が解答可能になるのかを事前に詳細に定めた仕様書を作成し，その仕様書にもとづいてテスト項目を開発することが望ましい。しかし，現実的には予め仮定したような解答反応を学習者が行うとは限らず，実際のテストを用いた項目反応モデルの選択の必要性が生じる。

1.1 多様な認知診断モデルの比較

CDM には非常に多くのモデルが包含され，モデル利用者は多様なモデルの中から適切なモデルの選択を迫られる。CDM の選択においては，実データによるモデル比較研究の蓄積が近年なされてきている（e.g., Li, Hunter, & Lei, 2016; 鈴木・豊田・山口・孫, 2015; 山口・岡田, 2017a）。例えば，山口・岡田 (2017a)では TIMSS 2007 の 4 年生の算数データを用いて，G-DINA モデル（generalized deterministic noisy “and” gate model; de la Torre, 2011）に包含されるモデル群を対象にモデル比較を行なった。具体的なモデルとしては，G-DINA モデル，R-RUM（re-parametarized reduced unified model; Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007），LLM（linear logistic model; Maris, 1999），A-CDM（additive CDM; de la Torre, 2011），DINA モデル（deterministic input noisy “and” gate model; Junker & Sitsuma, 2001），および DINO モデル（deterministic noisy “or” gate model; Templin & Henson, 2006）といったモデルを比較対象とし，とくに TIMSS のサンプルサイズ 639 の日本人データにおいては AIC, BIC の観点から R-RUM モデルの適合がよいということを示した。鈴木・豊田・山口・孫（2015）は教研式標準学力検査 NRT「中学 1 年数学」の 948 名の中学生の解答データに対して G-DINA モデル, A-CDM, DINA モデル，DINO モデルを適用し，AIC の観点から G-DINA モデル，BIC・CAIC の観点から A-CDM が支持されたことを報告している。さらに，Li et al.（2016）は英語のテストであるミシガン英語アセスメントバッテリー（Michigan English Language Assessment Battery, MELAB）のサンプルサイズ 2019 解答データを G-DINA モデル，A-CDM，LLM，R-RUM，DINA モデル，DINO モデルを比較し，AIC の観点からは G-DINA モデル, BIC の観点からは A-CDM が選択された。Chen and de la Torre（2014）はイギリスの 2012 人の PISA 2000 のリーディングアセスメントを用いて，山口・岡田（2017a）と同様のモデル比較を行い，AIC，BIC の観点から LLM が選択されることを示した。

上記のように実データを用いたモデル比較研究がなされてきおり，現実的なデータにどのようなモデルが適合するのか経験的な知見が蓄積されつつある。

1.2 目的

経験的な CDM のモデル選択研究は非常に重要である一方，実データでは真の項目反応関数はわからない。このため，モデル比較に一般的に利用される情報量規準（例えば AIC，BIC）といった指標がどれほど CDM のモデル選択において有効であるのか，明らかでない。

上記の問題を受けて，本研究ではシミュレーションを用いて，モデル選択に利用される種々の指標を評価し，データ生成モデルと解析モデルに乖離が生じている場合にその乖離をどの程

度検出できるのか，シミュレーション研究によって検証することを目的とする。先行研究では PISA 調査など比較的大規模なテストに CDM を適用しているが， CDM が本来利用されるべき状況は，１つの学級やクラスなど，比較的小さいサンプルサイズが想定される。この点を考慮して，本研究では比較的小さいサンプルサイズでの検証を行うことによって，現実的な状況に近い場合を想定する。

2 方法

本章では本研究に利用した CDM の定式化および，モデル比較に使用した指標（とくに情報量規準）を示した後に，実際のシミュレーションの手続きと評価方法をしめす。なお，CDM の定式化の詳細に関しては，山口・岡田（2017b）が日本語の文献としては詳しい。

2.1 認知診断モデルの定式化

現在提案されているものの中で最も高い表現力を持つモデルは G-DINA モデル（de la Torre, 2011）である。本研究では，G-DINA モデルおよびその下位モデルとして表現可能な，A-CDM（de la Torre, 2011），LLM（Maris, 1999），R-RUM（Roussos et al, 2007），DINA モデル（Junker & Sitjsma, 2001），そして DINO モデル（Templin & Henson, 2006）の 6 種類のモデルをデータ生成および解析モデルとして用いる。

各モデルの概略を示す。G-DINA モデルは正答確率をモデル化する際に，問題の解答に必要な能力であるアトリビュート複数の交互作用効果を考える点が特徴のモデルである。A-CDM は各問題項目の正答に必要なアトリビュートが加法的に項目の正答確率に影響を与えるモデルである。LLM は A-CDM と同様にアトリビュートが加法的に影響を与えるが，パラメタリゼーションとしては項目反応理論のようにロジット変換を行うことで線形になるモデルである。R-RUM はこれまでのモデルとは逆に，個別のアトリビュートを習得していないことによって正答率が積算的に減少するモデルである。DINA モデルは問題に必要なアトリビュートを全て習得していないと問題への正答確率が高くならないと仮定するモデルである。DINO モデルは逆に問題の正答に必要なアトリビュートを１つ以上習得している場合に正答確率が高くなることを仮定するモデルである。

次に，形式的にそれらのモデルを表現する。まず，$j(=1,\cdots,J)$ を，受験者を区別する添字とする。さらに，アトリビュート番号を $k(=1,\cdots,K)$ とする。本研究ではアトリビュート習得パタンの添字は $l(=1,\cdots,2^K)$ とした。これは，アトリビュート習得パタンに特に制約を掛けず，全ての習得パタンを推定することを意味している。

この G-DINA モデルを

$$\begin{aligned} \mathrm{P}_j(\alpha_{lj}^*) = \Pr(X_j = 1|\alpha_{lj}^*) &= \delta_{j0} \\ &+ \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \\ &+ \cdots \\ + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} &+ \cdots \\ \cdot + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} & \end{aligned} \quad (1$$

と定義する。K_j^* は項目 j に必要なアトリビュート数である。α_{lk} は l 番目のアトリビュート習得パタンの k 番目の要素を表しており，0・1 の 2 値を取る。0 は未習得，1 は習得に対応する。α_{lj}^*

は項目jで区別できるアトリビュート習得パタンlである。さらに，δ_{j0}は切片パラメタ，δ_{jk}はアトリビュートkを習得している場合の主効果パラメタ，$\delta_{jkk'}$はアトリビュートk,k'を習得している場合の交互作用効果パラメタ，$\delta_{j12\cdots K_j^*}$は最高次の交互作用項を表している。G-DINA モデルではさらに一般化線形モデルのようにリンク関数を指定することもできる。

G-DINA モデルの交互作用効果を 0 と制約することで A-CDM が得られ，さらにリンク関数として logit を選択すると LLM が得られる。R-RUM モデルは log リンク関数を選択し，多少の式変形を行うと R-RUM モデルが得られる。切片と最高次の交互作用のみを推定すると DINA モデルが表現できる。DINO モデルの制約は若干複雑であるが，１つ以上アトリビュートを習得している場合正答確率が上昇するような制約を課すことができる。山口・岡田（2017b）などを参考にこれらのモデルの特徴をまとめると Table 1 となる。

Table 1
Summary of The Models Used in This Study

Model	Compensatoy/Noncompensatory	Three Types	Link Function
G-DINA	Sturate	Saturate	Identity
R-RUM	Noncompensatory	Main effects	log
LLM	Compensatory	Main effects	logit
A-CDM	Compensatory	Main effects	Identity
DINA	Noncompensatory	Parsimonious	Identity
DINO	Compensatory	Parsimonious	Identity

2.2 情報量規準

本節では統計科学一般で利用されている情報量規準について述べる。日本語の文献に関しては小西・北川（2004）が情報量規準についての記述に関して詳しい。本研究では Deviance, AIC（Akaike information critetion; Akaike, 1974），BIC（Bayesian information criterion; Schwarz, 1978）の３つを用いる。まず，Deviance を

$$Deviance = -2logLik \quad (2$$

と定義する。$loglik$は最大化された対数尤度であり，-2 倍することで，小さいほどモデルの適合を示す指標となる。
また，AIC を

$$AIC = Deviance + 2 \times \#parametar \quad (3$$

とする。$\#parametar$推定されたパラメタ数である。最後に，BIC を

$$BIC = Deviance + \#parametar \log(samplesize) \quad (4$$

とする。ここで示した情報量規準に関しては，最尤推定法（MLE, maximum likelihood estimation）による方法であり，いずれも相対指標とされ，絶対的な値の解釈はできない。また，いずれの指標も値が小さいほど適合がよいことを示すものである。

2.3 シミュレーションの手続き

本研究ではサンプルサイズ条件は 200，400，600，800 の 4 条件を設定した。データ生成モデル，分析モデルには G-DINA モデル，R-RUM，LLM，A-CDM，DINA モデル，DINO モデルを用いた。データ生成に際して，真の項目パラメタを設定する必要があるが，DINA モデル，DINO モデルの slip, guessing パラメタを全ての項目で.2 に固定した。このとき，他の制約がゆるいモデルでは，当該の項目に必要なアトリビュートを全て習得していない場合（つまり切片パラメタ）を.2 とし，全てのアトリビュートを習得している場合の正答確率が.8（つまり当該項目のアトリビュートを全て習得している解答者のうち 2 割は誤答する）ように設定した。また，アトリビュート習得パタンは一様乱数から生成した。データ生成には GDINA パッケージ（Ma & de la Torre, 2017）の simGDINA 関数を利用した。

シミュレーションでは，特定のサンプルサイズ，データ生成モデルを設定したもとで，項目反応を生成し，1 つのデータに対して 6 種類の分析モデルを適用した。個別のモデル推定に際しては，初期値を変えて 10 回推定をし，もっとも尤度が高くなった条件を採用した。推定には GDINA 関数を利用した。

Q 行列に関しては，その設定によって項目パラメタの識別性（identifiability）の問題が存在することが指摘されている（Xu & Zhang, 2016）。本研究では，Xu and Zhang (2016)が示した DINA モデルにおける識別性の条件を満たすように設定した Q 行列を作成し，全ての条件で共通に利用した（Table 2）。具体的な識別性の条件については Xu and Zhang (2016)を参考にした。

各サンプルサイズ条件，データ生成条件ごとにシミュレーションは 100 回行った。その 100 回の推定結果を元に，情報量規準をもとに平均順位と標準偏差を算出した。Deviance, 情報量規準はすべて値が小さいほど適切であることを示している指標であるため，もっとも小さい値から順に 1 位，2 位と順位をつけた。

Table 2
Q-matrix Used in the Simulation Study

	# Attribute						# Attribute				
# Item	1	2	3	4	5	# Item	1	2	3	4	5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

3 結果

データ生成モデルが G-DINA モデル，R-RUM，LLM，A-CDM，DINA モデル，DINO モデルのそれぞれ場合におけるサンプルサイズ条件分析モデル別の Deviance，AIC，BIC の平均順位および標準偏差を Table 3~8 に示した。また，Deviance，AIC，BIC のそれぞれの観点で最もよい適合を示したモデルのセルに色をつけた。平均順位が最も低いセルに G-DINA モデルは飽和モデルであり，最も Deviance が低いことは理論的に整合している。また，本研究での設定でのパラメタ数は，G-DINA モデルが 175, R-RUM,

LLM，A-CDM が 124 で， DINA モデルおよび DINO モデルが 93 であった。

Table 3 から，G-DINA モデルが生成モデルである時, AIC は一貫して真のモデルを選択した。しかし，サンプルサイズが 200 の場合の平均順位は 1.370 であり必ずしも新のモデルが選択されない傾向がみられる。しかし，サンプルサイズが 600 以上では平均順位が 1.000 であり，今回のシミュレーション条件では正しいモデルを選択出来ることが示された。一方，G-DINA モデルが真のモデルの場合，BIC の観点からは常に G-DINA モデルは選択されない傾向が示され

Table 3
Mean Rank and SD When Data Gatagenerate Model was G-DINA

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	1.370	(0.800)	5.960	(0.197)
	R-RUM	3.950	(0.219)	3.900	(0.302)	3.340	(0.590)
	LLM	2.000	(0.000)	1.780	(0.416)	1.160	(0.368)
	A-CDM	3.050	(0.219)	2.950	(0.386)	2.320	(0.510)
	DINA	5.890	(0.314)	5.890	(0.314)	4.900	(0.461)
	DINO	5.110	(0.314)	5.110	(0.314)	3.320	(1.262)
400	G-DINA	1.000	(0.000)	1.060	(0.278)	4.360	(0.503)
	R-RUM	3.920	(0.273)	3.920	(0.273)	2.950	(0.330)
	LLM	2.000	(0.000)	1.950	(0.219)	1.000	(0.000)
	A-CDM	3.080	(0.273)	3.070	(0.293)	2.080	(0.273)
	DINA	5.900	(0.302)	5.900	(0.302)	5.880	(0.356)
	DINO	5.100	(0.302)	5.100	(0.302)	4.730	(0.679)
600	G-DINA	1.000	(0.000)	1.000	(0.000)	3.930	(0.293)
	R-RUM	3.950	(0.219)	3.950	(0.219)	3.010	(0.333)
	LLM	2.000	(0.000)	2.000	(0.000)	1.000	(0.000)
	A-CDM	3.050	(0.219)	3.050	(0.219)	2.060	(0.239)
	DINA	5.910	(0.288)	5.910	(0.288)	5.910	(0.288)
	DINO	5.090	(0.288)	5.090	(0.288)	5.090	(0.288)
800	G-DINA	1.000	(0.000)	1.000	(0.000)	3.500	(0.704)
	R-RUM	3.950	(0.219)	3.950	(0.219)	3.330	(0.570)
	LLM	2.000	(0.000)	2.000	(0.000)	1.000	(0.000)
	A-CDM	3.050	(0.219)	3.050	(0.219)	2.170	(0.378)
	DINA	5.930	(0.256)	5.930	(0.256)	5.930	(0.256)
	DINO	5.070	(0.256)	5.070	(0.256)	5.070	(0.256)

Table 4
Mean Rank and SD When Data Gatagenerate Model was R-RUM

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	3.690	(0.677)	5.350	(0.730)
	R-RUM	2.190	(0.545)	1.200	(0.550)	1.190	(0.545)
	LLM	3.010	(0.460)	2.120	(0.624)	2.010	(0.460)
	A-CDM	3.800	(0.449)	2.990	(0.611)	2.800	(0.449)
	DINA	5.030	(0.171)	5.030	(0.171)	4.190	(0.419)
	DINO	5.970	(0.171)	5.970	(0.171)	5.460	(0.558)
400	G-DINA	1.000	(0.000)	3.610	(0.634)	4.030	(0.171)
	R-RUM	2.010	(0.100)	1.010	(0.100)	1.010	(0.100)
	LLM	3.000	(0.142)	2.080	(0.307)	2.000	(0.142)
	A-CDM	3.990	(0.100)	3.300	(0.482)	2.990	(0.100)
	DINA	5.000	(0.000)	5.000	(0.000)	4.970	(0.171)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
600	G-DINA	1.000	(0.000)	3.290	(0.782)	4.000	(0.000)
	R-RUM	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
	LLM	3.000	(0.000)	2.200	(0.402)	2.000	(0.000)
	A-CDM	4.000	(0.000)	3.510	(0.502)	3.000	(0.000)
	DINA	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
800	G-DINA	1.000	(0.000)	2.770	(0.750)	4.000	(0.000)
	R-RUM	2.010	(0.100)	1.010	(0.100)	1.010	(0.100)
	LLM	2.990	(0.100)	2.410	(0.514)	1.990	(0.100)
	A-CDM	4.000	(0.000)	3.810	(0.394)	3.000	(0.000)
	DINA	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)

Table 5
Mean Rank and SD When Data Gatagenerate Model was LLM

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	3.890	(0.469)	5.340	(0.831)
	R-RUM	3.480	(0.835)	2.540	(0.858)	2.480	(0.835)
	LLM	2.910	(0.683)	1.940	(0.736)	1.910	(0.683)
	A-CDM	2.610	(0.680)	1.630	(0.720)	1.610	(0.680)
	DINA	5.430	(0.498)	5.430	(0.498)	4.770	(0.763)
	DINO	5.570	(0.498)	5.570	(0.498)	4.890	(0.751)
400	G-DINA	1.000	(0.000)	3.950	(0.219)	4.000	(0.000)
	R-RUM	3.830	(0.551)	2.870	(0.597)	2.830	(0.551)
	LLM	2.640	(0.523)	1.650	(0.557)	1.640	(0.523)
	A-CDM	2.530	(0.627)	1.530	(0.627)	1.530	(0.627)
	DINA	5.380	(0.488)	5.380	(0.488)	5.380	(0.488)
	DINO	5.620	(0.488)	5.620	(0.488)	5.620	(0.488)
600	G-DINA	1.000	(0.000)	3.970	(0.171)	4.000	(0.000)
	R-RUM	3.890	(0.424)	2.920	(0.464)	2.890	(0.424)
	LLM	2.580	(0.589)	1.580	(0.589)	1.580	(0.589)
	A-CDM	2.530	(0.540)	1.530	(0.540)	1.530	(0.540)
	DINA	5.260	(0.441)	5.260	(0.441)	5.260	(0.441)
	DINO	5.740	(0.441)	5.740	(0.441)	5.740	(0.441)
800	G-DINA	1.000	(0.000)	3.880	(0.327)	4.000	(0.000)
	R-RUM	3.940	(0.343)	3.060	(0.489)	2.940	(0.343)
	LLM	2.550	(0.520)	1.550	(0.520)	1.550	(0.520)
	A-CDM	2.510	(0.541)	1.510	(0.541)	1.510	(0.541)
	DINA	5.270	(0.446)	5.270	(0.446)	5.270	(0.446)
	DINO	5.730	(0.446)	5.730	(0.446)	5.730	(0.446)

Table 6
Mean Rank and SD When Data Gatagenerate Model was A-CDM

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	3.850	(0.500)	5.360	(0.835)
	R-RUM	3.610	(0.737)	2.720	(0.817)	2.610	(0.737)
	LLM	2.840	(0.647)	1.860	(0.682)	1.840	(0.647)
	A-CDM	2.550	(0.672)	1.570	(0.700)	1.550	(0.672)
	DINA	5.510	(0.502)	5.510	(0.502)	4.840	(0.721)
	DINO	5.490	(0.502)	5.490	(0.502)	4.800	(0.778)
400	G-DINA	1.000	(0.000)	3.940	(0.239)	4.000	(0.000)
	R-RUM	3.900	(0.414)	2.960	(0.491)	2.900	(0.414)
	LLM	2.550	(0.539)	1.550	(0.539)	1.550	(0.539)
	A-CDM	2.550	(0.575)	1.550	(0.575)	1.550	(0.575)
	DINA	5.520	(0.502)	5.520	(0.502)	5.520	(0.502)
	DINO	5.480	(0.502)	5.480	(0.502)	5.480	(0.502)
600	G-DINA	1.000	(0.000)	3.910	(0.288)	4.000	(0.000)
	R-RUM	3.960	(0.281)	3.050	(0.411)	2.960	(0.281)
	LLM	2.560	(0.519)	1.560	(0.519)	1.560	(0.519)
	A-CDM	2.480	(0.522)	1.480	(0.522)	1.480	(0.522)
	DINA	5.370	(0.485)	5.370	(0.485)	5.370	(0.485)
	DINO	5.630	(0.485)	5.630	(0.485)	5.630	(0.485)
800	G-DINA	1.000	(0.000)	3.730	(0.446)	4.000	(0.000)
	R-RUM	4.000	(0.000)	3.270	(0.446)	3.000	(0.000)
	LLM	2.550	(0.500)	1.550	(0.500)	1.550	(0.500)
	A-CDM	2.450	(0.500)	1.450	(0.500)	1.450	(0.500)
	DINA	5.440	(0.499)	5.440	(0.499)	5.440	(0.499)
	DINO	5.560	(0.499)	5.560	(0.499)	5.560	(0.499)

Table 7
Mean Rank and SD When Data Gatagenerate Model was DINA

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	2.000	(0.000)	3.870	(0.630)
	R-RUM	3.000	(0.000)	3.000	(0.000)	2.010	(0.100)
	LLM	4.000	(0.000)	4.000	(0.000)	3.250	(0.435)
	A-CDM	5.000	(0.000)	5.000	(0.000)	4.880	(0.356)
	DINA	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	5.990	(0.100)
400	G-DINA	1.000	(0.000)	2.000	(0.000)	2.570	(0.498)
	R-RUM	3.000	(0.000)	3.000	(0.000)	2.430	(0.498)
	LLM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	A-CDM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	DINA	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
600	G-DINA	1.000	(0.000)	2.000	(0.000)	2.000	(0.000)
	R-RUM	3.000	(0.000)	3.000	(0.000)	3.000	(0.000)
	LLM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	A-CDM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	DINA	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
800	G-DINA	1.000	(0.000)	2.000	(0.000)	2.000	(0.000)
	R-RUM	3.000	(0.000)	3.000	(0.000)	3.000	(0.000)
	LLM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	A-CDM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	DINA	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
	DINO	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)

Table 8
Mean Rank and SD When Data Gatagenerate Model was DINO

# Sample	Analysis model	Deviance Mean	Deviance SD	AIC Mean	AIC SD	BIC Mean	BIC SD
200	G-DINA	1.000	(0.000)	2.000	(0.000)	2.950	(0.687)
	R-RUM	5.000	(0.000)	5.000	(0.000)	5.170	(0.428)
	LLM	3.000	(0.000)	3.000	(0.000)	2.190	(0.394)
	A-CDM	4.000	(0.000)	4.000	(0.000)	3.920	(0.273)
	DINA	6.000	(0.000)	6.000	(0.000)	5.770	(0.489)
	DINO	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
400	G-DINA	1.000	(0.000)	2.000	(0.000)	2.010	(0.100)
	R-RUM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	LLM	3.000	(0.000)	3.000	(0.000)	2.990	(0.100)
	A-CDM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	DINA	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
	DINO	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
600	G-DINA	1.000	(0.000)	2.000	(0.000)	2.000	(0.000)
	R-RUM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	LLM	3.000	(0.000)	3.000	(0.000)	3.000	(0.000)
	A-CDM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	DINA	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
	DINO	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)
800	G-DINA	1.000	(0.000)	2.000	(0.000)	2.000	(0.000)
	R-RUM	5.000	(0.000)	5.000	(0.000)	5.000	(0.000)
	LLM	3.000	(0.000)	3.000	(0.000)	3.000	(0.000)
	A-CDM	4.000	(0.000)	4.000	(0.000)	4.000	(0.000)
	DINA	6.000	(0.000)	6.000	(0.000)	6.000	(0.000)
	DINO	2.000	(0.000)	1.000	(0.000)	1.000	(0.000)

た。むしろ，一貫して LLM が選択された。この結果は一考に値する。Table3 の BIC の欄を詳細に検証すると，LLM，A-CDM，R-RUM の平均順位が高い傾向がみられた。また，サンプルサイズが 200 の条件では G-DINA モデルの平均順位は 5.960 であり，むしろ最下位位に近い値でありまったく真のモデルを選択できておらず，制約の厳しい DINA モデルや DINO モデルよりも低い順位を示した。G–DINA モデルの選択される順位はサンプルサイズが上昇するに従って徐々に向上するものの，前述のように最適なモデルとしては選択されなかった。

R-RUM が真のモデルである場合には平均的には AIC も BIC も真のモデルを選択した（Table 4）。ただし，サンプルサイズが 200 のときは AIC の平均順位が 1.200，BIC の平均順位が 1.190 であり，若干モデル選択にゆらぎがあった。

真のモデルが LLM の場合，AIC も BIC も LLM を選択する傾向がみられた（Table 5）。しかし，真のモデルである LLM と A-CDM の平均順位はどちらも 1.5 前後であり 2 つのモデル選択は難しい可能性が示された。

A-CDM が真のモデルである場合には，A-CDM が正しく選択される傾向がみられたものの，やはり LLM も少なからず選択する傾向がみられた（Table 6）。LLM と A-CDM の結果は今回検証した条件下ではサンプルサイズが大きくなったとしても正しいモデル選択が行われなかった。

最後にデータ生成モデルが DINA, DINO モデルのどちらの場合であってもサンプルサイズが 200 程度であっても平均順位が 1 位であり，正確に真のモデルを選択することができている（Table 7，Table8）。

4 考察

本研究の目的は，比較的小サンプル下でデータ生成モデルと解析モデルが異なったときに，CDM の選択が正しくなされるのかを検証することであった。本研究の結果は次にまとめられ

る。１．G-DINA モデルが真のモデルの場合，AIC は真のモデルを選択するが，BIC は一貫して LLM を選択する。この傾向は小サンプル条件下では覆らなかった。２．R-RUM,が真のモデルの場合，AIC と BIC は同様に真のモデルを選択できる。３．LLM，A-CDM が真のモデルの場合，A-CDM と LLM の区別をデータから着けることは難しい可能性がある。また真のモデルが LLM であっても A-CDM であっても A-CDM が選択される可能性が若干高いと考えられる。4. DINA モデル, DINO モデルが真のモデルの場合はサンプルサイズが 200 であっても AIC, BIC は確実に真のモデルを当てることができる。

一般に，AIC は予測に適したモデルを選択し，BIC は真の構造に近いモデルを選択する傾向があるとされる。経験的には AIC はやや複雑なモデルを選択する傾向が知られている。また, BIC は罰則項にサンプルサイズがあるためサンプルサイズを大きくすることにより罰則が厳しくなる。本研究の結果からはサンプルサイズが小さい状況では AIC によるモデル選択が比較的有効である可能性が示唆された。AIC と BIC が示す最適なモデルが異なっていたのは，データ生成モデルが G-DINA モデルの場合であった。AIC は BIC よりも罰則が小さく，指標に占める尤度の比重が大きいため，パラメタが多く尤度が高くなるモデルの方が選択されやすいと考えられる。

G-DINA モデルが真のモデルの場合には，BIC は一貫して誤ったモデルを選択し，サンプルサイズが大きくなったとしても真のモデルを選択することはなかった。今回の設定では，slip・guessing パラメタのみを固定したため，データ生成の際に交互作用パラメタが必ずしも大きくならなかった可能性などが考えられる。G-DINA モデルを特徴づけるのは交互作用項であるので，この効果の大きさによっては，主効果のみのモデルで十分となる可能性が高くなると考えられる。

A-CDM と LLM は同じ非補償・主効果モデルであり，非常に類似したモデルであったため正しいモデルを選択することが困難であったと考えられる。一方，R-RUM は A-CDM，LLM と同様の主効果モデルであるものの，非補償的な性質をもち，他のモデルとは区別されやすかったと考えられる。

本研究の結果から，実データ解析におけるモデル選択において LLM や A-CDM が選択された場合には注意が必要であることを示している。また複数の情報量基準を組合せた統合的な判断を行う必要性も示唆された。

本研究の限界としては，アトリビュート数，項目数，項目パラメタ，Q 行列条件を固定し，アトリビュートの分布を一様分布に仮定した点が挙げられる。データ生成の条件を変更することで実際の結果が変化することは考えられるため，よりさまざまな条件を用いて今回の結果を精緻に検証する必要があるだろう。さらに，今回は全ての項目で同じ項目反応モデルを仮定したが，実際には項目によって項目反応モデルが異なっている可能性も考えられる。この点に関しては，実証的な研究も踏まえて，実際の学習者がどのような項目反応を持つのかを検証し，改めてシミュレーション研究を行うことも必要といえる。また，本研究では，最尤推定法を用いたが検討した状況は比較的小サンプルであるため，項目パラメタやアトリビュート習得パタンの推定値のゆらぎが大きい可能性がある。今後は小サンプル状況での項目パラメタやアトリビュート習得パタンの推定値がどのように振る

舞うのか，詳細な検討が必要である。

引用文献

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*, 716-723. doi: 10.1109/TAC.1974.1100705.

Chen, J., & de la Torre, J. (2014). A Procedure for Diagnostically Modeling Extant Large-Scale Assessment Data: The Case of the Programme for International Student Assessment in Reading. *Psychology, 5*, 1967. doi: 10.4236/psych.2014.518200

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199. doi: doi:10.1007/s11336-011-9207-7.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item responsetheory. *Applied Psychological Measurement, 25*, 258-272. doi: 10.1177/01466210122032064.

小西 貞則・北川 源四郎(2004). 情報量規準. 朝倉書店

Leighton, J., & Gierl, M. (Eds.). (2007). Cognitive diagnostic assessment for education: *Theory and applications*. New York, NY: Cambridge University Press.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*, 391-409. doi: 10.1177/0265532215590848.

Ma, W. & de la Torre, J. (2017). G-DINA: The generalized DINA model framework. R package version 1.4.2.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212. doi: 10.1007/BF02294535.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge: Cambridge University Press, pp. 275–318.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement. Theory, methods and applications.* New York, NY: Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464. doi: 10.1214/aos/1176344136.

鈴木 雅之・豊田 哲也・山口 一大・孫 媛 (2015). 認知診断モデルによる学習診断の有用性の検討―教研式標準学力検査 NRT「中学１年数学」への適用―. 日本テスト学会誌, *11*, 81-97.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305. doi: 10.1037/1082-989X.11.3.287

植野真臣・島宏二郎 (2010). 学習評価の新潮流. 朝倉書店.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625-649. doi: 10.1007/s11336-015-9471-z

山口 一大・岡田 謙介 (2017a). TIMSS2007 の日

本人サンプルを用いた認知診断モデルと項目反応理論モデルの比較. 日本テスト学会誌, *13*, 1-16.

山口 一大・岡田 謙介 (2017b). 近年の認知診断モデルの展開. 行動計量学, *13*, 17-32.