IRT による理系記述式テストデータの分析(3)
――項目パラメタ・能力パラメタによる IRT 適用課題の導出――

○泉 毅（ 株式会社教育測定研究所 ）・ 倉元 直樹 （ 東北大学 ）

1. 目的・背景

1.1. 目的

本稿では，倉元・泉(2016)の議論を受け理系記述式テストデータの IRT モデルとの整合性に関する問題を扱う。

本研究で対象とする理系記述式テストとは，短答式，数式を用いた解答，図や絵を用いた解答，穴埋め式などの解答形式で出題される理系分野の学力測定を目的とするテストを言う。本稿では，理系分野として高等学校の数学，理科を想定している。なお，一つのテストの中に記述式で解答を求める設問とともに，真偽式や多枝選択式などの選択式の解答形式の設問も含まれることがあるが，全体として理系記述式テストと呼ぶことにする。

本稿は，設問が形式的に局所独立の仮定を満たすことが難しい場合，逸脱がどの程度パラメタの推定に影響を及ぼすのか，実際のデータから項目パラメタ，能力パラメタを算出し，それらをもとに検討を加えることを目的とする。

1.2. 局所独立の仮定に関する技術的問題点

局所独立の仮定は，IRT を用いた分析を行う際に必要とされる重要な仮定であるが，理系記述式テストでは大問形式が主流である。大問形式での出題は長文問題のようなあるテーマに沿った出題，また前の問題が解けなければ後の問題が解けない出題等が考えられる。これらの形式では原理的に局所独立の仮定を満たすことができない。

Yen (1993) は，局所独立の仮定を満たさないことを局所依存 (LID) と呼び，LID を引き起こす原因を複数挙げている。記述式かつ大問形式のテストにおいて，LID を引き起こす原因として文脈依存と項目間の連鎖性が挙げられる。

文脈依存とは，複数の項目間の間接的な相互関係性を指す。大問形式のテストにおいては，一つのリード文のもとにある一定のテーマが与えられ，それに沿った複数の小問が出題される。文脈依存の典型的な出題形式と言える。

項目の連鎖性とは，二つの項目間のより直接的な相互関係性を表す。例えば，数式を展開させるような設問の場合，直前の問題に正解できなければその次の設問に正答できないというような構造の出題形式によくみられる。

1.3. 項目の連鎖性の分類

Yen (1993) の連鎖性の定義は「二項目間に正答率の影響があること」とされる。項目の関係性と影響の強さに関して明示的な言及はない。本稿では，現実のテスト場面に即して項目間の関係性の違いに着目して下位分類を行った。

テストに含まれる任意の二つの項目のペアを，出題順に「前の項目」「後の項目」と表現し，受験者群のそれぞれの項目への「正答，誤答」の解答パタンを考える。受験者の解答は，必ず「前の項目に正答かつ後の項目に正答」「前の項目に正答かつ後の項目に誤答」「前の項目に誤答かつ後の項目に正答」「前の項目に誤答かつ後の項目に誤答」パタンいずれかに属する。項目ペアの連鎖性の違いを 4 パタンの出現頻度の違いとして表現した。

1.3.1. 実質的同一項目

二つの項目に関する正答が同一の操作で同時に導かれる関係である。前の項目の正答情報さえあれば後の項目に必ず正答できるようなケースで，正誤情報はほぼ同一となる。モデル上の出現パタンは「前の項目に正答かつ後の項目に正答」と「前の項目に誤答かつ後の項目に誤答」の二つに限られる。「実質的同一項目による連鎖項目ペア」と呼ぶ。

1.3.2. 完全連鎖項目

前の項目の解答情報を用いて後の項目の解答を導く項目構造を指す。この場合，前の項目に不正解であると後の項目は必ず不正解となる。「完全連鎖項目ペア」とする。

完全連鎖項目ペアの場合，前の項目に正答した受験者は後の項目にも正答する可能性がある。しかし，前の項目に誤答した受験者は，後の項目には正答することが出来ない。よって，原理的には「前の項目に誤答かつ後の項目に正答」以外の三つの解答パタンの出現があり得る。

1.3.3. 部分連鎖項目

より連鎖性が弱い項目ペア「部分連鎖項目」と呼ぶ。4 パタンすべての解答パタンをとる可能性があるが，相対的に「前の項目に正答かつ後の項目に誤答」，「前の項目に誤答かつ後の項目に正答」のパタンが相対的に少ない。

1.3.4. 連鎖性がない項目

項目ペアに連鎖性がない場合は，4 パタンすべての解答パタンをとる。連鎖性がない項目ペアでも，同じリード文を共有する等，他のいくつかの項目と文脈依存性があるために，完全には局所独立を満たしていないケースもある。

2. 方法

2.1. 分析対象者

本研究では，倉元 (2003) で作成，実施されたテストデータを用いる。このテストは大学進学を目指す高校 3 年生の生徒, 約2, 900名の参加のもとに解答を得たものであり，数学分野，物理分野，化学分野，生物分野から出題がなされた (倉元, 2003)。全項目無解答者を除いた解答データを分析の対象とした。分析対象者数を表 1 に示す。

表 1. 各テストの分析対象者数

	分析対象者数
数学	2733
物理	1776
化学	2639
生物	946

2.2. 分析モデル

項目パラメタの推定に大きな負荷がかからない単純なモデルを用いることとした。具体的には，正誤のデータ，すなわち 2 カテゴリのデータには 2 パラメータロジスティックモデル (2PLM) を, 2 カテゴリよりも多い多値データには段階反応モデル(GRM: Samejima, 1969) を採用する。

2.3. 分析方法

まず, 各教科の一次元性をスクリープロットで確認する。次に二値データならば，2PLM，多値データならば GRM によって分析し, 項目パラメタ・能力パラメタの推定を行う。

同じデータを再分類し，カテゴリ数が 2 の場合と多値の場合に分けて分析した。内容から見て合理的な部分点とするため，各作題者へのヒアリングをもとに閾値を定めた。作題者には本テストの項目得点のヒストグラムを提示し，妥当な閾値について判断を仰いだ。本稿では, 二値型項目，多値型項目と呼ぶ。また，連鎖性の構造についてもヒアリングを行い，連鎖性をもつ項目の把握を行った。

IRT への適用に際して，能力パラメタが標準正規分布する際に想定される識別力パラメタ推定値の値を 0～2.0 とし，この値の範囲に収まるか否かという観点から検討を行った。2.0 以上の推定値を, 識別力の過大推定とみなすこととした。また，2PLM，GRM の能力パラメタ推定値と素点合計の散布図，相関係数を求め，分析法による違いを検討した。

識別力パラメタが過大推定された項目が連鎖性のある項目であれば, その項目群をテストレットとして再度 2PLM, GRM，能力パラメタに関する分析を行った。なお，分析には IRTPRO ver.2.1 (Cai, Thissen & du Toit, 2011) を用いた。

3. 結果

紙面の都合上，項目数が 4 分野の中で最も多かった物理分野の結果の一部について記載する。

3.1. 基礎統計量

物理分野のテストの構成は，大問 1 に 4 項目 (item2_1_01 ～ item2_1_04)，大問 2 に 2 項目 (item2_2_01, item2_2_02)，大問 3 に 5 項目 (item2_3_01 ～item2_3_05) が含まれ，計 11 項目であった。またそれぞれの項目には 3～9 点の配点がなされ，計 50 点満点であった。

また，物理分野のテストでは 11 項目中 4 項目が多枝選択式の解答形式であった。その他の 7 項目は記述式の解答を求める設問であり，図を用いた解答，短答式，論述式が各 1 項目，数式を用いた解答が 4 項目であった。各項目の解答形式・配点・基礎統計量を表 2 に示す。

3.2. 連鎖性

物理分野は大問 1 と大問 2 に含まれる項目，すなわち Item2_1_01～Item2_1_04，Item2_2_01，Item2_2_02 は独立した項目であることが分かった。

表 2. 各項目の解答形式・配点・基礎統計量 (物理分野)

項目	解答形式	配点	平均得点	得点率	標準偏差
Item2_1_01	多枝選択式	3	2.18	72.5%	1.34
Item2_1_02	多枝選択式	3	2.20	73.2%	1.33
Item2_1_03	多枝選択式	3	2.31	77.1%	1.26
Item2_1_04	多枝選択式	8	2.91	36.3%	2.65
Item2_2_01	図を用いた解答	3	0.06	2.0%	0.42
Item2_2_02	短答式	4	1.16	29.1%	1.63
Item2_3_01	数式を用いた解答	9	4.62	51.4%	3.80
Item2_3_02	論述式	3	0.38	12.7%	1.00
Item2_3_03	数式を用いた解答	4	0.86	21.6%	1.20
Item2_3_04	数式を用いた解答	5	0.65	13.1%	1.46
Item2_3_05	数式を用いた解答	5	0.04	0.8%	0.44

大問 3 に関しては，Item2_3_01 と Item2_3_02，Item2_3_03 と Item2_3_04, Item2_3_04 と Item2_3_05, が部分連鎖の項目ペアとされた。

3.3. スクリープロット

一次元性を確認するため，配点をもとにした素点合計点から求めたスクリープロットを描いた。物理分野のスクリープロットを図 1. に示す。第一固有値と第二固有値に格段の差があると判断し，分析を続けた。

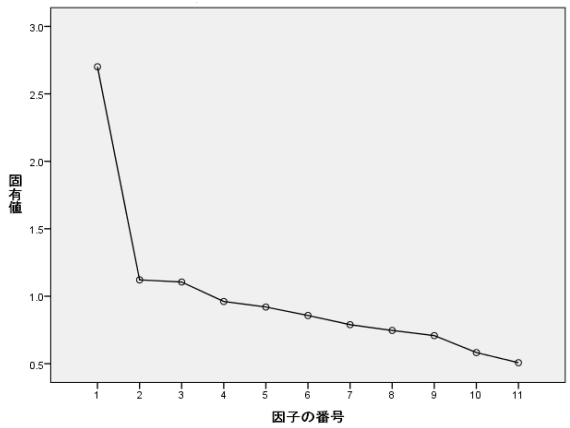

図 1. 物理分野 11 項目のスクリープロット

3.4. 項目パラメタ

二値型項目分析における，物理分野の項目パラメタの推定値を表3に示す。物理分野は item2_3_01, item2_3_03, item2_3_04，item2_3_05 の識別力パラメタが過大推定された。また，連鎖性がないが，item2_1_01，item2_1_02, item2_1_03 の識別力パラメタには，通常よりも小さい推定値が得られた。

表 3. 物理分野 11 項目（二値型）の項目パラメタ

項目	*a*	*s.e.*	*b*	*s.e.*
item2_1_01	0.17	0.06	-5.61	2.06
item2_1_02	0.35	0.07	-2.93	0.55
item2_1_03	0.09	0.07	-14.28	11.29
item2_1_04	1.11	0.11	0.25	0.09
item2_2_01	1.09	0.29	4.07	0.97
item2_2_02	1.11	0.11	0.60	0.12
item2_3_01	2.43	0.21	-0.08	0.06
item2_3_02	1.42	0.14	1.79	0.16
item2_3_03	2.51	0.23	0.40	0.08
item2_3_04	2.24	0.28	1.05	0.14
item2_3_05	3.29	2.16	2.70	0.55

各分野において一部の項目をテストレットとして扱い項目パラメタの算出を行った。テストレットを含む二値型項目分析における項目パラメタの算出結果を表 4 に示す。作題者への連鎖性の構造に関するヒアリングの結果に基づき，物理分野は部分連鎖の関係にあるとされた item2_3_01 と item2_3_02 をテストレット (testlet2_3_01_02) とし，また，同様に部分連鎖の関係にあるとされた item2_3_03, item2_3_04，item2_3_05 の 3 項目をテストレット (testlet2_3_03_04_05) とした。識別力が通常より小さく推定された item2_1_01，item2_1_02，item2_1_03 は連鎖性のない項目であった。これらの項目はテストレットとして扱うことは不適切であると判断し，テストレットとせず分析を行った。

表 4. テストレットを含む
物理分野 8 項目（二値型）の項目パラメタ

項目	*a*	*s.e.*	*b1*	*s.e.*	*b2*	*s.e.*	*b3*	*s.e.*
item2_1_01	0.17	0.06	-5.64	2.11				
item2_1_02	0.36	0.07	-2.89	0.54				
item2_1_03	0.08	0.07	-14.85	12.30				
item2_1_04	1.14	0.09	0.24	0.05				
item2_2_01	1.12	0.23	4.01	0.65				
item2_2_02	1.15	0.09	0.59	0.06				
testlet2_3_01_02	2.17	0.16	-0.14	0.04	1.58	0.07		
testlet2_3_03_04_05	2.53	0.22	0.26	0.04	1.21	0.06	3.01	0.17

その結果，二つのテストレットとした項目 (testlet2_3_01_02 と testlet2_3_03_04_05) で，再び識別力パラメタが過大推定された。また，二値型項目分析と同様，識別力パラメタが非常に低く，かつ困難度パラメタと標準誤差が非常に高い，推定結果が極めて不安定である項目がみられた (item2_1_01，item2_1_02，item2_1_03) 。

3.5. 能力パラメタ

二値型項目における能力パラメタ，多値型項目における能力パラメタ推定値の散布図を図 2 に示す。これは，2PLM と GRM の二つのモデルに基づいて各受験者の能力パラメタを推定し，その結果を散布図に表したものである。

その結果，能力パラメタ推定値が 0.5 以上，-0.5 以下の部分において，推定された能力パラメタ値の絶対値が大きいほど二つの分析による能力値の差に開きがある受験者が増加していく様子が見られる。

また，二値型項目における能力パラメタ，多値型項目における能力パラメタ推定値の相関係数の値は r=.966 であった。

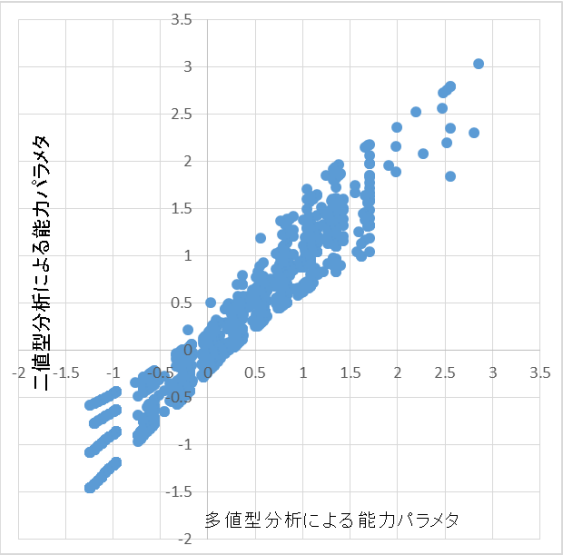

図 2. 物理二値型と多値型の能力パラメタ散布図(r=. 966)

4. 考察

物理分野においては，二値型にせよ，多値型にせよ，局所独立の仮定の逸脱は識別力パラメタの過大推定につながることが示唆された。さらにテストレット化によって局所独立を解消ないしは軽減するという方略も効果がなかった。

識別力パラメタの安定した推定という観点からは，二値型項目として分析した場合でも多値型項目として分析した場合でも，分析結果は変わらなかった。したがって，本研究の結果からは，部分点の扱いの違いが大きく識別力パラメタの推定に影響する可能性は小さいことが示唆された。

一方，他の項目との連鎖性があり，局所独立の仮定を満たさない項目については，項目パラメタの推定に悪影響を及ぼすことが実証的に示された。構造的に局所独立の仮定を逸脱する項目のペアが存在しない生物分野のテストデータに対する分析 (泉・倉元，2014) は，受容可能な範囲に識別力パラメタが収まるなど，項目パラメタの推定が比較的上手く行っていたのに対し，連鎖性がある項目を含む数学分野，物理分野，化学分野は，いずれも識別力パラメタの過大推定がみられた。しかも，識別力パラメタが過大推定された項目は，いずれも他の項目と構造的に完全連鎖，ないしは，部分連鎖という形で直接的な連鎖性がある項目であった。

能力パラメタの観点からは，どの分野においても，二値型項目分析と多値型項目分析から得られた能力パラメタ推定値の相関係数の値は.95 程度の，高い値となった。しかし，散布図を描くと，能力パラメタ推定値の絶対値が大きくなるにつれ，二値型項目分析と多値型項目分析の推定の差異が見られ，直線的な関係とは言えないことが示された。

5. IRT 適用の課題

本研究の結果からは，どの程度の強さの関係性が識別力パラメタの過大推定につながるのか，識別力パラメタの推定という観点からは，どの程度の局所独立の仮定からの逸脱であれば許されるのか，といった点に関しては，課題が残った。

また，一連の分析に用いた項目数は 6〜11 項目と一般的な客観式テストと比較すると非常に少なかった。IRT モデルによる分析の上で，項目数の確保が難しいということは，理系記述式テストの宿命とも言える本質的な構造的問題である。識別力パラメタの過大推定をはじめとして，項目パラメタの推定が不安定になるという問題は，項目の局所独立の仮定の逸脱という問題のみならず，推定に必要な項目数が確保できていないことに由来する可能性が考えられる。

少なくとも局所独立の仮定の逸脱の程度による影響の把握，また，項目数確保は，記述式テストへ IRT 適用を行うことを考えるならば，避けて通れない難題であると言えるだろう。

引用文献

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows. [Computer software]. Lincolnwood, IL: Scientific Software International

泉毅・倉元直樹 (2014). IRT による理系記述式テストデータの分析――高校生対象の生物テストデータを用いて――, 日本テスト学会第 12 回大会発表論文集, 170-173.

倉元直樹 (2003). 高校と大学の教育接続を重視した試験問題開発研究――モニター調査結果報告――, 夏目達也 (編) 高校と大学のアーティキュレーションに寄与する新しい大学入試についての実践的研究, 平成12〜14年度日本学術振興会科学研究費補助金 (基盤研究[A]), 研究課題番号 12301014, 研究代表者 夏目達也, 研究成果報告書, 99-175.

倉元直樹・泉毅 (2016). IRTによる理系記述式テストデータの分析(2)――記述式テストIRT化検討の背景――, 日本テスト学会第14回大会発表論文集.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometorika Monograph Supplement*, **34**, 100-114.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, **30**, 187-213.

付記

本稿は JSPS 科研費 15K13124, 16H02051 の助成に基づく研究成果の一部である。