教育情報学研究　第16号（2017）
Educational Informatics Research, 2017, No.16, 77-94

項目反応理論による理系記述式テストデータの分析
――項目間の連鎖性と項目得点のカテゴリ化を巡って――

泉　毅 *，倉元　直樹 **

*株式会社教育測定研究所
**東北大学

要旨：本研究では既存の理系記述式テストデータの分析を通して，大学入試等の我が国の大規模テストにおける記述式テストへの項目反応理論 (IRT) 適用に関する課題を探る．理系記述式テストに対する IRT 適用時の問題点としては，項目間の連鎖性による局所独立の仮定の逸脱と部分点の扱いが大きな障害となることが挙げられる．本研究では，局所独立の仮定への逸脱に対する IRT モデルの頑健性と，部分点の扱いがパラメタ推定に及ぼす影響について，複数の IRT モデルの比較を行った．その結果，いずれのモデルにおいても項目パラメタ推定の不安定さの解消が難しいことが分かった．局所依存構造を持つ項目のテストレット化も十分な推定精度の改善にはつながらなかった．本研究において，理系記述式テストへの IRT モデル適用は困難であることが示された．

キーワード：理系記述式テスト，テストレット，項目反応理論，局所独立，連鎖性

1. 目的
1.1. 背景

項目反応理論 (項目応答理論，Item Response Theory : IRT，以後，IRT と表記する) が適用される場面の典型は，客観式かつ一問一答型の大規模テストである．しかし，近年の我が国の教育改革の議論を概観する限りでは，従来からの客観式テストの枠を超えた形式のテストに対しても，IRT 適用の期待が高まっているように思われる．近い将来，様々なテスト場面における IRT 適用の可能性を検討すべき状況の到来が予想される．

従来，少数の例外を除いて，IRT が我が国のハイステークスな試験で積極的に活用されることはなかった．しかし，ここ数年の教育改革論議の中でにわかに大学入試への IRT 適用に対する期待が広範囲に広がってきた．2013 (平成25) 年10月，教育再生実行会議の第四次提言 (教育再生実行会議，2013) において，大学入試の共通試験として，いわゆる達成度テスト (仮称) の導入が提言された．提言によれば，達成度テスト (仮称) には，基礎レベルと発展レベルがあり，発展レベルは，大学の入学者選抜を実施する際の基礎資格として利用する，とされている．また，受験回数は，一発勝負ではなく「複数回挑戦を可能」とするために，年複数回実施が提言され，コンピュータ式テスト (Computer Based Testing: CBT，以後，CBT と表記する) による試験実施が示唆された．

上記の提言を受け，2014(平成26) 年3月に発表された中央教育審議会高大接続特別部会の審議経過報告 (中央教育審議会，2014a) では，達成度テスト (仮称) の在り方として「記述式を導入すること」「紙媒体ではなくコンピュータによる出題・解答の方式を導入すること」に関する専門的な検討を進める，との方針が示された．また，複数の試験機会の得点を比較可能とするには「IRT (項目反応理論) 等を用いた得点調整，得点表示方式」についての検討が必要であるとされた．さらに，答申 (中央教育審議会，2014b) では，それまで「達成度テスト (仮称) 基礎レベル」として議論されてきた構想の共通試験が「高等学校基礎学力テスト (仮称) 」，「達成度テスト (仮称) 発展レベル」として議論されてきた構想の共通試験が「大学入

学希望者学力評価テスト (仮称) 」と名称変更されて，その後の議論が展開されている．

さらに，以上の議論を受けて2015 (平成27) 年9月の高大接続システム改革会議「中間まとめ」(高大接続システム改革会議，2015) においては，大学入試センター試験に代わる試験として，大学入試希望者学力評価テスト (仮称) の導入が提言され，IRT，CBT の導入に関する検討と同時に記述式テストの導入に関する議論がなされた．その一方で，例えば，村上 (2003) は，受験機会が複数設定され，それぞれの成績を同一の尺度上での比較することに関して，我が国では理論的基盤と試験を実施するためのバックアップ態勢が弱い，と指摘しているが，そのような社会的基盤に関する議論が十分なされた痕跡は認められない．

高大接続システム改革会議の「最終報告」(高大接続システム改革会議，2016) では，IRT や CBT の適用は後景に退き，大学入学希望者学力評価テストに対する国語記述式問題の導入が主たる検討の対象となっている．本稿執筆時点では，記述式テストに対して IRT 適用が必要であるという議論は生じていない．しかし，状況の変化に応じて，近い将来，記述式テストを念頭に置いた IRT の運用に関する議論が再燃することも十分に考えられる．

1.2. 問題の所在

基本的に記述式テストは IRT になじまないとされる．その理由は枚挙にいとまがない．例えば一つの設問に対する解答に長時間を要し，一人の受験者が多数の項目に解答できない上に客観的な採点が行えず，採点者に起因する誤差が混入するなど IRT が想定している項目プールを用いた大規模テストを構成するには難しい性質を持っている．したがって従来の我が国の大学入試の慣行に照らせば，記述式テストに IRT を適用することは考えられなかった．

一方，共通試験である大学入学希望者学力評価テスト (仮称) では，将来的な IRT の実用化と同時に記述式の導入が議論されている．現時点では両者は別々の問題と考えられている．導入の是非が議論されている試験で即座に記述式テストを IRT で運用することが求められているわけではないが，近い将来，導入に向けての検討が必要となる可能性は皆無ではない．大学入学希望者学力評価テスト (仮称) に関わる議論を視野に入れた場合，記述式テストデータへの IRT 適用は我が国における現実のテスト場面が直面する課題の一つである．現在，最も焦点が当たっているのは国語における記述式の設問であるが，同時に数学の記述式問題についても検討が行われている．

しかし，IRT の応用場面として，上記のような状況を念頭に置いた検討は，我が国ではまだ始まっていない．村上 (2003) でも指摘されたように，我が国で IRT を用いたテストを本格的に導入するためには，様々な観点からの基礎研究をあらかじめ行っておく必要がある．

そこで本研究では，我が国の大学入試において，記述式テストへの IRT の活用に関して潜在的な期待が高まっていることを受け，従来は適用されていなかった解答形式のテストに対する IRT 適用の検討を行う．本研究の対象は，いわゆる記述式テストの中でも，理系分野の問題である．

大学入試の場面で理系分野の記述式テストが実施されているのは，大学独自に行われる個別試験である．宮本他 (2016) によれば，2015 (平成27) 年度入試における国公立大学の一般入学者選抜試験において出題された数学の解答形式のほぼ全てが数式を展開する形の記述式，ないしは，図・絵等で解答する形式による出題であった．

本研究においては，数学を含む理系記述式テストの IRT 適用に関する問題点について，実際のテストデータを用い，テスト理論的な観点から検証することを試みる．理系記述式テストへの IRT 適用を検討し，その適用事例を提示することは，記述式テストへの IRT 活用のフィージビリティの側面において社会的な意義があると考えられる．

1.3. 理系記述式テスト

本研究で対象とする理系記述式テストは，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．一部の項目に多枝選択式や真偽式などの選択式の解答形式も含まれることがあるが，全体をとして理系記述式テストと考えることとする．

多くの場合は数学や物理，化学，生物，地学などの理科の科目ないしは科目群を出題範囲とする，教科・科目型の学力検査として出題されることが多い．それ以外に合教科・科目型の総合問題として出題される場合もある．いずれも選択式のみのテストと比べて一項目の解答に要する時間が長く，多数の項目を出題することが難しい．項目間の関係も各設問が相互に独立な内容となる形式ではなく，一群の設問が一つのテーマを構成するいわゆる大問形式であることが多い．

1.4. 本研究の目的

大学入学希望者学力評価テスト (仮称) のような大学入試の共通試験における記述式テスト導入の実現には，複数回実施した場合の等化，採点の公平性，明確な採点基準の設定，多数の項目数の確保等，運用上の問題は多岐にわたるが，本研究における検討の対象とはしない．本研究で検討の対象とするのは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である．特に，項目パラメタの合理的な推定が可能か否かといった点に焦点を当てる．本研究における課題は以下の3点である．

1点目は，部分点の扱いに関する課題である．記述式テスト項目の特徴として，解答結果が正誤の二値ではなく，部分点を含む多値になる場合が多いことが挙げられる．したがって，部分点の扱いもパラメタ推定に影響を及ぼすことが考えられる．

2点目は，局所独立の仮定に関する課題である．記述式テストの場合，テスト項目の構造的な問題として，IRT によって根本的な条件である局所独立の仮定を満たすことは難しい．それでも IRT の適用を行わなければならない場合，その逸脱がどの程度パラメタの推定に影響を及ぼすのか，といったことは検討に値する．

3点目は，項目数に関する課題である．そもそも記述式テストは，客観式テストと比べ，項目数の確保が難しい．このことは本研究で扱う理系記述式テストにおいても同様である．少ない項目数で，項目パラメタの合理的な推定が可能か否か検討する必要がある．

1点目と，2点目の課題に関して，次節で詳述する．

1.5. IRT 適用に際し予想される技術的課題

⑴部分点の扱いに関する技術的問題

部分点の扱いに関して検討しておかなければならない必要性は，一般的な配点及び採点基準に由来する．

記述式テスト項目においては，欠損値となる部分点ないしは極めて少数の答案しか該当しない部分点が存在する場合が多い．すなわち，配点規則上，特定の部分点を取りうる答案がほとんどないというケースが考えられる．したがって，二値モデルはもちろんのこと部分点に対応できないが，多値モデルを適用する場合でも部分点が与えられたデータをそのままのカテゴリとして IRT を適用することには技術的な問題が生じると予想される．

本研究では，これらを解決するために便宜的に部分点と部分点の間に閾値を設定し，カテゴリ数を減少させる方法を採ることとした．まず，一つ目の方法として，部分得点の閾値によって項目得点を正答と誤答に二分する方法が考えられる．その場合には基本的な二値の IRT モデルが適用可能となる．もう一つの方法は，カテゴリ数を元の部分点の数よりも少なくするが，二値までには絞らない方法である．その場合には多値データに対応する IRT モデルを適用することが必要となる．

日本のテストにおいて IRT によってテストデータを分析する場合には，基本的に多枝選択式の二値データを扱う場合が多い．しかし，解答結果が正答・誤答のみではなく，部分点が与えられ2カテゴリ以上になる場合，段階反応モデル (Graded Response Model, GRM: Samejima, 1969, 以後，GRM と表記する)，一般化部分得点モデル (Generalized Partial Credit Model: GPCM, Muraki, 1992)，名義反応モデル (Nominal Response Model: NRM, Bock, 1972) 等の多値型のモデルが適用される．

多枝選択式のテストデータに対し多値型のモデルを用いて分析を行った国内の研究としては，平井 (1993)，平井･渡部 (1994)，石塚他 (2001)，御園･水町 (2011) が挙げられる．これらの先行研究には全て GRM が用いられた．また，これらの先行研究では記述式のテストデータに関する言及はなされていない．

記述式のテストデータに段階反応モデルを適用した研究は，管見の限りでは平井・渡部 (1994) の小論文の採点データを GRM により分析した研究，また国語の記述式テストデータの GRM による分析 (柴山，2012) のみである．しかし，平井・渡部 (1994) は，カテゴリ採点の有効性の検討，また評定者の評価の特徴を把握することを主眼としており，評定者11人を項目と見立てた上で分析を行っている．そのため記述式テストデータへの段階反応モデルの適用とは言いがたい．また，柴山 (2012) は記述式テストデータに対し GRM を適用したが，局所独立の仮定を満たさない項目に関する観点からの分析はない．

⑵局所独立の仮定に関する技術的問題

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基本的な前提条件である．豊田 (2002) の表現を借りれば，局所独立の仮定とは「能力パラメタ θi が所与である場合には，項目反応は互いに独立である」ことを意味する．これは，J 個の項目への項目反応 $x_1,x_2,\cdots x_J$ の条件付き同時確率 $P\,(x_1,x_2,\cdots,x_J|\theta i)$ が条件付き周辺確率の積に等しくなることである．つまり局所独立の仮定とは反応データについて，

$$P\left(x_1,x_2,\cdots,x_J\middle|\theta_i\right)=\prod_{j=1}^{J}P\,(x_j|\theta_i) \qquad (1)$$

となることを仮定するというものである．

Yen (1993) は，局所独立の仮定を満たさないことを局所依存 (Local item dependence : LID) と呼び，LID を引き起こす原因を複数挙げている．その中で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考えられるものとしては，文脈依存 (Passage dependence) と項目間の連鎖性 (Item chaining) 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．すなわち，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のが文脈依存の典型的な出題形式と言える．

それに対して，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．

また，文脈依存の関係にせよ，連鎖性にせよ，局所独立を仮定できないデータに IRT を適用すると，能力パラメタの推定誤差の増加 (登藤，2010) や項目識別力の過大推定 (Chen & Wang, 2007; Tuerlinckx & De Boeck, 2001) が起こるとされている．したがって，一つの目安としては，局所独立の仮定への侵犯の有無を項目識別力の推定値が通常のテストで想定される値の範囲に収まっているか否かで判断することが考えられる．

局所依存性の問題を解決する方法の一つとして，一つの大問に含まれる項目群をテストレット (Wainer & Kiely, 1987) にまとめる方法が考えられる．テストレットとは，複数の項目をひとまとめとし，一つの項目とみなして扱うことを指す．具体的な分析の際には，局所独立性を満たさない項目群をテストレットとし，テストレットごとに多値型モデルによる分析を行うというものである．

テストレットに対して適用する多値型モデルとして，石塚他 (2001) は GRM，Sireci, Thissen and Wainer (1991) は NRM，Yen (1993) は GPCM を用いた．この分析の利点として石塚他 (2001) は，項目間の局所独立性の仮定を項目の集合であるテストレット間の局所独立性で置き換えることができるとしている．しかし，テストレットを用いるデメリットとして，複数の項目を一つの項目として扱うため項目数が減ることが挙げられる．本研究で分析の対象とする理系記述式テストにおいては，元来，多くの項目数を確保するのが難しい．したがって，分析過程の中で，モデルの精緻化と項目数のいずれを取るか，という選択に迫られることが考えられる．

1.6. 項目の連鎖性の分類

Yen (1993) の局所依存に関する議論のうち，連鎖性と呼ばれるものは「二項目間に正答率の影響があること」とされるにとどまっており，項目の関係性と影響の強さに関しては，明示的には言及されていない．そこで本節では，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，項目の連鎖性をその関係性と強さに応じて分類することを試みる．

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．すなわち，項目ペアの連鎖性の違いをこの4パタンの出現頻度の違いとして表現する試みである．

本研究では，連鎖性の強さに基づき，項目ペアの連鎖性について概念的に以下の四種類に分類することを試みる．

⑴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ような関係を考える．前の項目の正答情報さえあれば，ケアレスミスさえなければ後の項目には必ず正答できるような形式である．すなわち，実質的に全く同じ設問が続けて出題されているようなケースとみなすことができる．

その結果，二つの項目の正誤情報は，ほぼ同一となり「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」という二種類の解答パタンは原則として見られない．したがって，モデル上は，実際に出現するパタンが「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」という二つに限られることとなる．このような項目の連鎖のパタンを「実質的同一項目による連鎖項目ペア」と呼ぶ．

⑵完全連鎖項目

例えばセンター試験の数学のように前の項目の解答情報を用いて後の項目の解答を導くような項目構造が考えられる．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．このように，後の項目に正答するためには前の項目で正答していなければならない，といったような形で強い連鎖性を有している項目ペアを「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，完全連鎖項目ペアの場合，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．

⑶部分連鎖項目

「完全連鎖項目」ペアと比べると明示的な連鎖性が弱いと考えられる項目ペアの関係性を「部分連鎖項目」と呼ぶ．すなわち，構造的には完全連鎖項目のように見えても，前の項目の正答を経由しないで後の項目に正答可能な別解が存在するようなケースである．

部分連鎖項目ペアの場合，四種類すべての解答パタンをとる可能性がある．部分連鎖項目ペアの特徴としては，連鎖の度合いが強くなるにつれて「前の項目に正答かつ後の項目に誤答」の解答パタンが増加し，逆に「前の項目に誤答かつ後の項目に正答」のパタンが減少する．すなわち，完全連鎖項目ペアに近づいていく．逆に，連鎖の度合いが弱くなるにつれて「前の項目に誤答かつ後の項目に正答」の解答パタンが増加し，逆に「前の項目に正答かつ後の項目に誤答」のパタンが減少する．

⑷連鎖性がない項目

項目ペアに明示的な連鎖性がない場合は，四種類すべての解答パタンをとる．

連鎖性がない項目ペアの場合，全ての項目について相互に構造的な関係性が見られない局所独立性が満たされたケースが典型的である．しかし，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考えられる．

2. 方法

2.1. 分析対象者

本研究では，倉元 (2003) で用いられた，大学入学試験問題開発研究のために実施されたテストデータを用いて分析を行う．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る (倉元 , 2003).

このテストは，数学分野，物理分野，化学分野，生物分野からの出題がなされた．本研究では，こ

の4分野への，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．各テストへの分析対象者数を表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. テストの構成

テストが実施された際には各60分の「総合問題Ⅰ」，「総合問題Ⅱ」として出題されたが，実質的には四つの分野がそれぞれ独立した大問を構成していた．

総合問題Ⅰは実質的には数学分野の出題である．大問2問で構成され，大問1が小問4問，大問2が小問3問から成る．受験者は全ての設問に解答することとなっていた．

総合問題Ⅱはそれぞれ物理分野，化学分野，生物分野から二つの分野を選択解答する形式であった．表1で示したようにほとんどの受験者が化学を選択した．物理分野は大問3問から構成され，大問1が小問4問，大問2が小問2問であった．大問3は，見かけは7問だが，そのうちの3問は一つの設問としてまとめて採点されており，実質5問の小問から成る．化学分野は大問2問から構成され，第1問が小問3問，第2問が小問6問から成る．生物分野は大問1問であり，含まれる小問は6問であった．

以後，「分野 *，大問 **，第 *** 問」に該当する小問の表記を「Item*_**_***」のように表す．なお，それぞれ，分野1は数学分野，分野2は物理分野，分野3は化学分野，分野4は生物分野を表す．なお，大問内における小問の実際の表記には統一したフォーマットはなく，様々に表現されていた．

2.3. 分析モデル

IRT では受験者の項目への解答に基づいて，被験者の能力値，項目のパラメタを推定する．本研究では，正誤のデータ，すなわち2カテゴリのデータを扱う二値モデルに加え，部分点が存在する等の条件の下で3カテゴリ以上の値を取るデータを扱う多値モデルを扱うケースを想定する．

二値データを分析するモデルとしては，2パラメタ・ロジスティック・モデル (two parameter logistic model: 2PLM，以後，2PLM と表記する) を使用した．2PLM とは，能力パラメタが θ_i である受験者 i が項目 j に正答する確率 $P_j(\theta_i)$ を

$$P_j(\theta_i) = \frac{1}{1+\exp\left[-a_j(\theta_i - b_j)\right]} \tag{2}$$

とするものである．ここで，a_j，b_j はそれぞれ，項目 j の識別力パラメタおよび困難度パラメタである．

また多値データを分析するモデルとして GRM を使用した．

$$u_j = 0,1,2,\cdots,c,\cdots,C-1 \tag{3}$$

$$P(u_j = c|\theta_i) = P_{jc}(\theta_i) = P^*_{jc}(\theta_i) - P^*_{jc+1}(\theta_i) \tag{4}$$

$$P^*_{j0}(\theta_i) = 1 \tag{5}$$

$$P^*_{jC}(\theta_i) = 0 \tag{6}$$

$$P^*_{jc}(\theta_i) = \frac{1}{1+\exp\left[-a_j(\theta_i - b^*_{jc})\right]} \tag{7}$$

u_j は，C 個の値をとる順序尺度の離散変数である．$P(u_j=c|\theta_i)$ は，能力パラメタが θ_i である受験者が $u_j=c$ と反応する確率を表している．$P_{jc}(\theta_i)$ は，能力パラメタが θ_i である受験者 i が項目 j において c と反応する確率を表している．また，$P^*_{jc}(\theta_i)$ は，能力パラメタが θ_i である受験者 i が項目 j において c 以上と反応する確率を表している．a_j は，項目 j における識別力パラメタであり，b^*_{jc} は項目 j において c 以上のカテゴリをとることに対する困難度パラメタある．

本研究で分析の対象とする理系記述式テストに対しては，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，部分点を正誤に振り分けたデータ，すなわち2カテゴリのデータには2PLM を採用することとした．また，部分点にカテゴリを設けて2カテゴリよりも多いデータには項目パラメタの解釈の容易さから GRM を採用することとした．

2.4. 分析方法

⑴一次元性の確認

各分野のテストについて，それぞれ一次元の構造を持ち，IRT 適用の前提を満たしていることを確認した．具体的には，倉元 (2003) のデータを用いて，スクリープロットを描いた．

⑵部分点の扱いと局所依存の構造の同定

次に，部分点の扱いを決定し，局所依存の構造について把握することとした．それにはその分野と当該の試験問題に関する専門的知識と判断が必要となるため，作題者に対するヒアリングを実施した．なお，数学分野の第2問の作題者はすでに故人となっていたため，一緒に作題を行って出題意図と採点基準を熟知している第1問の作題者をヒアリングの対象者とした．作題者には日時を指定して，対面，ないしは，インターネットコミュニケーションソフトスカイプ (Skype) を用いて遠隔でインタビューを行った．

インタビューにおいては，まず，本研究で用いたテスト項目の内容と項目得点のヒストグラムを提示して，作題意図と採点基準について思い出してもらうこととした．その上で，作題者が採点基準に沿って，個々の設問ごとに部分点のカテゴリ合併に関わる閾値を定め，作題意図に沿った合理的な再カテゴライズを行った．最初は，一部の部分点を同一のカテゴリとみなした多値データの作成基準を定めた．このような形で多値データに加工したテスト結果を「多値型テスト」と呼ぶ．なお，あえて部分点のカテゴリを作らずに二値とした方が自然な場合もあったため，多値型テストには正誤の2カテゴリしか持たない項目も含まれる．次いで，部分点のカテゴリをさらに正答，誤答の二値に分類する閾値を定めた．本研究では，このような形で二値データに加工したものを「二値型テスト」と呼ぶ．

さらに，各テストの作題者には，項目間の連鎖性の構造に関するヒアリングを行った．明らかに「実質的同一項目」とみなすことのできる項目ペアは存在しなかったが，ある項目の解がそれより後の別の項目に正答するための前提として必要な場合，先述のように「完全連鎖項目」ないしは「部分連鎖項目」とみなされる．作題者には大問内の各項目の構造について，完全連鎖ないしは部分連鎖の関係にある項目について特定することを求めた．

⑶Q_3統計量の算出

作題者へのヒアリングによる局所依存の構造の結果との違いを把握するため，項目間の局所依存の度合いを測る指標として，それぞれのモデルから推定された項目パラメタ推定値と能力パラメタ推定値と反応データを用いてQ_3統計量 (Yen, 1984) を算出する．Q_3統計量を用いる利点として，Chen and Wang (2007) は，算出が容易であり，また，他の局所依存を測る指標では，実際のデータでは値を求めることができない場合があるのに対し，Q_3統計量は，そのような制限が少ない，としている．

二値データにおけるQ_3統計量は (8) 〜 (12) 式を用いて求められる．

$$P_j(\hat{\theta}_i) = \frac{1}{1+\exp\left[-\hat{a}_j(\hat{\theta}_i - \hat{b}_j)\right]} \tag{8}$$

$$E_{ji} = P_j(\hat{\theta}_i) \tag{9}$$

$$d_{ji} = x_{ji} - E_{ji} \tag{10}$$

$$\boldsymbol{d}_j = \{d_{j1}, d_{j2}, \cdots, d_{jN}\} \tag{11}$$

$$Q_{3jj'} = r(\boldsymbol{d}_j, \boldsymbol{d}_{j'}) \tag{12}$$

E_{ji} は，能力パラメタが$\hat{\theta}_i$である受験者 i が項目 j に反応した場合の期待値 $P_j(\hat{\theta}_i)$ である．d_{ji} は，受験者 i の観測得点 x_{ji} と期待値 E_{ji} の差を表す．$\boldsymbol{d}_j$，$\boldsymbol{d}_{j'}$ は，それぞれ，全ての受験者から得られた d_{ji}，$d_{j'i}$) を要素とするベクトルである．N は受験者数を表す．$Q_{3jj'}$ は，$\boldsymbol{d}_j$, $\boldsymbol{d}_{j'}$) の相関係数である．

また，多値データにおけるQ_3は (13) 式から (18) 式を用いて求められる．

$$P_{jk}(\hat{\theta}_i) = P^*_{jk}(\hat{\theta}_i) - P^*_{jk+1}(\hat{\theta}_i) \tag{13}$$

$$P^*_{jk}(\hat{\theta}_i) = \frac{1}{1+\exp\left[-\hat{a}_j(\hat{\theta}_i - \hat{b}^*_{jk})\right]} \tag{14}$$

$$E_{ji} = \sum_{k=0}^{K} k\, P_{jk}(\hat{\theta}_i) \tag{15}$$

$$d_{ji} = x_{ji} - E_{ji} \tag{16}$$

$$\boldsymbol{d}_j = \{d_{j1}, d_{j2}, \cdots, d_{jN}\} \tag{17}$$

$$Q_{3jj'} = r(\boldsymbol{d}_j, \boldsymbol{d}_{j'}) \tag{18}$$

ここで，項目 j を含むテストレットから得られる得点を k=0,…,K とする．E_{ji} は，能力パラメタが $\hat{\theta}_i$ である受験者 i の，項目 j を含むテストレットへの期待得点である．d_{ji} は，受験者 i の観測得点 x_{ji} と期待値 E_{ji} の差を表す．d_j，$d_{j'}$) は，それぞれ全ての受験者から得られた d_{ji}，$d_{j'i}$) を要素とするベクトルである．$Q_{3jj'}$ は，d_j，$d_{j'}$ の相関係数である．

Q_3統計量の値が 0.2を超えると，項目間の局所依存の度合いが高いと判断される (Chen & Thissen, 1997)．本研究においては, 二値型テスト，多値型テストにおける Q_3統計量の値を求め，0.2を超える項目ペアを確認する．

⑷正常な識別力パラメタの範囲の決定

IRT への適用に際して，識別力パラメタ推定値の通常想定される値を 0.0 ～ 2.0 とし，この値の範囲に収まるか, という観点からの検討を行った．経験的に2.0以上の識別力を持つ項目は客観式テストでも多くない．まして，主観的評価による測定誤差が入り込む記述式テストで2.0を超える識別力は期待できないと考え，便宜的に2.0を基準として判断することとした．識別力パラメタの推定値 0.0 ～ 2.0 に収まらない場合，識別力の過大推定とみなすこととした．推定に際し，能力パラメタに対しては標準正規分布を仮定することとした．

⑸異常な識別力パラメタに関する処理

識別力パラメタ推定値が通常の範囲に収まらなかったケースが生じ，さらにその項目が，他の項目と連鎖性のある項目であれば，泉他 (2013) の方法にしたがい，連鎖性のある項目群を一つのテストレットとみなして分析を行うこととした．その際には，一つのテストレットに含まれる複数の項目の合計得点を GRM における一つの項目のカテゴリとみなして扱う．

なお，本研究で用いたテストは分野によっては複数の大問を含む構成となっており，同一の大問の中には文脈依存性が認められる．しかし，本研究では，局所依存からの逸脱として文脈依存性よりも強い関係性を持つ連鎖性の構造を問題とするため，大問をテストレットとみなした分析は行わない．

以上のことから，本研究においては，

ⅰ) 二値型テストに対する2PLM
ⅱ) 多値型テストに対する GRM
ⅲ) テストレットを含む二値型テストに対する GRM
ⅳ) テストレットを含む多値型テストに対する GRM

の4種類の分析を行い，理系記述式テストデータへの IRT 適用課題を検討する．

分析には IRTPRO ver.2.1 (Cai, Thissen & du Toit, 2011) を用いる．なお，多値型テストの中に二値型テストが含まれる場合には，2PLM とみなして分析されることになる．

3. 結果・考察

3.1. 一次元性の確認

図1 ～図4は，数学分野，物理分野，化学分野．生物分野のスクリープロットである．

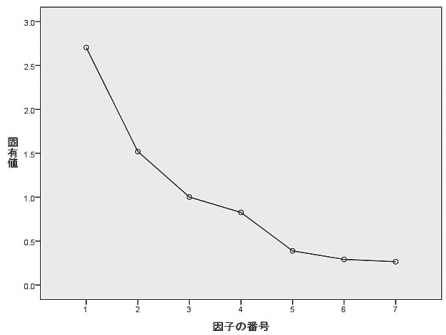

図1　数学分野7項目のスクリープロット

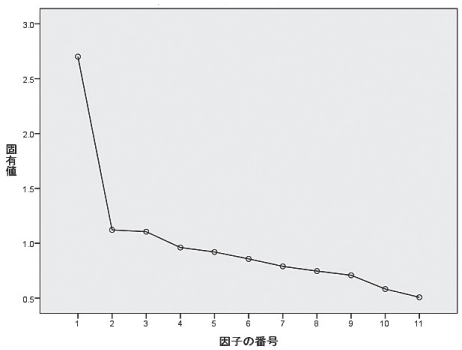

図2　物理分野11項目のスクリープロット

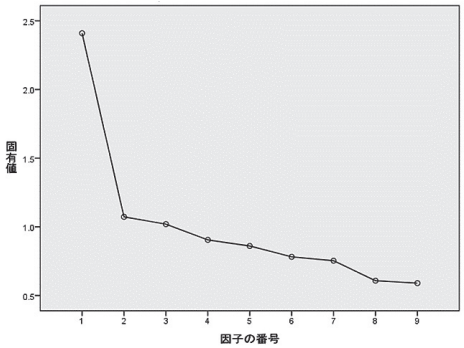

図3　化学分野9項目のスクリープロット

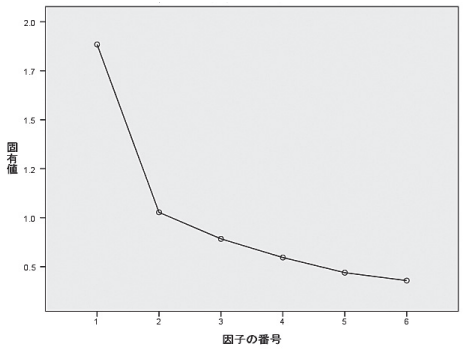

図4　生物分野9項目のスクリープロット

各分野とも，第1固有値の寄与が十分に大きかった．数学の第2固有値が比較的大きく，積極的に一次元性が満たされているとは言い難いが，第1固有値と第2固有値の差が大きいため，テストとして一次元性が保たれていると判断した．

3.2. 連鎖性の構造

先述の手続きにしたがって，各テストの作題者に各教科の連鎖性の構造に関するヒアリングを行った結果を図5 ～図8に示す．黒の項目は連鎖性のある項目を示す．矢印は連鎖の向きを示す．また，細い矢印は部分連鎖を指し，太い矢印は完全連鎖を指す．灰色の項目は連鎖性のない項目である．

数学分野においては，Item1_1_01とItem1_1_02，Item1_1_02とItem1_1_03が部分連鎖の項目ペア，Item1_1_03とItem1_1_04が完全連鎖の項目ペアと判断された．また，Item1_2_01は独立した項目であるが，Item1_2_02とItem1_2_03が完全連鎖の項目ペアであると判断された．

図5　数学7項目の連鎖性の構造

物理分野においては大問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と判断された．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判断された．

図6　物理11項目の連鎖性の構造

化学分野では，大問1においてItem3_1_01とItem3_1_02，Item3_1_01とItem3_1_03が部分連鎖の項目ペアと判断された．また，大問2に関してはItem3_2_01，Item3_2_02，Item3_2_04が独立した項目であった．Item3_2_03とItem3_2_06，Item3_2_05とItem3_2_06は部分連鎖の項目ペアと判断された．

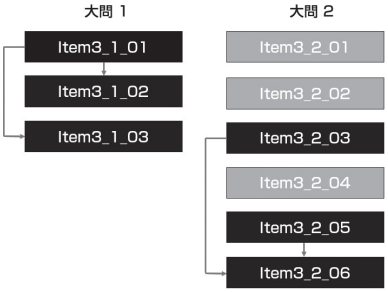

図7　化学9項目の連鎖性の構造

生物分野はItem4_01 ～ Item4_06すべてが同じリード文を共有するものの，明示的な関連性が見られない「連鎖性がない項目」で構成されていた．すなわち，大問としては1問のみの出題であるために，生物単体のテストとしては，構造的に局所独立の仮定を侵犯しない構造となっていた．

図8　生物6項目の連鎖性の構造

各分野の連鎖性の構造には特徴がみられた．数学分野は，大問1に含まれる全ての項目がそれぞれ一つ前の項目の解答を前提として解く形式の，ひとつながりの構造を成していた．さらに，大問2の3項目のうちの2項目に連鎖性があり，全体として項目間の連鎖性が強い構造となっていた．物理分野は大問1と大問2では構造的に明示的な連鎖性は見られなかったが，大問3は，5項目中3項目が一つ前の項目と連鎖性を持つ構造となっていた．化学では，大問1は全ての項目が連鎖していたが，連続する項目間ではなく2問目と3問目が1問目の項目の解答を前提とした構造となっており，数学分野や物理分野とはやや性質の違う連鎖性を持つ構造が見出された．さらに，大問2に関しては，連鎖性のない項目と連鎖性のある項目が順序を超えて混在する複雑な構造を有していた．生物分野は全ての項目がリード文を共有していたものの，構造としては連鎖性がなく，形式的には局所独立が保たれている可能性が示唆された．

3.3. Q_3統計量による局所依存構造の検出

定量的に局所依存構造を把握するため，Q_3統計量の算出を行った．

表2　Q_3の値が0.2を超えた項目ペアの数

	二値型	多値型
数学分野	2	3
物理分野	0	0
化学分野	0	0
生物分野	0	0

Q_3統計量の値が0.2を超えた項目ペアの数を二値型テスト，多値型テストの各分野に関して表2に示す．

数学分野では，二値型テストにおいて，item1_2_01とitem1_2_02，item1_2_02とitem1_2_03の二つの項目ペアのQ_3統計量が0.2を超えた．また，多値型テストにおいては，二値型項目における二つの項目ペアに加えてitem1_1_01とitem1_1_02の三つの項目ペアのQ_3統計量が0.2を超えた．item1_2_01とitem1_2_02の項目ペアは，作問者へのヒアリングによる連鎖性の構造と異なるが，その他の項目ペアは一致する結果となった．

物理分野，化学分野，に関しては，作題者のヒアリングによる結果とは異なり，Q_3統計量の観点からは，局所独立の仮定を満たさない項目ペアは見出せなかった．

生物分野に関しては，作題者のヒアリングによる連鎖性の構造と一致する結果となった．

3.4. 項目分析と項目パラメタの推定

各分野に関して，項目ごとの配点，解答形式，古典的な項目分析で用いられる得点率とIT相関，さらに，二値型テスト，多値型テストで得られたaパラメタ，bパラメタを表3 ～表6に示す．配点，平均得点，得点率，IT相関に関して，これらは再カテゴライズがなされる前のデータを用いて求められた．また，IT相関は，各項目で得られた得点と，各テストの合計得点との相関係数である．

数学分野の二値型テストの結果に関して，数学分野は，item1_1_03，item1_1_04の識別力パラメタが過大推定された．これらの項目は完全連鎖の関係にある項目ペアであった．多値型テストにおいても，同様にitem1_1_03，item1_1_04の識別力パラメタが過大推定された．さらに，困難度パラメタについても極端に低い値（item1_2_01のb_1）

と高い値（item1_2_02，item1_2_03の b_3）が見られた．

以後，識別力パラメタが過大推定されたと考えられる項目を過大推定項目と呼ぶ．

物理分野は，二値型テストにおいて 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．item2_3_01は item2_3_02と部分連鎖の関係にあり，item2_3_03と item2_3_04，item2_3_04と item2_3_05も部分連鎖の関係にある項目ペアであった．また，過大推定とは逆に，item2_1_01，item2_1_02，item2_1_03の識別力パラメタは，十分な識別性能が得られていなかった．多値型テストにおいても同様の結果となった．さらに，困難度パラメタについても極端に低い値（item2_1_01 ～ 03の b，b_1）と高い値（item2_2_01，item2_3_05の b，b_1，item2_1_04の b_4）が見られた．

化学分野は二値型テストにおいて，item3_1_02が過大推定された．item3_1_02は item3_1_01と部分連鎖の関係にある項目ペアであった．多値型テストにおいても同様の結果となった．さらに，困難度パラメタについても極端に高い値（item3_2_04，item3_2_06の b，b_1．item3_2_02の b，b_1，b_2．item3_2_01の b_2）が見られた．

全ての項目が構造的に局所独立と判断された生物分野では，多値型テストにおいて高い値の困難度パラメタ（item4_01，item4_04の b2）が見出されたが，識別力パラメタが異常な値を示した項目はなかった．

全ての分野において，多値型項目として分析した場合でも，二値型項目とした場合と同様に項目パラメタの推定が不安定であった．これらの結果から，部分点のカテゴリ化の工夫といった方法によって，局所独立の仮定の逸脱による識別力パラメタの過大推定の問題を解消することは困難であることが示唆された．

表3　各項目の解答形式･配点･基礎統計量･項目パラメタ (数学分野)

項目	解答形式	配点	平均得点	得点率	IT相関	二値型		多値型			
						a	*b*	*a*	*b1*	*b2*	*b3*
item1_1_01	数式を用いた解答	10	8.17	0.82	0.56	1.68	-0.63	1.69	-1.84	-0.62	
item1_1_02	数式を用いた解答	10	6.71	0.67	0.61	1.28	-0.18	1.38	-1.66	-0.15	
item1_1_03	数式を用いた解答	20	6.82	0.34	0.79	3.26	-0.02	4.82	-0.39	0.89	
item1_1_04	数式を用いた解答	10	1.95	0.20	0.71	4.34	0.78	37.11	0.73	1.01	
item1_2_01	図と数式を用いた解答	15	12.29	0.82	0.47	0.63	-1.61	0.51	-4.16	-1.99	
item1_2_02	数式を用いた解答	15	2.57	0.17	0.59	0.83	1.07	0.69	1.21	1.23	4.80
item1_2_03	数式を用いた解答	20	2.19	0.11	0.56	0.85	1.51	0.68	1.80	2.50	9.94

表4　各項目の解答形式･配点･基礎統計量･項目パラメタ (物理分野)

項目	解答形式	配点	平均得点	得点率	IT相関	二値型		多値型				
						a	*b*	*a*	*b1*	*b2*	*b3*	*b4*
Item2_1_01	多枝選択式	3	2.18	0.73	0.22	0.17	-5.61	0.17	-5.77			
Item2_1_02	多枝選択式	3	2.20	0.73	0.28	0.35	-2.93	0.36	-2.91			
Item2_1_03	多枝選択式	3	2.31	0.77	0.18	0.09	-14.28	0.08	-14.96			
Item2_1_04	多枝選択式	8	2.91	0.36	0.66	1.11	0.25	1.08	-0.87	0.23	1.43	2.32
Item2_2_01	図を用いた解答	3	0.06	0.02	0.18	1.09	4.07	1.09	4.09			
Item2_2_02	短答式	4	1.16	0.29	0.53	1.11	0.60	1.06	0.62			
Item2_3_01	数式を用いた解答	9	4.62	0.51	0.79	2.43	-0.08	2.05	-0.57	-0.09	0.53	
Item2_3_02	論述式	3	0.38	0.13	0.43	1.42	1.79	1.43	1.79			
Item2_3_03	数式を用いた解答	4	0.86	0.22	0.66	2.51	0.40	2.43	0.40	1.95		
Item2_3_04	数式を用いた解答	5	0.65	0.13	0.57	2.24	1.05	2.32	1.05	1.73		
Item2_3_05	数式を用いた解答	5	0.04	0.01	0.25	3.29	2.70	3.42	2.66			

表5　各項目の解答形式･配点･基礎統計量･項目パラメタ (化学分野)

項目	解答形式	配点	平均得点	得点率	IT相関	二値型		多値型				
						a	*b*	*a*	*b1*	*b2*	*b3*	*b4*
Item3_1_01	穴埋め式	14	6.49	0.46	0.75	1.34	0.32	1.59	-1.65	-0.78	0.30	2.03
Item3_1_02	数式を用いた解答	6	0.71	0.12	0.60	3.19	1.27	2.93	1.29			
Item3_1_03	数式を用いた解答	6	2.28	0.38	0.70	1.75	0.28	1.74	0.29	0.59		
Item3_2_01	論述式	3	0.78	0.26	0.41	0.89	0.80	0.90	0.81	2.31		
Item3_2_02	論述式	3	0.36	0.12	0.25	0.61	2.23	0.61	2.23	4.30		
Item3_2_03	穴埋め式	3	0.57	0.19	0.41	0.89	1.87	0.94	1.79			
Item3_2_04	短答式	3	0.27	0.09	0.34	1.08	2.53	1.10	2.50			
Item3_2_05	穴埋め式	6	2.96	0.49	0.66	1.18	0.26	1.26	-0.19	0.26		
Item3_2_06	数式を用いた解答	6	0.02	0.00	0.09	1.20	5.24	1.20	5.22			

表6　各項目の解答形式･配点･基礎統計量･項目パラメタ (生物分野)

項目	解答形式	配点	平均得点	得点率	IT相関	二値型		多値型					
						a	*b*	*a*	*b1*	*b2*	*b3*	*b4*	*b5*
Item4_01	短答式	10	3.42	0.34	0.54	0.83	-0.42	0.94	-0.40	2.57			
Item4_02	論述式	10	4.48	0.45	0.67	1.01	-0.03	1.05	-0.09	-0.03	0.34	0.48	0.54
Item4_03	短答式	10	3.11	0.31	0.65	0.96	0.74	1.05	0.70	1.16			
Item4_04	短答式	10	1.37	0.14	0.50	1.09	1.86	1.02	1.95	2.36			
Item4_05	短答式	5	1.70	0.34	0.41	0.88	0.87	0.85	0.90				
Item4_06	短答式	5	3.34	0.67	0.53	1.82	-0.59	1.69	-0.61				

3.5. テストレットモデルによる項目パラメタの推定

(1) テストレットを含む二値型テスト

二値型テストの分析結果として，数学分野，物理分野，化学分野においては過大推定項目がみられた．作題者のヒアリングに基づけば，これらの項目はすべて，他の項目と完全連鎖，ないしは，部分連鎖の関係にある項目であった．構造的な局所独立の仮定への侵犯が識別力パラメタの過大推定につながったことが考えられる．

そこで，連鎖性への対処として，これらの項目を含む項目ペアをテストレットとし，テストレットを含む二値データでの分析を行った．

連鎖性の構造に関する作題者へのヒアリング結果に基づき，数学分野は完全連鎖の関係にあるとされたitem1_1_03とitem1_1_04の項目ペアをテストレット(testlet1_1_03_04) として分析を行うこととした．数学分野における，テストレットを含む二値型テストの項目パラメタ推定値を表7に示す．その結果，テストレット項目とは別の連鎖性のある項目ペアそれぞれ (item1_2_02，item1_2_03) の識別力パラメタが過大推定された．物理分野は部分連鎖の関係にあるitem2_3_01とitem2_3_02をテストレット項目 (testlet2_3_01_02) とし，部分連鎖の関係にあるitem2_3_03，item2_3_04，item2_3_05を一つにまとめ，3項目を含むテストレット項目 (testlet2_3_03_04_05) とした．識別力パラメタに十分な識別性能が得られなかった3項目item2_1_01，item2_1_02，item2_1_03は構造的な連鎖性がない項目であった．これらの項目をテストレットとして扱うことは不適切であると考え，二値型テストとしたまま分析に加えることとした．物理分野における，テストレットを含む二値型テストの項目パラメタ推定値を表8に示す．結果として，二つのテストレット項目 (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．さらに，item2_1_01，item2_1_02，item2_1_03の識別力パラメタの推定結果が小さく，極端な値を取る困難度パラメタ（item2_1_01 ～ 03，item2_2_01の b_1，testlet2_3_03_04_05の b_3）の問題も解消されなかった．項目パラメタ推定は，またしても全体的に不安定であった．

表7　テストレットを含む二値型テストの項目パラメタ推定値 (数学分野)

項目	*a*	*b1*	*b2*
item1_1_01	0.78	-1.04	
item1_1_02	0.80	-0.25	
testlet1_1_03_04	0.77	-0.06	1.73
item1_2_01	1.21	-0.98	
item1_2_02	5.20	0.50	
item1_2_03	3.56	0.76	

表8　テストレットを含む二値型テストの項目パラメタ推定値 (物理分野)

項目	*a*	*b1*	*b2*	*b3*
item2_1_01	0.17	-5.64		
item2_1_02	0.36	-2.89		
item2_1_03	0.08	-14.85		
item2_1_04	1.14	0.24		
item2_2_01	1.12	4.01		
item2_2_02	1.15	0.59		
testlet2_3_01_02	2.17	-0.14	1.58	
testlet2_3_03_04_05	2.53	0.26	1.21	3.01

化学分野は部分連鎖の関係にある item3_1_01 と item3_1_02の項目ペア (testlet3_1_01_02) をテストレット項目として分析を行うこととした．化学分野における，テストレットを含む二値型テストの項目パラメタ推定値を表9に示す．すべての項目において識別力パラメタは過大推定されなかったが，極端に高い値を示す困難度パラメタ（item3_2_02，item3_2_04．item3_2_06の b_1）の問題は解消されなかった．

表9　テストレットを含む二値型テストの項目パラメタ推定値 (化学分野)

項目	*a*	*b1*	*b2*
testlet3_1_01_02	1.68	0.18	1.77
item3_1_03	1.69	0.28	
item3_2_01	0.90	0.79	
item3_2_02	0.64	2.13	
item3_2_03	0.94	1.79	
item3_2_04	1.11	2.49	
item3_2_05	1.25	0.25	
item3_2_06	1.23	5.12	

生物分野については，二値型テストにおける分析で問題ない推定値が得られたので，テストレットを含む分析は行わないこととした．

(2)　テストレットを含む多値型テスト

二値型テストと同様に，項目パラメタが過大推定された項目を含む連鎖性のある項目ペアをテストレットとし，テストレットを含む多値データとしての分析を行うこととした．テストレット化の判断は二値型テストと同様である．

数学分野の結果を表10に示す．過大推定項目は二値型テストの分析結果と同様であった．さらに，二値型テストでは解消された極端に高い値示す困難度パラメタが再び現れた（testlet1_1_03_04の b_4，item1_2_03の b_3）

表10　テストレットを含む多値型テストの項目パラメタ推定値 (数学分野)

項目	*a*	*b1*	*b2*	*b3*	*b4*
item1_1_01	1.06	-2.46	-0.81		
item1_1_02	1.07	-1.95	-0.18		
testlet1_1_03_04	0.90	-0.78	1.46	1.76	2.33
item1_2_01	0.98	-2.34	-1.13		
item1_2_02	3.75	0.54	0.54	1.88	
item1_2_03	3.83	0.75	1.02	3.23	

物理分野もテストレット化の判断は二値型テストと同様である．結果を表11に示す．過大推定項目は二値型テストの分析結果と同様であった．極端な値を取る困難度パラメタ（item2_1_01 ～ 03の b_1，item2_1_04の b_4，item2_2_01の b_1，testlet2_3_03_04_05の b_4，b_5）の問題も解消されなかった．

表11　テストレットを含む多値型テストの項目パラメタ推定値 (物理分野)

項目	*a*	*b1*	*b2*	*b3*	*b4*	*b5*
item2_1_01	0.17	-5.79				
item2_1_02	0.36	-2.88				
item2_1_03	0.08	-15.71				
item2_1_04	1.08	-0.87	0.23	1.42	2.31	
item2_2_01	1.08	4.13				
item2_2_02	1.07	0.62				
testlet2_3_01_02	2.02	-0.61	-0.12	0.46	1.81	
testlet2_3_03_04_05	2.86	0.25	1.02	1.63	2.23	3.18

化学分野もテストレット化の判断は二値型テストと同様である．結果を表12に示す．二値型テストの分析結果と同様に過大推定項目はなかったが，極端に高い値を示す困難度パラメタ（testlet3_1_01_02の b_5．item3_2_01，item3_2_02の b_2，item3_2_04．item3_2_06の b_1）の問題は解消されなかった．

表12　テストレットを含む多値型テストの項目パラメタ推定値 (化学分野)

項目	*a*	*b1*	*b2*	*b3*	*b4*	*b5*
testlet3_1_01_02	1.80	-1.55	-0.76	0.20	1.38	2.50
item3_1_03	1.70	0.29	0.59			
item3_2_01	0.90	0.80	2.30			
item3_2_02	0.63	2.18	4.20			
item3_2_03	0.99	1.73				
item3_2_04	1.11	2.50				
item3_2_05	1.31	-0.19	0.25			
item3_2_06	1.19	5.27				

数学分野，物理分野，化学分野のすべてにおいて，二値型テストの分析結果と類似の結果が得られた．テストレットを含まない分析で過大推定項目が1項目であった化学分野では過大推定が解消されたものの，困難度パラメタの値は改善しなかった．さらに，複数項目に過大推定が見られた数学分野，物理分野ではテストレット化による過大推定の解消は上手く行かなかった．特に数学では，テストレット項目ではなく，新たに他の項目で過大推定となるなど，極めて不安定な構造が見られた．

4．総括

本研究の出発点として，現在行われている大学入試改革の政策的な議論が，はたしてテストの学術的な研究成果を踏まえた上で現実的に実現可能な条件を設定した上で行われているものであるのか，という問題意識があった．現実には様々な課題，論点が交錯する中，本研究では理系記述式テストに焦点を絞ることとした．しかも，網羅的に実施条件を検討するのではなく，純粋にテスト理論的な観点から三つの課題に絞り込んで検討を加えた．いずれも理系記述式テストが，IRT モデルが要請するテストの性質に合致しないことを前提に，それを克服することができるのかどうか，といった課題設定である．

また，本研究では，評価の指標を極めて単純な視点を設定した．すなわち，識別力パラメタの過大推定や極端な値を取る困難度パラメタが発生するか否か，項目パラメタ推定に問題が起こるとすればそれを防ぐ方法があるのか，ということである．そもそも，多枝選択式のような客観式テストと比較した場合，理系記述式テストは部分点を与えることができるのが利点であるが，部分点を再カテゴリ化することでその利点にあらかじめ制約をかける代わりに，可能な限り精度の良い推定を試みようとしたのが，本研究の姿勢であった．

二値型テストの分析においては，二値モデルの IRT による分析を試みた．項目の構造が局所独立の仮定と矛盾しなかった生物分野を除き，連鎖性のある項目を含んだ数学分野，物理分野，化学分野では識別力パラメタの推定は安定しなかった．また，部分点の再カテゴリ化の問題に注目し，多値型テストとしての分析を行ったが，結果は二値型テストとほぼ同様であった．連鎖性のある項目をテストレットとみなした場合でも，過大推定の解消には至らなかった．

二値型テストとしての分析，テストレットを含む二値型テストの分析，多値型テストの分析，テストレットを含む多値型テストの分析，いずれにおいても共通の弱点は項目数を十分に確保できないことであった．二値型テスト，多値型テストにおいて分析に用いたデータは，6 ～ 11項目である．一般的な客観式テストと比較すると項目数が非常に少ない．Q_3統計量の値が0.2を超えた項目ペアが数学分野の2 ～ 3組しか見いだせなかったということは，局所依存構造が存在しないことの証明ではなく，項目数不足によって局所依存個所の特定に失敗した結果とみるべきである．もともと，得点率が極端に低い項目（得点率0.1未満の項目が物理分野の item2_2_01，item2_3_05，化学分野の item3_2_06）が含まれていたことも項目パラメタ推定の不安定さを助長した一因であろう．しかし，通常行われるように不良項目を除いた分析も極めて難しい．さらに項目数を減らすことになるとともに，一つ一つの設問に解答する時間と労力の大きさを考えると，そのこと自体が受験者の

パフォーマンスを落とすことにつながる可能性が高いからである．

以上のことから，識別力パラメタの過大推定をはじめとする，項目パラメタの推定の不安定さの本質的な要因として，推定に必要な項目数が確保できていないことが示唆された．また，このことを加味すると，本研究の分析結果から，カテゴリ化の工夫や，項目連鎖性による局所独立性が満たされない場合への対処が，どの程度パラメタ推定へ影響を及ぼしたのか判断することは，分析結果の解釈の限界を超えるものであり，困難であると言える．

本研究は，理系記述式テストのような複雑な構造を備えるテスト形式の出題に IRT に基づく CBT を適用するといった斬新かつ大胆な構想に対して，フィージビリティ・スタディに踏み込む意味があるかどうかを検討するための最初の試金石と位置づけられるものである．すなわち，テスト理論的に理系記述式テストに IRT を適用しても問題がないこと，また，適切に運用するための現実的な最低条件等を提示できなければ，構想自体が机上の空論に過ぎない．本研究では，理系記述式テストへの IRT モデル適用を試みたが，その最初の段階を踏むことすら容易ではないことが示唆される結果となった．それが本研究によって見出された最大の成果と言える．

本研究は単なる一つのケーススタディに過ぎない．しかし，シミュレーションではなく，実際に受験者が解答を行ったデータに基づく実証研究であるところに意義がある．もちろん，各設問の難易度が適切であったか否か，というような，本研究で用いられたデータに固有の問題点が残るかもしれない．受験者の能力分布に対して，その全範囲を適切に識別するような設問を工夫して，新たにデータを取って再分析を行うことも可能であろう．しかし，本研究の結果から見ると，得られた知見が本研究で用いられたデータに固有の問題であるとも言い切れない．すなわち，本研究で焦点を当てた課題は，典型的な理系記述式テストに共通する性質と考えられるからである．

式・グラフ等を描くことを通じてより論理的な思考力・表現力の発揮が期待できる (高大接続システム改革会議 , 2016)，といった数学や物理等の理系記述式テストの利点を生かしながら，IRT モデルに適合するようなテストを設計することは容易な作業ではない．IRT モデルによく適合する多数の項目の中の一部に連鎖性のある項目が存在するような状況では，相互に連鎖する項目をテストレットとして一つにまとめることによって，項目パラメタの異常推定の問題はある程度解決できる可能性はある．しかし，理系記述式テストにおいては，肝心の項目数の確保という課題がほぼ克服不可能な難題である．短時間で解答可能な設問を多数集めるような設計のテストを考えるならば，あえて記述式にする意味はない．さらに，客観式テストと異なり，記述式の形式では採点者が必要となる．理系記述式テストで期待されるような高度な思考力や表現力の発露を適切に評価するには，その分野のエキスパートが採点作業に相当の時間を費やす必要がある．その上，複数の採点者が採点に当たったとしても，採点プロセスにおける誤差の混入はまぬがれない．コンピュータによる自動採点を開発しようにも，定型的で標準的な解答が想定されるような設問では，あえて記述式を採用することの意義が問われる事態となるであろう．

このように，IRT モデルによる理系記述式テストの開発という課題は，相互に矛盾した条件が重なっており，万人が満足できる解決策の得られないような構想と言える．

もちろん，将来的にこれらの問題への解決策が提示され，IRT モデルに基づく理系記述式テストが運営されている状況が招来される可能性は否定しようがないであろう．しかし，大学入試のハイステークスなテストというものは，単なる調査と異なり，個人の命運がかかるものである．何らかの失敗があれば，受験者にとって不幸なだけでなく，社会的に激しく糾弾されることになる．展望のない可能性に依拠して安易に手を付けられるものではない．まして，IRT モデルに基づく大規模テストには，事前に項目パラメタが推定された膨大な数の項目を持つ秘匿された項目プールが必要となる．予備調査のためにテスト項目が人目にさらされても設問が測定しようとする特性や能力の性質に変化はないのか，項目を秘匿したままに予備調査が可能なのか，といった類の問題に対する

検討は，全く着手されていない状況である．

将来的に起こりうる問題を未然に防ぐことも重要な研究課題の一つである．その点において，本研究が試みた分析は，限りなく成功の可能性が低い上に高いコストが伴う道に踏み込むことを防止するために設けられる道標の一つとしての役割を担うことになると考える．

付記

本研究は第 1 著者の博士学位請求論文（泉, 2016）の一部の章に対して大幅に加筆修正を加えたものである．

また，本研究はJSPS科研費，課題番号15K13124の助成に基づく研究成果の一部である．

引用文献

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46,443-459.

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

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Chen, C., & Wang, W. (2007). Effect of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31,* 388-411.

中央教育審議会 (2014a). 中央教育審議会高大接続特別部会審議経過報告 , 平成26年3月25日 (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2014/04/01/1346157_1.pdf, 最終閲覧日 2016年12月5日).

中央教育審議会 (2014b). 新しい時代にふさわしい高大接続の実現に向けた高等学校教育 , 大学教育 , 大学入学者選抜の一体的改革について――すべての若者が夢や目標を芽吹かせ , 未来に花開かせるために――(答申), 平成26年12月22日 (http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2015/01/14/1354191.pdf, 最終閲覧日 2016年12月5日).

平井洋子 (1993). 多肢選択式テストによる測定の精緻化の試み . 東京大学教育学部紀要 , 33, 167-175.

平井洋子・渡部洋 (1994). 小論文評点のカテゴリ化に関する測定論的考察 . 行動計量学 , 21, 21-31.

石塚智一・中畝菜穂子・内田照久・前川眞一 (2001). テストレットモデルによる英語試験問題の分析 . 大学入試センター研究紀要 , 30, 1-24.

泉毅 (2016). 理系記述式テストへの IRT 適用課題の検討，博士学位論文，東北大学大学院教育情報学教育部.

泉毅・山野井真児・山田剛史・白川隆朋・対馬英樹 (2013). 局所独立性を満たさないテストデータに対する段階反応モデルの適用―2PLM による分析との比較検討―. 日本テスト学会誌 , 9, 37-55.

高大接続システム改革会議 (2015). 高大接続システム改革会議「中間まとめ」, 平成27年9月15日 (http://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/__icsFiles/afieldfile/2015/09/15/1362096_01_2_1.pdf, 最終閲覧日 2016年12月5日).

高大接続システム改革会議 (2016). 高大接続システム改革会議「最終報告」, 平成28年3月31日 (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2016/06/02/1369232_01_2.pdf, 最終閲覧日 2016年12月5日).

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.

教育再生実行会議 (2013). 高等学校教育と大学教育の接続・大学入学者選抜の在り方について (第四次提言), 平成25年10月31日 (http://www.kantei.go.jp/jp/singi/kyouikusaisei/pdf/

dai4_1.pdf, 最終閲覧日 2016年12月5日).
御園真史・水町龍一 (2011). テストレットモデルによる数学分野の問題項目分析 . 日本教育工学会研究報告集 , 4, 177-180.
宮本友弘・庄司強・田中光晴・石上正敏・倉元直樹 (2016). 国立大学における個別学力試験の解答形式に関する研究 (1). 日本テスト学会第14回大会抄録集 , 40-41.
村上隆 (2003). 研究の背景と目的 . 村上隆 (編) 我が国の公的機関における得点等化の導入に向けた心理・教育測定的研究 , 平成12 〜 14年度日本学術振興会科学研究費補助金 (特別研究推進費 [1]) , 課題番号12800015, 研究代表者 村上隆 , 研究成果報告書 ,1-11.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometorika Monograph Supplement, 34,* 100-114.
柴山直 (2012). 全国規模の学力調査における重複テスト分冊法の展開可能性について . 柴山直 (編) 平成23年度文部科学省委託研究「学力調査を活用した専門的課題分析に関する調査研究」研究成果報告書
Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement,* 28, 237-247.
登藤直弥 (2010). 局所独立性の仮定が満たされない場合の潜在特性推定への影響 . 日本テスト学会誌 , 6, 17-28.
豊田秀樹 (2002). 項目反応理論 [入門編]―テストと測定の科学― 朝倉書店
Tuerlinckx, F., & De Boeck, P. (2001). The Effect of Ignoring Item Interactions on the Estimated Discrimination Parameters in Item Response Theory. *Psychological Methods,* 6, 181-195.
Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement,* 24, 185-202.
Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

An Analysis of Science Constructed-Response Tests by Item Response Theory: On the Problem of Item Chaining and Categorization of Item Scores

Tsuyoshi Izumi*, Kuramoto Naoki**

* The Japan Institute for Educational Measurement, Inc., ** TohokuUniversity

ABSTRACT

The present study tried to apply item response theory (IRT) models to preexisting science writing tests, for the sake of applying IRT models to large-scale high-stakes examinations in Japan such as those used for university admissions. The difficulty in applying IRT to science constructed-response items is that items do not usually satisfy the local independence assumption. In addition, scoring of partially correct responses is another point at issue. The present study compared several IRT models in terms of item parameter estimation. The results revealed instability in the estimates of the discrimination and difficulty parameters, especially when testlets of chained items were included in the models. The present study indicated the difficulty of applying IRT models for science constructed-response test.

Key words: science writing test, testlet, item response theory , local independence, item chaining