Bulletin of the Graduate School of Education and Human Development, Nagoya University (*Psychology and Human Development Sciences*)
2018, Vol. 65, 21 − 35. http://dx.doi.org/10.18999/nupsych.65.1.2

局所依存性をもつテストに対する項目応答理論の適用可能性

坪田彩乃[1)]　石井秀宗[2)]　野口裕之[3)]

問題と目的

項目応答理論

項目応答理論（Item Response Theory: IRT）（Lord & Novick, 1968; Hambleton & Swaminathan, 1985）とは，テストを作成・実施・分析するための数理モデルである。個々の項目ごとに受験者の能力値と正答率との数学的な関連が示されているという特徴があり（植野･荘島, 2010），発達データや学力の経年比較をするために異なる問題項目によって作成されたテストを比較するという目的から考え出された（日本テスト学会，2010）。IRTはテストで測定している能力（特性）についての受験者の能力値（特性尺度値）を，項目反応パタン（テストの項目に対する受検者の正誤についての情報）に基づき推定する。項目の統計的特性を共通尺度上で事前に推定しておくことで，受検者によって解答する項目が異なっても複数の受検者の能力値を同一尺度上で推定することが可能となる。

IRTに基づくテストを作成・実施・分析する際に満たされる必要性がある仮定の一つに，局所独立性（Local Independence）の仮定（Lord & Novick, 1968）が挙げられる。この仮定は，受検者iの能力パラメタθ_iについて，ある項目に正答（誤答）する確率が，他のある項目に正答（誤答）したことによらず一定であり，独立であるというものである。この仮定により，複数の項目に対する項目反応は，それぞれの項目についての正誤の確率の積，すなわち以下の式で表すことが可能となる。

$$P(x_1, x_2, \cdots, x_j|\theta_i) = \Pi_{j=1}^{J} P(x_j|\theta_i) \quad (1)$$

ここで，$P(x_1, x_2, \cdots, x_j|\theta_i)$は特性値$\theta_i$を持つ受検者$i$が，$J$個ある項目$x_1$, x_2, …, x_Jに対する反応確率を，$P(x_j|\theta_i)$は特性値θ_iを持つ受検者iが，j番目の項目x_jに対する反応確率を表している。

こうした局所独立性の仮定は実用場面に適用するには厳しく，非常に強い仮定であるという指摘もある（McDonald, 1981）。IRTに基づくテストの利点の一つとして，事前に項目プールを作成しておくことにより，受験者によって異なる項目を解答しても同一尺度上で受験者の能力推定値の比較を可能となることが挙げられる。予備テストで良い特性を示した，特性値が既知である項目によって成り立つものが項目プールであり，テストを作成するときには，項目プール内の項目を選択して編集することで，テスト開発者が意図した性能のテストを編集・実施することが可能である（野口・大隅，2014）。Stout（1990）は異なる項目プールに属するペア項目i, i'について共分散が0になるとき，局所独立性の仮定が満たされているとみなすことができるとしている。ある項目プール内の項目間に局所依存性が生じていたとしても，一つのテストとして成形する段階で項目プールから複数の項目を選択するとき，最終的に用いられる項目間の共分散が0であれば，局所独立性の仮定とみなすことができるということである。Stout（2002）は，このペア項目i, i'について以下の式が成り立つときを弱い局所独立性（weak local independence）と呼んでいる。

$$P(U_i = u_i, U'_{i'} = u'_{i'}|\Theta = \theta) = P(U_i = u_i|\Theta = \theta)P(U'_{i'} = u'_{i'}|\Theta = \theta) \quad (2)$$

ここで，$P(U_i = u_i, U_{i'} = u_{i'}|\Theta = \theta)$は特性値$\theta$である受験者が，$i$, i'番目のペア項目U_i, $U_{i'}$に対する反応確率を表し，$P(U_i = u_i|\Theta = \theta)$は特性値$\theta$である受検者が，$i$番目の項目$U_i$に対する反応確率を表している。つまり，ペア項目$i$, i'間で局所独立性の仮定が成り立つときを弱い局所独立性として表している。

項目間に局所独立性の仮定が成り立たないとき，局所依存性（local depedence）が生じていることになる。実際のテストで局所依存性が生じている項目を取り除くことは難しく，PISA調査でも局所依存性が生じている項目が存在していることが指摘されている（Monseur,

1）名古屋大学大学院教育発達科学研究科博士後期課程（指導教員：石井秀宗教授）
2）名古屋大学大学院教育発達科学研究科
3）名古屋大学

Baye, Lafontaine, & Valeérie, 2011)。局所依存性が生じる原因は，項目間の直接的な依存性，間接的な依存性，そして，測定対象とは異なる特性の影響という3パタンに大別することが可能である（加藤・山田・川端, 2014)。この中で，項目間の直接的な依存性ならびに間接的な依存性は，テストに含まれる項目そのものに原因を見出すことが可能である。たとえば, 数学や物理といった科目では，前の設問の解答を用いて後の設問に解答させる問題形式で出題することがある。こうした項目では直接的な依存性が生じる。また，国語や英語といった科目では，同一の文章を用いて複数の設問を解答させる問題形式で出題することがある。こうした項目では間接的な依存性が生じる。このように出題される問題形式の影響で，項目間に局所依存性が生じることとなる。

さて，こうした項目間に局所依存性が生じている項目を含むテストにIRTを当てはめるとき，我々がとる手段は以下の二つのどちらかとなる。一つは，項目の局所依存性を無視する形で，項目間に局所独立性を仮定する方法である。このとき，ある項目へ正答（誤答）することと他のある項目へ正答（誤答）することの関連性が実際にはあるにも関わらず, ないと仮定している。二つ目は，項目間の局所依存性を考慮したモデルを用いることである。しかし，項目間の局所依存性をモデル内でどのように表すのか，また，そのようなときに用いられるパラメタの意味を明確にする必要がある。更に，それらのモデルを用いた時に，得られた結果をどのように解釈するべきであるのかを十分に留意する必要も生じる。

局所依存性が生じている項目に局所独立性を仮定することによる影響は，Sireci, Thissen, & Wainer（1991）やLee（2000）でも検討されている。中でも，登藤（2012）では，局所依存性が生じている項目群について，局所依存性を考慮したモデルと局所独立性を仮定するTwo-parameter Logistic Model（2PLM）での比較検討を行っている。その結果，項目パラメタの推定では影響があるものの，能力推定では，2PLMであっても，局所依存性を考慮したモデルを用いたときと遜色のない結果となっている。そのため，項目に局所依存性が生じている場合についても，実用場面でより簡便なモデルを適用することは十分に可能であると考えられる。

テストレットモデル

Wainer & Kiely（1987） は CBT（computer based testing）の文脈から，同一の関連性をもつ複数の項目を一つの群として捉えた項目群について「テストレット」と呼んだ。日本におけるテストでは，大問形式の項目を含むテストを多く用いており，これらの項目群の中には局所独立性の仮定を置くことが困難な場合もある。こうした項目をテストレットとして扱うテストレットモデルを適用することにより，IRTでの分析が可能となる。

国内における実データを用いたモデル間の比較研究としては，石塚・前川・菊池・中畝・内田（2001）や泉・山野井・山田・白川・対馬（2013）が挙げられる。石塚他（2001）は英語科目の大問形式の項目にテストレットモデルを当てはめ，分析を行っている。その結果，能力パラメタの推定誤差はテストレットモデルの方が2PLMよりも大きくなったが，能力パラメタの推定値は両者に大きな違いが見られなかったため，推定精度の問題を除けば，どちらのモデルであっても実用上問題がないとしている。また，泉他（2013）では，数学科目の連鎖性のある項目群にテストレットモデルを当てはめ，分析を行っている。その結果，テストレットモデルと2PLMでは能力推定値の相関係数の値は大きくなったが，散布図を描くと差異が見られたとしており，極端な識別力パラメタの推定値が得られたことを原因の一端として考えている。

これらの研究では，局所依存性が生じると考えられる項目群を一塊のテストレットとして，その正答数を数え上げることによって，段階反応モデル（Graded Response Model: GRM）（Samejima, 1969）を当てはめIRTでの分析を可能としている。

GRMとは，解答が正誤の二値に限られず，二値以上の順序性を持つ段階的なカテゴリを許す多値型IRTモデルである。

このモデルでは，潜在特性値θ_iをもつ受検者が，項目jのN個のカテゴリのうちから，カテゴリkを選択する確率$P_{j,k}(\theta_i)$を以下の式で表す。

$$P_{j,k}(\theta_i) = P^{+}_{j,k}(\theta_i) - P^{+}_{j,k+1}(\theta_i) \tag{3}$$

$$P^{+}_{j,k} = \frac{1}{1+\exp\left(-1.7a_j(\theta - b_{j,k})\right)} \tag{4}$$

$$P^{+}_{0} = 1.0 \tag{5}$$

$$P^{+}_{N+1} = 0.0 \tag{6}$$

このとき，$P^{+}_{j,k}$は，項目jにおいて，k番目以上のカテゴリを選択する確率を表している。

局所依存性が生じる項目群をテストレットと見なし，その合計得点を多値型の項目として段階反応モデルを適用し扱うことは，他の局所依存性を考慮した複雑なモデルより，比較的扱いやすいと考えられる。他の複雑なモデルでは，局所依存性を表すパラメタを加えることにより，新たな推定プログラムの構築や，パラメタ自体の特別な解釈をする必要性が生じるのに対し，テストレットに段階反応モデルを当てはめる場合では該当する大問の

正答数を数え上げるという操作を行うことのみで推定を可能とする。また，IRTでの分析が可能なソフトウェアでも，GRMでの分析は複雑なプログラムを課すことなく，既存のパッケージを用いることにより可能である。

テストレットモデルでは，項目群の正答数を数え上げるため，一つのテストレットの中で異なる正誤パタンの組み合わせであっても正答数が同じならば同じ得点として扱うこととなる。しかし，項目間に直接的な関係があるとき，前の項目に誤答したことによって，連続する後の項目への正答確率が0になることがある。こうした場合，前の項目への正誤が後の項目への正誤に影響を及ぼし，一つの正答数に対し一つの正誤パタンを示す。

従来考えられてきた表面的な局所依存性（Surface local independence）は，連続する2項目について，項目1が正答（誤答）であるとき，項目2も正答（誤答）する確率を操作するものであった（Chen & Thissen, 1997）。しかし，このモデルの場合，前の項目に正答することで, 後の項目に正答する確率を高めることになる。そのため本研究では，前の項目での誤答が後の項目への正答確率には影響を及ぼすが，前の項目での正答は後の項目の正答確率へは影響を及ぼさない局所依存性について検討を行う。

簡便なモデルが必要となるテスト場面

テストには様々な目的があり, その目的に即した作成，採点，評価が必要となる。現在，IRTを適用して実施されているテストは，TOEFL，日本語能力試験などが存在する。こうしたテストは，大規模なものであり，非常にハイステークスなテストである。しかし，受験者集団の規模が小さかったり，結果によりその後の処遇が左右されないようなテストであっても，IRTを適用させることが可能な場面は存在する。

たとえば，心理テストを作成・実施・分析する場面が挙げられる。並川他（2012）では，Big-Five尺度の短縮版を作成するために，IRTによる検討を行っている。また，渡辺（2009）では，方向感覚における尺度構成を，渡辺・平林（2009）では抑うつ感の尺度構成を，IRTを用いて行っている。

既存の心理尺度についての検討も，IRTを用いて行われている。酒井・野口（2015）では，学生相談で用いるUPI，GHQ-30，K10のカットオフポイントについての妥当性の検証を，Havaei & Dahinten（2017）では，職務エンパワメントの尺度であるCWEQ IIについての信頼性・妥当性の検証を行っている。

しかし，従来のIRT研究では，受検者集団の規模が大きく，テストの専門家が関与することが可能となる場面がほとんどである。そのため，小規模な集団や，従来ではIRTを適用することの少なかった場面における適用の可否については，あまり検討されてこなかった。

本研究の目的

本研究では（a）項目間に局所独立性が成り立っているテスト，(b)項目間で局所依存性が生じているテスト，の両者で項目パラメタ値がすべて相互に等しいケースを想定して受検者の能力パラメタ推定値がどのように異なるかを検討する。これにより，局所依存性が生じているテストの能力推定の精度が項目間の局所依存性によるものであるのか，局所独立性を仮定できるテストにおいても生じる程度の誤差であるのかを検討する。加えて，簡便なIRTモデルを適用することが可能なレベルで能力パラメタ値の推定を行える条件について検討するために，困難度，受検者数の観点から推定精度への影響を検討する。どのような条件下においてIRTによる分析が可能であるかを探り，従来ではIRTによる分析を用いることが不可能と考えられていたテスト場面においての適用可能性の是非を検討する。

なお本研究では，局所独立性の仮定より逸脱する場面でのIRTの適用可否について検討を行うために，誤答の後に正答を許さない完全な局所依存性が生じている場合について取り上げる。

方法

全ての項目で局所独立の項目反応パタン発生モデル

全ての項目で局所独立を仮定する項目を12項目もつテストを仮定した。テストの項目反応モデルは2PLMを仮定する。

$$P_j(\theta_i) = \frac{1}{1+\exp\left(-1.7a_j(\theta_i - b_j)\right)} \tag{7}$$

$$Q_j(\theta_i) = 1 - P_j(\theta_i) \tag{8}$$

ここで，θ_iは受検者iの潜在特性値，a_jは項目jの識別力パラメタ，b_jは項目jの困難度パラメタをそれぞれ表している。

局所依存性が生じている項目反応パタン発生モデル

項目間に局所依存性を仮定するテストでは，誤答の後に正答が生じることがない局所依存性を置く。こうした局所依存性が生じている3項目を一塊として，大問1題と見なし，大問4題から成る12項目について，項目反応が二値型のテストを想定した。更に，テストレットモデルでの分析を行うために，局所依存性が生じている3項目を一塊のテストレットと見なし，大問1題の正答数を多値型の項目反応とするテストを想定した。

受検者の潜在特性値に応じて項目反応パタンの確率を算出する。項目に対する正答を1，誤答を0とした場合，本研究で生じる項目反応パタンと項目反応確率は，大問一題につき

$$(0,0,0) \quad Q_1(\theta) \tag{9}$$
$$(1,0,0) \quad P_1(\theta)Q_2(\theta) \tag{10}$$
$$(1,1,0) \quad P_1(\theta)P_2(\theta)Q_3(\theta) \tag{11}$$
$$(1,1,1) \quad P_1(\theta)P_2(\theta)P_3(\theta) \tag{12}$$

の4パタンになる。$P_j(\theta_i)$は潜在特性値がθ_iである受検者が項目jに正答する確率を表している。これらについて，先に述べた式（7），（8）を用いて求めた。

また，大問をテストレットとして扱うテストレットモデルでは，大問一題ごとの正答数を0〜3の4段階で表したものを項目反応パタンとして用いた。

シミュレーションの条件設定

局所独立性を仮定できるテスト，局所依存性が生じているテストのいずれも，人数，項目識別力，項目困難度を同一条件下に置き，項目反応パタンを発生させた。

人数は，100，200，500，1,000名の4パタンについて検討する。

識別力パラメタの値は，登藤（2012）のシミュレーション実験で用いられている値ならびに，石塚他（2001），泉他（2013）の実際のテストの項目識別力として報告されている値を参考に設定した。全ての項目で固定させ，識別力が低い場合（$a_j = 0.3$），中程度の場合（$a_j = 0.9$），高い場合（$a_j = 1.5$）の3パタンを検討する。

困難度パラメタは，局所依存性を仮定する項目群の3項目を一塊と考える。3項目の困難度パラメタの組み合わせ（b_1,b_2,b_3）については，一般的に考えられるような段階的に難しくなる項目困難度の組み合わせから現実場面では想定することが難しい段階的に易しくなる組み合わせとして，（–1,0,1）から（1,0,–1）にかけてb_1を0.2ずつ大きく，b_3を0.2ずつ小さくする11パタンについて検討した。つまり，（b_1,b_2,b_3）=（–1,0,1），（–0.8,0,0.8），（–0.6,0,0.6），（–0.4,0,0.4），（–0.2,0,0.2），（0,0,0），（0.2,0,–0.2），（0.4,0,–0.4），（0.6,0,–0.6），（0.8,0,–0.8），（1,0,–1）の11パタンである。これらの組み合わせについて，困難度を各値±0.25の範囲で，一様分布を用いて発生させた。

シミュレーションの手続き

1. 受検者（N = 100,200,500,1,000）の潜在特性値θ_iを標準正規分布$N(0,1)$から発生させた。
2. 困難度パラメタb_iの真値を，一様分布U(b_i–0.25, b_i+0.25)により発生させた。なお，b_i = –1, –0.8, –0.6, –0.4, –0.2,0,0.2,0.4,0.6,0.8,1の11の値をとる。
3. θ_iに応じて式（7），（8）を用いて局所独立性を仮定できるテスト，式（9）〜（12）を用いて局所依存性が生じているテストのそれぞれの項目反応パタンの確率を求めた。
4. 解答パタン作成のため，一様分布$U(0,1)$を発生させ，3. で求めた確率に応じてN名分の項目反応パタンを作成し，N×3の解答パタン行列を作成した。また，局所依存性が生じているテストでは，大問をテストレットとして扱い，能力パラメタ値の推定を可能にするために，これら3項目の正答数を数え，N×1のテストレット解答パタン行列も作成した。
5. 3, 4について大問4つ分, すなわち4回繰り返した。（計12項目，テストレット4つを作成した。）
6. iについて，局所独立性を仮定できるテストでは12項目，局所依存性が生じているテストでは12項目，テストレットモデルを用いたテストでは局所依存性が生じているテストの3項目ごとの正答数を集計した4項目からなるテストデータを発生させた。このデータにもとづいて，局所独立性を仮定できるテストには2PLM，局所依存性が生じているテストには2PLM，テストレットモデルを用いたテストにはGRMをあてはめ，最尤推定法を用いて項目パラメタ値の推定をし，その結果に基づき2PLMでは最尤推定法，GRMではベイズ推定法で潜在特性値の推定を行った。
7. 2〜6について100回繰り返した。

なお，項目反応パタンの発生および潜在特性値の推定にはR2. 9. 2（R Core Team, 2009）ならびにRizopoulos（2006）よりltmパッケージ，Partchev（2014）よりirtoysパッケージを用いた。

用いる指標

潜在特性の推定値と真値との差を検討するための指標として，Chen & Wang（2007），登藤（2010）を参考にして，$\overline{Bias}$，$\overline{RMSE}$，$\overline{cor\theta}$を用いた。

*Bias*とは，100回繰り返して算出した推定値の平均が，真値から正負どちらの方向にどのくらい離れているかを示すものである。$\overline{Bias}$とは，*Bias*のN人分の平均を表す。*RMSE*（Root Mean Square Error）とは，繰り返して算出した100回分の推定値が，真値から平均的にどのくらい離れるかを示すものである。$\overline{RMSE}$とは，*RMSE*のN人分の平均を表す。$\overline{cor\theta}$とは，潜在特性の推定値と真値との相関係数の100回分の平均を表す。

*Bias*と*RMSE*は，以下のように算出される。

$$Bias(\hat{\zeta}) = \frac{1}{100}\sum_{r=1}^{100}(\hat{\zeta}_r - \zeta) \quad (13)$$

$$RMSE(\hat{\zeta}) = \sqrt{\frac{1}{100}\sum_{r=1}^{100}(\hat{\zeta}_r - \zeta)^2} \quad (14)$$

ここで，ζはパラメタの真値を表し，$\hat{\zeta}$はパラメタζの推定量を表す。また，$\hat{\zeta}_r$はそれらのr回目の推定値をそれぞれ表す。今回，ζは各受検者の潜在特性値の真値が，$\hat{\zeta}$は各受検者の潜在特性値の推定値となる。

結果

能力推定値について，適用したモデルならびに受験者数による差を識別力ごとに検討した。

識別力が a ＝0.3のとき

はじめに，低い識別力の条件について検討を行った。項目困難度の組み合わせが段階的に下がるとき，テストレットモデルを用いたテストでは受検者数がN = 100のとき，項目パラメタの推定ができず，測定が不可能となる条件があった。

各タイプの分析モデルにおいて，受検者数が100名，200名，500名，1,000名である場合に，11パタンの項目困難度の組み合わせによって，$\overline{Bias}$，$\overline{RMSE}$，$\overline{cor\theta}$がどのように変化するかについて，Figure 1〜3の値となった。

局所独立性を仮定できるテストでは，項目困難度が段階的に難しくなる組み合わせにおいて$\overline{Bias}$の値が0付近の値を示した。$\overline{cor\theta}$は項目困難度の組み合わせによらず，安定はしていないものの一貫して低い値をとり，受検者数の大小に影響を受けない困難度条件もあった。

テストレットモデルを用いたテストでは$\overline{Bias}$について局所独立性を仮定できるテストと同様の結果であったが，項目困難度が段階的に易しくなる組み合わせでは受験者数が小さくなるにつれて正の方向に大きい値をとった。$\overline{RMSE}$は一貫して他のモデルに対し最も小さい値をとり，N = 100の値であっても，局所独立性を仮定できるテスト，局所依存性の項目に2PLMを適用したテストのN = 1000よりも小さくなった。$\overline{cor\theta}$の値も他のモデルに対し最も大きい値をとった。しかし，項目困難度が段階的に易しくなる組み合わせにおいて，段階的に難しくなる組み合わせよりも小さい値となるときがあり，受検者数が小さくなるにつれて$\overline{cor\theta}$の値も小さくなった。

局所依存性を仮定できる項目に2PLMを適用したテストでは，$\overline{Bias}$の値は一貫して負の値をとった。$\overline{RMSE}$の値は，局所独立性を仮定できるテストとの差は小さいも

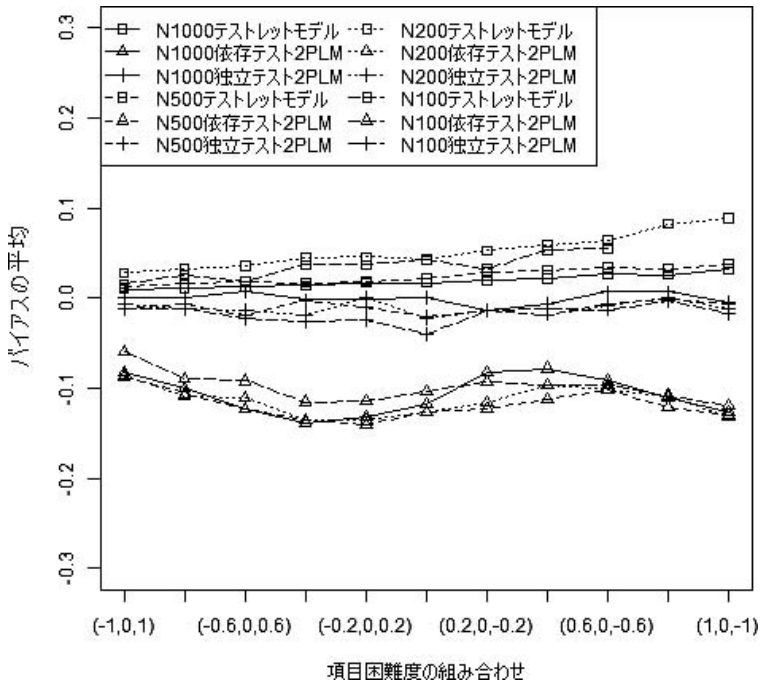

Figure 1　低識別条件における$\overline{Bias}$

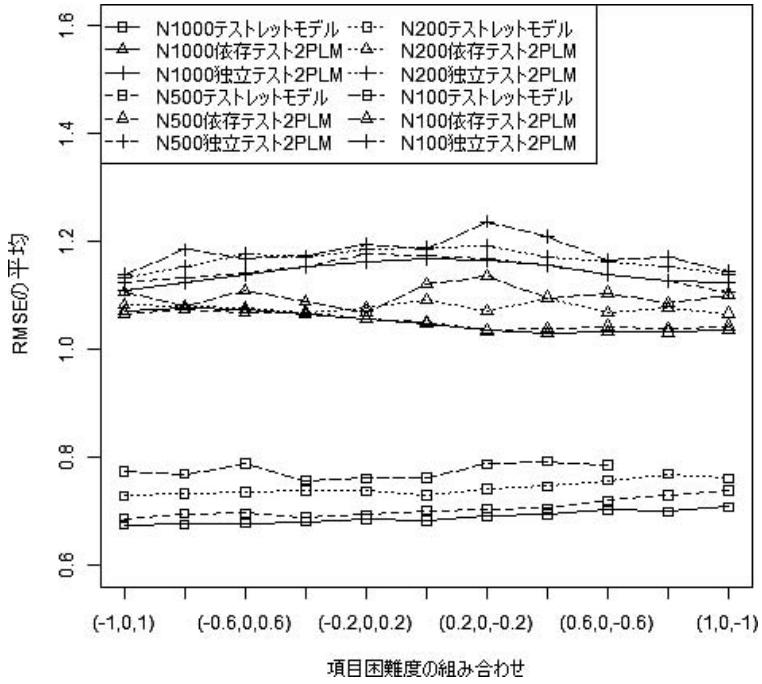

Figure 2　低識別条件における $\overline{RMSE}$

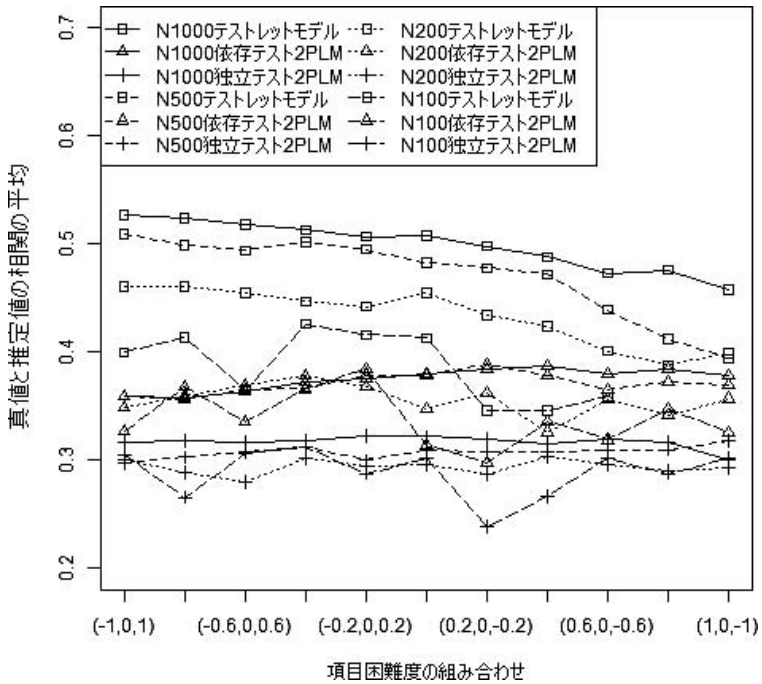

Figure 3　低識別条件における $\overline{cor\theta}$

のの，一貫して局所独立性を仮定できるテストよりも大きい値となった。$\overline{cor\theta}$はテストレットモデルを用いたテストと同様に，項目困難度が段階的に易しくなる組み合わせにおいて，段階的に難しくなる組み合わせよりも小さい値となる組み合わせもあり，更に受験者数の大小に影響を受けない条件もあった。

識別力が a ＝0.9のとき

次に，中程度の識別力の条件について検討を行った。項目困難度の組み合わせが段階的に下がるとき，テストレットモデルを用いたテストでは受検者数がN = 100, 200のとき，項目パラメタの推定ができず，測定が不可能となる条件があった。

各タイプの分析モデルにおいて，受検者数が100名，200名，500名，1,000名である場合に，11パタンの項目困難度の組み合わせによって，$\overline{Bias}$，$\overline{RMSE}$，$\overline{cor\theta}$がどのように変化するかについて，Figure 4～6の値となった。

局所独立性を仮定できるテストでは，$\overline{Bias}$の値が0付近の値を示した。$\overline{RMSE}$は項目困難度の組み合わせが同程度であるときに向かって，グラフの値が山なりに動いた。しかし，全体的に大きな値を示した。$\overline{cor\theta}$は項目困難度の組み合わせが同程度のとき，低い値となったが，0.7から0.9の高い相関を示した。

テストレットモデルを用いたテストでは，項目困難度が段階的に易しくなる組み合わせにおいて，能力パラメタ値の推定が不可能となる条件があった。$\overline{Bias}$は，項目困難度が段階的に難しくなる組み合わせであれば，局所独立性を仮定できるテストと同様に0付近の値を示した。しかし，段階的に易しくなる組み合わせにおいては，値が徐々に大きくなった。3つのモデルの中ではどの項目困難度の組み合わせにおいても，最も小さい$\overline{RMSE}$の値となったが，項目困難度が段階的に易しくなる組み合わせにおいて，その困難度の差が大きくなるにつれて$\overline{RMSE}$の値も大きくなった。$\overline{cor\theta}$は項目困難度が段階的に易しくなる組み合わせであるほど低い値となったが，0.6から0.9の高い相関を示した。

局所依存性が生じているテストでは，一貫して，$\overline{Bias}$の値が負を示しており，更に，項目困難度の組み合わせについて，段階的に難しくなるものであっても，その組み合わせの差が小さくなるとき，また，段階的に易しくなるものでは，その差が大きくなるほど，$\overline{Bias}$の値の絶対値が大きくなった。また，項目困難度の組み合わせが段階的に易しくなるとき，受検者の人数が小さいほど0から離れる値をとる傾向にあった。$\overline{RMSE}$についても，

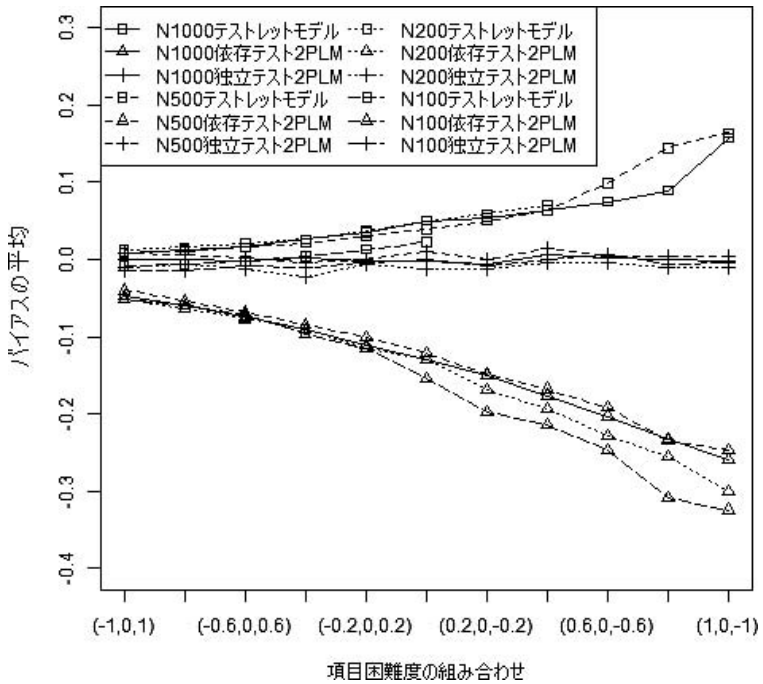

Figure 4　中識別条件における$\overline{Bias}$

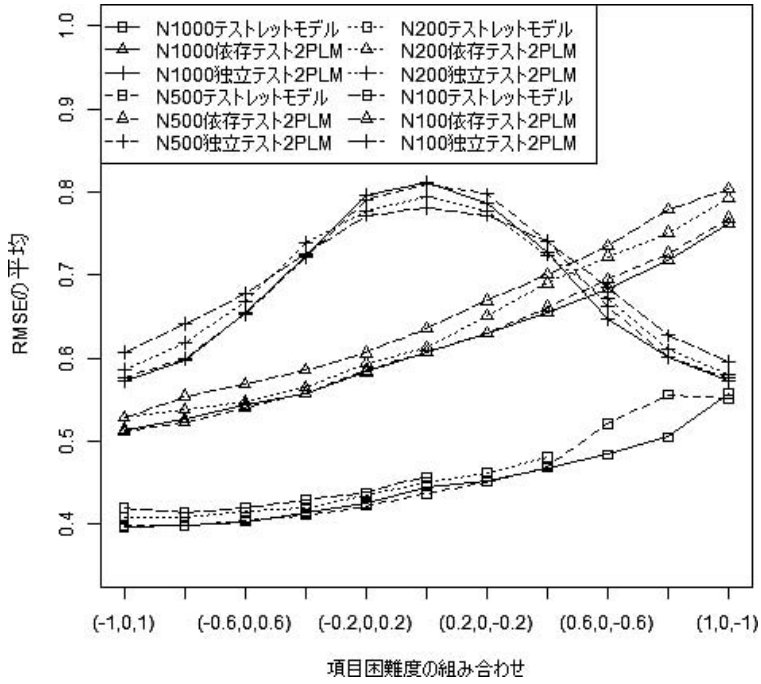

Figure 5　中識別条件における $\overline{RMSE}$

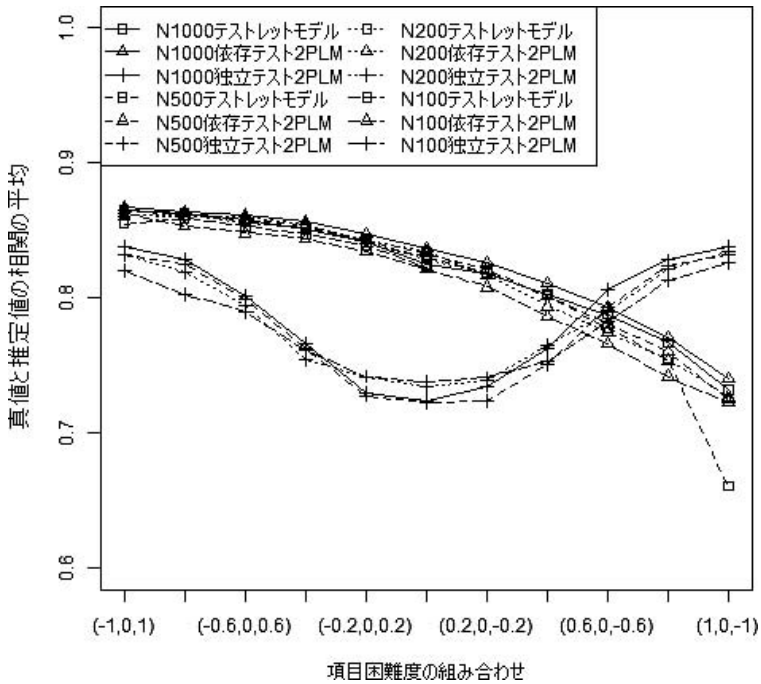

Figure 6　中識別条件における $\overline{cor\theta}$

項目困難度の組み合わせが段階的に易しくなるにつれて値が大きくなり，テストレットモデルを用いたテストと同様の動きを見せたが，$\overline{RMSE}$の値はテストレットを用いたものよりも大きくなった。$\overline{cor\theta}$は項目困難度が段階的に易しくなる組み合わせにおいて，段階的に難しくなる組み合わせよりも小さい値となるときがあった。受検者数の大小によらない困難度条件もあった。しかし，0.7から0.9の高い相関を示した。

特に受検者の人数に着目すると，テストレットモデルを用いたテスト，局所依存性が生じているテストでは，受検者数が多くなるほど強い相関になったのに対し，局所独立性を仮定できるテストでは項目困難度の差が小さい条件において，受検者数が小さいほど強い相関を示すことがあった。

識別力が a ＝1.5のとき

最後に，高い識別力の条件について検討を行った。項目困難度の組み合わせが段階的に下がるいくつかの条件において，かつ，人数が100名程度では，項目困難度の組み合わせによらず，テストレットモデルを用いたテストでは計算が収束しない場合があり，測定が不可能であった。

各タイプの分析モデルにおいて，受検者数が100名，200名，500名，1,000名である場合に，11パタンの項目困難度の組み合わせによって，$\overline{Bias}$，$\overline{RMSE}$，$\overline{cor\theta}$がどのように変化するかについて，Figure 7～9のような値となった。

局所独立性を仮定できるテストでは，$\overline{Bias}$の値が0付近の値を示した。$\overline{RMSE}$は項目困難度の組み合わせが同程度であるときに向かって，値が山なりに動いた。項目困難度の組み合わせが同程度のとき，$\overline{cor\theta}$が低い値となったが，0.8から0.95程度の高い相関を示した。

テストレットモデルを用いたテストでは，項目困難度が段階的に易しくなる組み合わせにおいて，能力パラメタ値の推定が不可能となる条件があった。$\overline{Bias}$は，項目困難度が段階的に難しくなる組み合わせであれば，0付近の値を示しており，困難度が段階的に易しくなる組み合わせにおいては，テストレットモデルを用いたテストでは$\overline{Bias}$が大きくなる傾向が見られた。計算できた範囲内では，$\overline{RMSE}$の値について，3つのモデルの中ではどの項目困難度の組み合わせにおいても，最も小さい値となった。$\overline{cor\theta}$は，項目困難度が段階的に易しくなる組み合わせであるほど低い値となったが，0.8から0.95程度の高い相関を示した。

局所依存性が生じているテストでは，一貫して$\overline{Bias}$

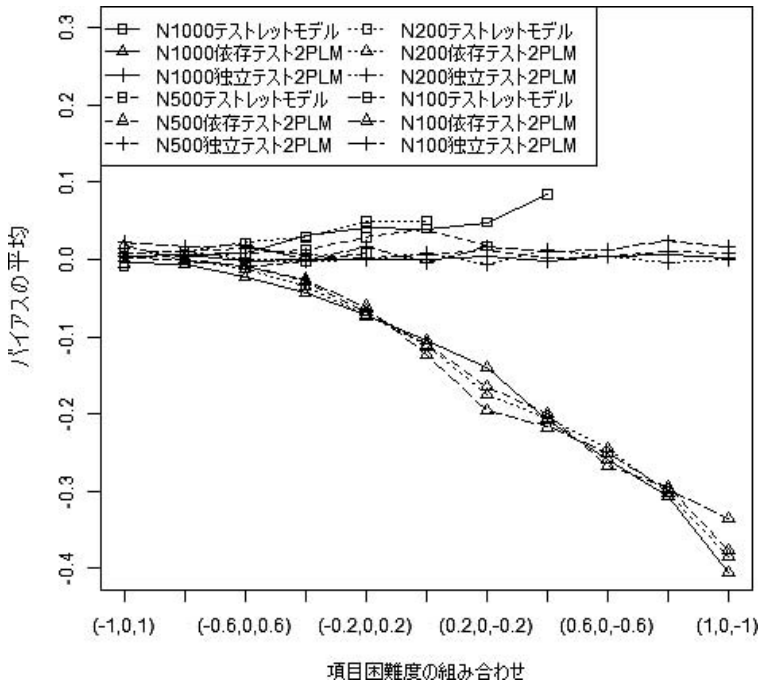

Figure 7　高識別条件における$\overline{Bias}$

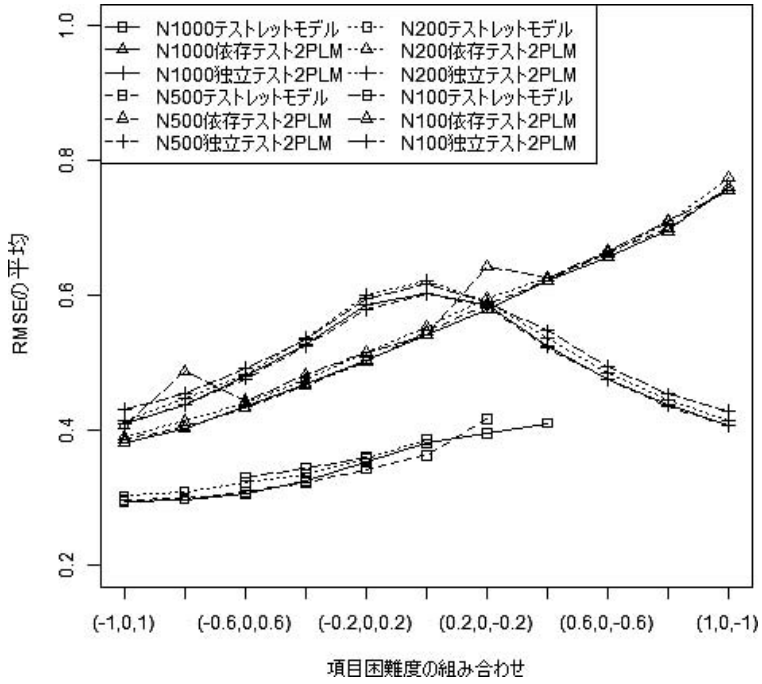

Figure 8　高識別条件における $\overline{RMSE}$

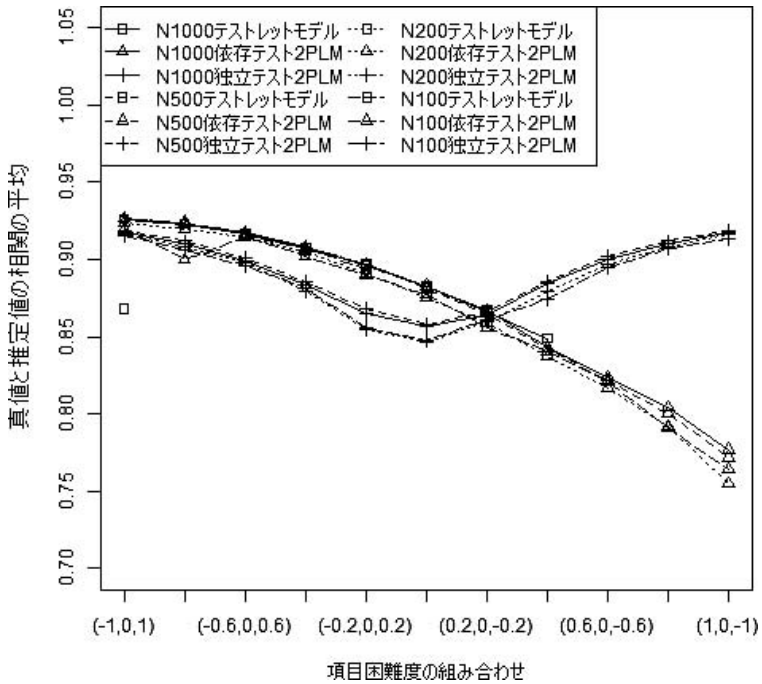

Figure 9　高識別条件における $\overline{cor\theta}$

の値が負を示しており，更に，項目困難度の組み合わせについて，段階的に難しくなるものであっても，その組み合わせの差が小さくなるとき，また，段階的に易しくなるものでは，その差が大きくなるほど，$\overline{Bias}$の値の絶対値が大きくなった。$\overline{RMSE}$は，項目困難度の組み合わせが段階的に易しくなるにつれて値が大きくなり，テストレットモデルを用いたテストと同様の動きを見せたが，テストレットを用いたものよりも大きい値をとった。$\overline{cor\theta}$は項目困難度が段階的に易しくなる組み合わせであるほど低い値となったが，0.75から0.95程度の高い相関を示した。

局所独立性を仮定できるテストと局所依存性が生じているテストでは，受検者数が大きくなるほど$\overline{RMSE}$の値が小さくなり，$\overline{cor\theta}$の値が大きくなる傾向があったのに対し，テストレットモデルを用いたテストでは受検者数による影響はみられなかった。

考察

本研究では，局所独立性を仮定できるテストと局所依存性が生じているテストという2種類のテストを想定して，2PLMとGRMを用いて能力パラメタ値の推定を行い，テストに含まれる項目の特性，受験者数を変化させ，局所依存性が生じているテストに簡便なモデルを用いたIRTの適用可能な条件について探索的に検討を行った。その結果，識別力が低いときには，テストレットモデルが他のモデルよりも推定精度が高くなったものの，全体的に推定精度は低くなった。また，受検者数による影響が大きくなった。識別力が中程度から高いときには，識別力が低い条件に比べて受検者数による影響は小さくなり，局所依存性が生じている項目間の項目困難度の組み合わせによって，局所依存性が生じているテストの能力パラメタ値の推定精度が低くなった。また，局所独立性を仮定できるテストであっても，項目困難度の組み合わせによっては能力パラメタ値の推定精度が下がった。これらのことより，項目困難度の組み合わせ，受検者数より，局所独立性を仮定できるテストへの2PLMとの比較から簡便なIRTモデルを使える条件について検討を行う。

項目困難度の組み合わせによる影響

まず，項目困難度の組み合わせによる影響について検討する。局所依存性が生じているテストでは，中程度から高い識別力において，全ての条件において，項目困難度の組み合わせが段階的に下がるほど，つまり，項目の難易度が段階的に易しくなるという条件において，系統的な影響が見受けられた。また，局所独立性を仮定できるテストでも，中程度から高い識別力において，項目困難度の組み合わせが一定の値であるほど，$\overline{RMSE}$が大きくなり，$\overline{cor\theta}$が小さくなるという影響が見受けられた。

具体的には，局所依存性を仮定するテストでは，中程度から高い識別力であるとき，テストレットモデルを用いたテスト，局所依存性が生じている項目に2PLMを適用したテストのどちらでも，$\overline{Bias}$の絶対値が大きくなり，$\overline{RMSE}$も大きくなった。また，これらのモデル間で，$\overline{RMSE}$の値に差があるものの，$\overline{cor\theta}$の値がほぼ同じ値であるということは，推定精度に差はあるものの，能力パラメタの真値と推定値の間の線型性は，どちらのモデルでも一定で保たれていることを示している。

本研究におけるシミュレーションで設定している項目間の局所依存性は，前の設問に誤答したら，それ以降の設問にも必ず誤答するというものである。本研究で用いた2PLMとGRMは，項目の困難度と受検者の潜在特性値が一致している場合，正答確率が0.5になるようにロジスティック曲線を設定する確率モデルである。受検者の潜在特性値が高くなるにつれ正答確率は1に近似し，逆に低くなるにつれ0に近似するという特徴がある。しかし，前の設問に正答しない限り，後の設問の正誤について確率モデルをあてはめないという局所依存性がある条件下では，受検者の正答確率を低く見積もることとなる。項目困難度が受検者の能力値よりも遥かに低い項目についても，確実に誤答することになる。そのため，このような局所依存性が生じている場合では，局所独立性を仮定して能力パラメタ値を適切に推定することは困難となる。

項目困難度の組み合わせが段階的に上がる項目群について，$\overline{Bias}$では大きく困難度が上がるとき，0付近の値をとり，3つのモデル間には大きな差はみられなかった。困難度の上り方が緩やかになるにしたがい，局所依存性が生じている項目に2PLMを適用したテストは局所独立性を仮定できるテストに2PLMを適用したテストよりも絶対値として大きな負の値をとり，テストレットモデルを適用したものでは正に大きな値をとった。$\overline{RMSE}$は，一貫してテストレットモデルを適用したものが0.3〜0.4程度と小さな値をとり，2PLMのものを大きく下回った。しかし，$\overline{cor\theta}$は局所依存性が生じているテストのどちらも，局所独立性を仮定できるテストよりも大きな値をとった。識別力が高いとき，困難度の開きが大きいほど，後の項目で誤答する確率が低い受検者を強く弁別することとなり，結果として真値との線形性が保たれたのだと考えられる。

また，局所独立性を仮定できるテストでは，識別力が中・高程度（a_j = 0.9,1.5）のとき，（b_1,b_2,b_3）=（0,0,0）

では$\overline{RMSE}$が大きくなり，$\overline{cor\theta}$が低くなる傾向が見られた。IRTでは，受検者の反応パタンと，各設問に対する理論的な正答確率から潜在特性の推定を行う。項目困難度b_iが全て0に近い値であることによって，$\theta = 0$から離れた受検者では，能力値の真値に差があったとしても全項目に正答／誤答といった同様の反応パタンを示すこととなり，能力値の推定精度が下がる。これにより，結果的に$\overline{RMSE}$が大きくなり，$\overline{cor\theta}$が低くなったと考えられる。

受検者数による影響

次に，受検者数が能力推定とモデル選択に及ぼす影響について検討する。受検者数が100名であるとき，高い識別力では項目困難度の組み合わせによらず，テストレットモデルでは能力推定がほとんど不可能であった。一方，同一条件のテストについて，2PLMでは能力推定を行うことが可能であった。受検者数が少なかったことにより，テストレットモデルでは項目の特徴を十分に反映するだけの項目反応パタンを得られなかった可能性が考えられる。また，識別力・項目困難度の組み合わせによらず，推定精度が安定しなかった。このことから，100名程度の受検者数では十分な測定が行えないことが示唆された。

受検者が200名以上の条件であれば，項目困難度の組み合わせが段階的に下がらない限り，テストレットモデルを用いたテスト，局所依存性が生じている項目に2PLMを適用したテストともに能力推定が可能であった。

受検者が200名以上であり識別力が低いとき，局所独立性を仮定できるテストと比較すると，テストレットモデルを用いたテスト，局所依存性が生じている項目に2PLMを適用したテストのどちらも$\overline{Bias}$の絶対値は大きくなった。$\overline{RMSE}$の値はテストレットモデルが他のモデルに対し小さい値であるものの，大きな値をとった。$\overline{cor\theta}$は，局所独立性を仮定できるテストが最も低くなっており，モデルを問わず能力パラメタの推定精度が低いことが示唆される。また，テストレットモデルを用いたテスト，局所独立性を仮定できるテストでは受検者数が増えるほど$\overline{Bias}$，$\overline{RMSE}$が0に近づき，$\overline{cor\theta}$が1に近づく傾向が見られた。

受検者数が200名以上であり識別力が中・高程度のとき，項目困難度の組み合わせが段階的に上がる項目群では，局所独立性を仮定できるテストと比較すると，テストレットモデルを用いたテスト，局所依存性が生じている項目に2PLMを適用したときのどちらも$\overline{Bias}$の絶対値は大きくなった。しかし，$\overline{RMSE}$は小さくなり，$\overline{cor\theta}$は高くなった。中程度の識別力では受検者数が500名以上，高い識別力では受検者数が200名以上であればそれぞれの指標の値に受検者数による大きな差は見られなかった。一方で，項目困難度が段階的に易しくなる組み合わせにおいては，局所独立性を仮定できるテストの方が，局所依存性が生じている項目に2PLMを適用させたテストや，テストレットモデルを用いたテストよりも推定精度が高くなった。また，受検者数が多いほど推定精度は上がったが，十分とは言えない精度であった。更に，識別力が高いとき，テストレットモデルでは受検者数を問わず測定が不可能であった。

しかしながら，局所依存性を含むテスト項目において，項目が段階的に易しくなっていくということは些か現実場面から乖離している条件設定である。そのため，現実的に考えうる徐々に難易度が上がる条件内であれば，局所依存性が生じている項目であっても，簡便なモデルを用いて能力パラメタの推定を行うことの可能性が示唆された。これについて，局所依存性を考慮しているモデルは異なるものの，登藤（2012）を支持する結果となった。

実際のテストに適用するために

本研究では，項目間に局所依存性が生じている項目を含むテストに対して，どのような条件下であれば実用場面で簡便なモデルを用いることが可能であるかについて，シミュレーションを通して検討を行った。項目困難度が段階的に下がることなく，受検者数が200名程度であれば，局所独立性を仮定できるテストとほぼ変わらない精度での能力推定が可能であることが示唆された。

しかし，本研究で得られた知見が，実際のテスト現場へも適用可能であるかどうかは，実際のテストデータを用いた更なる検討を要する。たとえば，本研究ではすべての項目について識別力パラメタが一定であると仮定したが，実際のテストでは項目ごとに識別力パラメタは異なる。そうした実際のテスト場面特有の状況を想定したとき，シミュレーションで得られた知見がどこまで適用可能であり，どういった点において乖離が生じるかを明らかにし，その上でIRTの適用可能性を唱えることが必要となる。

加えて，今回のシミュレーションで検討しているテストは，大問4題，小問12項目からなるものである。項目数が増えれば，得られる項目反応パタン数が増えることで，項目パラメタを精度よく推定するためには更に多くの受検者が必要となり，項目パラメタ値の推定精度が良くないとき，受検者の能力パラメタ値の推定にも影響を及ぼすこととなる。そのため，実際のテスト現場に適用する際には，どの程度の受検者を想定したテストであるのかを考慮した上で，テストに含める最低の項目数を検討していく必要があるだろう。

また，テストレットモデルを用いることで，局所依

存性をもつテストへのIRTの適用可能性が示されたものの，能力推定値が過小評価されるということは留意するべき点である。実際のテスト場面では，受検者の真の能力値は未知である。識別力が中程度以上であれば能力値の真値と推定値との相関は強かったとはいえ，現実場面へ適用したときには順位の入れ替わりが生じることになる。局所依存性をもつを項目にテストレットモデルを適用したとき，2PLMでの能力推定値との散布図を描くと，能力値の高い受験者において線形性が保たれない（泉他，2013）という報告があることから，一部の受検者において適用するモデルの違いにより，推定誤差が大きくなることが考えられる。しかし，本研究では受験者の能力パラメタ値の大小や局所依存性をもつ項目の項目パラメタとの関連から，どのような受検者の推定誤差が大きくなるのかという検討はなされていない。そのため，項目間に局所依存性が生じているテストでは，どのような受検者において特に推定誤差が大きくなる傾向があるのかを明らかにすることにより，推定誤差が大きくなる受検者を最小限にするための項目パラメタの設定を検討することが今度の課題として挙げられる。

なお，200名程度という人数は，例えば学校現場であれば一学年程度の人数であると考えられる。そのため，仮に学年単位での教科・科目にIRTを適用したい場合には，検討する余地はあると考えられる。しかしながら，極端に受検者数が少ないテストについては，適用が困難であると考えられる。今後は，IRTではない他のテスト理論についても検討を重ね，受検者集団が小さいテスト場面ではどのようなテスト理論が有用であるかについても検討を要する。

最後に，本研究で用いた能力パラメタ値の推定法について，2PLMでは最尤推定法，GRMではベイズ推定法と異なる手法を用いている。そのため，本研究で得られた知見が推定法の違いによるものであるのか，モデル選択の違いによるものであるのかを明確に切り離すことは難しい。この点についても，今後の課題として残された部分である。

引用文献

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

Chen, W. H. & Thissen, D. (1992). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, ***22***, 265-289.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and application*. Boston: Kluwer Nijhoff.

Havaei, F & Dahinten, V. S. (2017). How Well Does the CWEQ II Measure Structural Empowerment? Findings from Applying Item Response Theory. *Administrative Sciences*, ***2***, 1-20.

石塚　智一・前川　眞一・菊池　賢一・中畝　菜穂子・内田　照久（2001）．テストレットモデルによる英語試験問題の分析　大学入試センター紀要，***30***，21-38.

Ivailo Partchev (2014). irtoys: Simple interface to the estimation and plotting of IRT models. *R package version 0.1.7*. Retrieved from http://CRAN.R-project.org/package=irtoys

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.

加藤　健太郎・山田　剛史・川端　一光（2014）．Rによる項目反応理論　オーム社

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement gor testlet-based test scores using simulation techniques. *Journal of Educational Measurement*, ***36***, 91-112.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*.: Reading, Mass. : Addison-Wesley Publishing Co.

McDonald, R. P. (1981). The dimentionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, ***34***, 100-117.

Monseur, C., Baye, A., Lafontaine, D., & Valeérie, Q. (2011). PISA test format assessment and the local independence assumption. *IERI Monographs Series. Issues and Methodologies in Large-Scale assessments*, ***4***, 131-158.

並川　努･谷　伊織･脇田　貴文･熊谷　龍一･中根　愛･野口　裕之（2012）．Big Five尺度短縮版の開発と信頼性と妥当性の検討　心理学研究，***83***，91-99.

日本テスト学会（編）（2010）．見直そう，テストを支える基本の技術と教育　金子書房

野口　裕之・大隅　敦子（2014）．テスティングの基礎理論　研究社

Partchev, I. (2014). irtoys: Simple interface to the

estimation and plotting of IRT models. *R package version 0.1.7*. Retrieved from http://CRAN.R-project.org/package=irtoys (October 24, 2018).

R Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizopoulos, D. (2006). ltm: An R package for Latent VariableModelling and Item Response Theory Analyses, Journal of Statistical Software, *17*(5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/. (October 24, 2018).

酒井　渉・野口　裕之（2015）．大学生を対象とした精神的健康度調査の共通尺度化による比較検討　教育心理学研究，*63*，111-120.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 185-201.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimentionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Stout, W. F. (2002). Psychometrics:From practice to theory and back. *Psychometrika*, *67*, 485-518.

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*，18-28.

登藤　直弥（2012）．大問形式の問題の項目群への項目反応に対する確率モデルの比較　日本テスト学会誌，*8*，85-100.

植野　真臣・荘島　宏二郎（2010）．学習評価の新潮流　朝倉書店

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptivetesting: A case for testlets. *Journal of Educational Measurement*, *24*, 185-202.

渡辺　直登・野口　裕之（1999）．組織心理測定論―項目反応理論のフロンティア　白桃書房

渡辺　利夫（2009）．項目反応理論を用いた方向感覚の尺度構成（1）日本心理学会大会発表論文集（73），621.

渡辺　利夫・平林　輝幸（2009）．項目反応理論を用いた日常的抑うつ感の尺度構成（1）日本教育心理学会総会発表論文集（51）

ABSTRACT

Consideration on applicability of item response theory to the test items of which have local dependence

Ayano TSUBOTA, Hidetoki ISHII and Hiroyuki NOGUCHI

Item response theory (IRT) is a mathematical model for creating, conducting and analyzing test items. In this model, both item difficulties and examinee's latent trait are mesured on the same scale. Item response theory required the assumption of local independence between items. This assumption means that the response to an item is independent from responses to other items when ith examinee's ability parameter θ_i is given. However, we often conduct a test which is broken the assumption of local independence. Although complex IRT models have been proposed to solve this problem, we hope to use a simpler IRT model in real test cases if available. Testlet model is one of simpler models. A testlet is an item group items of which have relationship each other. By counting the number of correct answers in a testlet, items with local dependence can be analyzed by using GRM. Therefore it is meaningful to investigate to what extent a simpler IRT model and GRM are applicable to items with local dependence.The purpose of this study is to examine the effects of the number of examinees and the combination of item difficulties on the estimation of ability parameters in such situations. As simpler IRT models we employe the two-parameter logistic model (2PLM) and the testlet model, we assume to item response patterns items of which have local dependence. In the dependent pattern, we set that correct answers do not follow incorrect answers. The number of examinee is set to be 100, 200, 500, and 1000. Item discriminations are set at fixed values of 0.3, 0.9, and 1.5. Item difficulties are set in eleven patterns. The indices of differences between the true ability values and their estimates are $\overline{Bias}$, $\overline{RMSE}$, and $\overline{cor\theta}$. Simulation results showed the low item discrimination was precision of estimation generally poor. When the values of the item difficulties parameter decrease in a testlet, precision of estimate was also poor. When the values of the item difficulties parameter increase in a testlet, and the number of examinee was 200 the testlet model works well although it requires high item discrimination.When the values of the item difficulties parameter increase in a testlet and the number of examinee more than or equal to 500, the testlet model was applicable, while the 2PLM was acceptable. Therefore we can conclude that simpler IRT models such as 2PLM and testlet model can be applied to analyze test data items of which have local dependence if the values of the item difficulties parameter increase in a testlet and the number of examinee more than or equal to 200.

Key words: item response theory, local independence, local dependence, testlet model