兵庫教育大学　教育実践学論集　第15号　2014年３月　pp.1－12

テストレットの長さが項目反応理論のパラメタ推定に与える影響

山野井　真　児*，泉　　毅**，山　田　剛　史***

（平成25年６月18日受付，平成25年12月３日受理）

The Influence of the Length of Testlets on Parameter Estimates of the Item Response Theory

YAMANOI Shinji *, IZUMI Tsuyoshi **, YAMADA Tsuyoshi ***

The purpose of this study is to examine the accuracy of parameter estimates of IRT (Item Response theory) under the influence of the length of testlets. In this study, we examined the three features of parameter estimates by using simulation test data: (1) length of testlets, (2) strength of local dependency, (3) analysis models. We adopt Mean Difference (MD) and Root Mean Square Error (RMSE) as the indices of accuracy of parameter estimates. As a result, accuracy of parameter estimate got worse in cases where testlets was longer and analysis model that ignores local dependence was used for analyzing data that include local dependency. We concluded that when testlets of the test data is long and strength of local dependency is not weak, using analysis models that takes account of local dependency is superior in terms of parameter estimates.

Key Words : item response theory, local dependence, testlet, graded response model

１．問題と目的

１）項目反応理論のモデル　テスト分析の場面において，問題の難しさを表すために正答率，受験者の学力を表すために合計得点や偏差値が用いられることがある。このようなテスト分析について，芝 (1991)(1) は，「学力テストなどでは，いわゆる得点によって学力をあらわすが，テストの中に含まれる問題の難易によって正答数が変わるため，学力をあらわす得点も変化する」と，学力とテスト得点の分離ができない問題について述べている。

この問題を解決するための手法として，項目反応理論 (Item Response Theory: IRT) を用いたテスト分析が挙げられる。IRT について，芝 (1991)(1) は，「個々のテストの難易に依存しない尺度で各被験者の学力を推定したり，各項目の特徴を捉えたり，またテスト得点の理論分布を求めたりすることが可能になる」という利点を述べている。IRT のモデルの例として，2 パラメータ・ロジスティック・モデル (2 Parameter Logistic Model, 以下 2PLM) を式 (1) に示す。

$$P_j(\theta_i)=\frac{1}{1+\exp[-a_j(\theta_i-b_j)]} \quad (1)$$

$P_j(\theta_i)$ は，能力パラメタ θ_i を持つ受験者 i の，項目 j に対する正答確率を表す。能力パラメタは，テストで測定される能力を示す。また，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表し，これらのパラメタを項目パラメタと呼ぶ。識別力パラメタは，能力パラメタが高くなるにつれて正答確率がどの程度急激に変化するかを示す値，困難度パラメタは，2PLM の場合，正答確率が 50% であるときの能力パラメタの値である。

２）局所独立性の仮定　IRT のモデルを用いるためには，テストデータに局所独立性が仮定される必要がある。局所独立性の仮定について，豊田 (2002)(2) は，「θ_i が所与である場合には，項目反応は互いに独立である」という仮定であると説明している。Yen (1993)(3) は，局所独立性の仮定が満たされない項目の関係を局所依存 (Local Item Dependence) と呼び，局所依存が起こる状況の一つとして文脈への依存 (passage dependence) を挙げている。文脈への依存は，複数の項目が共通の文脈からなるときに起こり得る。例えば，テスト全体で測定される能力とは異なる，特定の文脈のみに関する並はずれた背景知識を持っている者がいる場合や，文脈の中の異なる項目間で，解答に用いられる情報が相互に影響している場合に局所依存が起こるというものである。文脈への依存が考えられるテストとして，英語や国語における大問形式の読解問題が挙げられる。荒井・前川 (2005)(4) は，

*　岡山大学大学院教育学研究科修了生 (Master of Education, Graduate School of Education, Okayama University)
**　東北大学大学院教育情報学教育部 (Graduate School of Educational Informatics Education Division, Tohoku University)
*** 岡山大学 (Okayama University)

日本における公的な大規模学力テストについて，大問形式による出題が多いことを指摘し，このことは，日本のテスト文化の特徴の一つであると捉えている。石塚・中畝・内田・前川 (2001)[5] は，IRT において局所独立性の仮定が前提となることを説明した上で，「我が国のように大問形式で作題された試験には，そのような項目反応理論に基づく分析が馴染まないと考えられて来た」と述べている。

局所依存性の問題を解決する方法の一つとして，大問に含まれる項目群をテストレット (Wainer & Kiely, 1987)[6] にまとめて分析するという方法がある。テストレットとは，項目を一塊の項目群としたものを指す。Wainer & Kiely (1987)[6] は，局所独立性の仮定を満たさない項目群をテストレットとして，項目群の合計点を反応データとした上で多値型モデルによる分析を行うことを提案している。多値型モデルの一つである，段階反応モデル (Graded Response Model: GRM, Samejima, 1969)[7] について，式 (2) から式 (3) に示す。

$$P_{jk}(\theta_i)=P^*_{jk}(\theta_i)-P^*_{jk+1}(\theta_i) \tag{2}$$

$$P^*_{jk}(\theta_i)=\frac{1}{1+\exp[-a_j(\theta_i-b^*_{jk})]} \tag{3}$$

$P_{jk}(\theta_i)$ は，能力パラメタが θ_i である受験者 i が，項目 j を含むテストレットにおいて k 個の項目に正答する確率を表している。また，$P^*_{jk}(\theta_i)$ は，能力パラメタが θ_i である受験者 i が，項目 j を含むテストレットにおいて k 個以上の項目に正答する確率を表している。a_j は，項目 j を含むテストレットにおける識別力パラメタであり，b^*_{jk} は，項目 j を含むテストレットにおいて k 個以上の項目に正答することに対する困難度パラメタである。

3）局所依存性がある場合のパラメタの推定誤差の問題

2PLM のように局所依存性を考慮しないモデルを用いる場合，データに局所依存性があるとき，局所独立性が満たされる場合と比較して，能力パラメタの推定誤差が大きくなることが報告されている (Bradlow, Wainer & Wang, 1999[8]; 登藤 , 2010[9])。また，識別力パラメタ・困難度パラメタの推定誤差が大きくなることが報告されている (Chen & Wang, 2007)[10]。特に，局所依存性を持つ項目と局所独立の仮定を満たす項目を両方含むデータに対し局所依存性を考慮せず分析した場合，局所依存性を持つ項目の識別力パラメタは過大推定され，局所独立性を満たす項目の識別力パラメタは過小推定されることが示されている (Tuerlinckx & De Boeck, 2001)[11]。

しかし，局所依存性を考慮するモデルを用いる場合においても，データに局所依存性があるとき，能力パラメタや項目パラメタの推定誤差が大きくなることが報告されている。DeMars (2006)[12] や登藤 (2012a)[13] では，データに局所依存性が見られる場合に，局所依存性を考慮するモデルから得られる能力パラメタの推定誤差について，局所依存性を考慮しないモデルから得られる推定誤差と比較し，大きな差が見られなかったことを報告した。また，登藤 (2012b)[14] では，データに局所依存性があるとき，困難度パラメタの推定誤差について，局所依存性を考慮するモデルと考慮しないモデルとの間に大きな差が見られなかったことを示している。例えば DeMars (2006)[12] では，受験者数 2000，項目数 25 の場合では 2PLM の能力パラメタの推定誤差が 0.23, GRM では 0.25 となり，GRM の推定誤差がわずかに大きくなったことを示している。このことについて DeMars (2006)[12] は，テストレットに含まれる項目群の合計点を分析の対象とすることで，情報量が少なくなったことを一つの要因として考察している。DeMars (2006)[12], 登藤 (2012a)[13], 登藤 (2012b)[14] の検討した条件において，局所依存性を考慮するモデルと局所依存性を考慮しないモデルとで，能力パラメタ，困難度パラメタの推定誤差の大きさが同程度であることが示された。また，局所依存性を考慮するモデルを用いた場合に，能力パラメタの推定誤差がわずかに大きくなる可能性があることが示唆された。したがって，困難度パラメタの推定誤差が大きく変わらず，さらに能力パラメタの推定誤差がより小さくなることを根拠に，テストに局所依存性が想定される場合であっても局所依存性を考慮しないモデルを用いるほうが推定精度の観点で優れているという可能性がある。ただし，このことは先行研究で検討されているシミュレーションの条件の範囲においてのみ考えられるものである。テストの性質によっては，局所依存性を考慮するモデルを用いたほうが能力パラメタや項目パラメタの推定誤差が小さくなる場合も考えられる。しかし，どのような場合に局所依存性を考慮するモデルを用いると，より推定誤差が小さくなるのかはこれまで明らかになっていない。このことは，テスト分析に用いるモデルを選択する場面での一つの問題として挙げられる。また，シミュレーションの条件によって，局所依存性を考慮するモデルを用いた場合のパラメタの推定誤差が特に大きくなる場合が無いかどうかについても確認する必要がある。したがって，先行研究に加え，より詳細な条件を設けた上で能力パラメタや項目パラメタの推定誤差について検討することが重要であると考えられる。

先行研究で検討された条件において不十分であると考えられる点として，テストレットに含まれる項目数 (テストレットの長さ) が最大で 5 項目であったことが挙げられる。実際のテストデータにおいては，テストレットが 5 項目より多い項目数にまとめられる場合がある。例えば，Wainer & Wang (2000)[15] は TOEFL の問題について，リーディング問題は 13 項目からなるテストレット，リスニング問題は 5 または 10 項目からなるテストレット

にまとめられたことを報告した。また，石塚他 (2001)[5] は，2000 年度の大学入試センター試験の英語の試験問題について 2 項目から 8 項目からなるテストレットにまとめて GRM による分析を行った。したがって，現実のテスト場面では，5 項目以上からなるテストレットが用いられる場合があることから，5 項目より多いテストレットの長さがパラメタの推定誤差の大きさに与える影響について検討する必要があると考えられる。

また，登藤 (2012a)[13]，登藤 (2012b)[14] ではテストレットが長くなるにつれて能力パラメタ，項目パラメタの推定誤差が大きくなることが示されている。しかし，登藤 (2012a)[13]，登藤 (2012b)[14] では，テストレットの長さの条件を変化させると同時に，テスト全体に含まれる局所依存性を持つ項目数も変化していた。そのため，これらの研究で能力パラメタ，項目パラメタの推定誤差が大きくなったのは，テストレットが長いことによるものか，テスト全体に含まれる局所依存性を持つ項目数が多いことによるものかが区別できない。よって，テスト全体に含まれる局所依存性を持つ項目数を統制した上で，テストレットの長さが IRT のパラメタの推定誤差の大きさに与える影響について検討する必要があると考えられる。

4）本研究の目的　石塚他 (2001)[5] は，大問形式によるテストについて，「選ばれたテーマによって出来不出来の個人差が決まってしまう」こと，「最初の設問への正誤によって続く設問への解答が誘導され易い」という問題を挙げる一方で，「断片的な知識だけでなく，思考力を図るのにも適した形式である」という点を指摘している。思考力について中央教育審議会 (2013)[16] は，高等教育段階で培うべき要素としており，「知識や技能を活用して複雑な事柄を問題として理解し，答えのない問題に解を見出していくための批判的，合理的な思考力をはじめとする認知的能力」を育むことが重要であると述べている。以上のことを踏まえると，今後，思考力を測定するテストが研究場面や実践場面で多くなることが想定される。このとき，大問形式のテストが実施された場合に，どのような分析モデルを適用するかが問題となる。

本研究の目的は，テスト全体に含まれる局所依存性を持つ項目数を統制し，テストレットの長さの条件によって，局所依存性を考慮する場合と考慮しない場合との間で，項目パラメタや能力パラメタの推定誤差の大きさに差が見られるかを検討することである。本研究の意義として，大問形式のテストに対して IRT による分析を適用する際において，テストレットの長さがどの程度の場合において局所依存性を考慮するモデルを用いることが IRT のパラメタの推定誤差の大きさという観点から有効であるのかという知見が得られることが挙げられる。このことにより，思考力や読解力を問う大問形式のテストを実施するという実践場面において，より推定精度の高い分析モデルを選択できることが本研究の意義の一つであると考えられる。

2．方法

1）データ生成のモデル　本研究では，モンテカルロ法を用いたシミュレーションによって，能力パラメタ，項目パラメタの推定精度について検討する。モンテカルロ法は解析的に解を求めることが難しい問題においても実証的に解が求められることや，パラメタの値を操作することで，複数の要因について検討できるという利点がある (Harwell, Stone, Hsu & Kirisci, 1996)[17]。ただし，Harwell et al. (1996)[17] はモンテカルロ法の欠点として，シミュレーションモデルの条件がどれだけ現実的であるかによって結果が変わることを挙げている。

シミュレーションによりデータを生成する際に，テストレットの長さと局所依存性の強さについて複数の条件をおいた。シミュレーションの条件については，登藤 (2012a)[13] および Zu & Liu (2010)[18] を参考に決定した。また，本研究では，受験者数 1000 人，40 項目のテストを想定したシミュレーションを行った。40 項目のうち，20 項目が局所依存性を持つ項目の数であるとし，ベイズ変量効果モデル (Bayesian Random Effects Model: BREM, Bradlow, Wainer & Wang, 1999)[8] にしたがってデータを生成した。

BREM は，次の式 (4) で表わされる。

$$P_j(\theta_i)=\frac{1}{1+\exp[-a_j(\theta_i-b_j-\gamma_{id(j)})]} \tag{4}$$

$P_j(\theta_i)$ は，受験者 i の，項目 j に対する正答確率を表す。また，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表し，これらのパラメタを項目パラメタと呼ぶ。θ_i は受験者 i の能力パラメタを表す。$\gamma_{id(j)}$ は，θ_i とは別の受験者 i の能力パラメタを表す。この能力パラメタは，項目 j が含まれる大問 $d(j)$ のみに関する能力を表す。

$\gamma_{id(j)}$ は事前分布として，

$$\gamma_{id(j)} \sim N(0, \sigma^2_{\gamma_{id(j)}}) \tag{5}$$

が仮定される。$\sigma^2_{\gamma_{id(j)}}$ は，大問 $d(j)$ における局所依存性の強さを示す。$\sigma^2_{\gamma_{id(j)}}=0$ の場合，大問 $d(j)$ に対する $\gamma_{id(j)}$ が全ての受験者に対して 0 となる。このとき，大問 $d(j)$ に含まれるすべての項目で局所独立性の仮定が満たされる。

生成されたデータに対し，局所依存性を考慮しないモデルとして 2PLM，局所依存性を考慮するモデルとして GRM を適用してパラメタの推定を行った。

局所依存性を考慮しないモデルとして 2PLM を用いたのは，本研究の先行研究である Tuerlinckx & De Boeck (2001)[11]，登藤 (2012a)[13]，登藤 (2012b)[14] が用いているモデルであり，これらの研究結果との比較を可能にするためである。

また，局所依存性を考慮するモデルとしてGRMを用いたのは，先行研究であるDeMars (2006)[12]や登藤(2012a)[13]が用いているモデルであり，結果の比較を行うためである。また，IRTの分析ソフトウェアであるPARSCALE (Muraki & Bock, 1997)[19]やMULTILOG (Thissen, 1991)[20]，Rのltmパッケージで GRMに対応していることから，BREMよりもGRMがテストデータの分析の場面でより広く用いられていると考えられるためである。

これら2つのモデルのそれぞれから得られるパラメタの推定値とパラメタの真値を比較し，パラメタの推定誤差の大きさについて検討する。

2）シミュレーションモデル　データの発生にはR ver. 2. 15. 2を用いた。また, 2PLM, GRMにおける項目パラメタ，能力パラメタの推定にはIRTPRO ver. 2. 1 (Cai, Thissen & du Toit, 2011)[21]を用いた。

シミュレーションに用いるデータは，BREMにより生成した。式(4)および式(5)より，BREMからデータを生成する際に必要となるパラメタはa_j，b_j，θ_i，$\gamma_{id(j)}$，$\sigma^2_{\gamma_{id(j)}}$である。能力パラメタは標準正規分布にしたがい，1000人分生成した。困難度パラメタは標準正規分布にしたがって40項目分，識別力パラメタは一様分布U(0.5,2.5)にしたがって40項目分，生成した。

識別力パラメタについては，BREMの真値を2PLMの分析から得られた推定値と比較可能にするため，Ip (2010)[22]に示される項目パラメタの変換を行った。Ip (2010)[22]は，BREMの識別力パラメタを次式(6)に示す係数λを乗じることにより，2PLMの識別力パラメタと比較可能にする方法を示している。

$$\lambda=\frac{1}{\sqrt{(\frac{16\sqrt{3}}{15\pi})^2 a_j^2\sigma^2_{\gamma_{id(j)}}+1}} \tag{6}$$

なお，BREMと2PLMの困難度パラメタは比較可能であることがIp (2010)[22]により示されている。

テストレットパラメタ$\gamma_{id(j)}$は，$N(0,\sigma^2_{\gamma_{id(j)}})$から，テストレット1つあたり1000人分生成した。

3）シミュレーション条件　局所依存性の強さを表す$\sigma^2_{\gamma_{id(j)}}$の値について, $\sigma^2_{\gamma_{id(j)}}$=0.2,0.8,1.4の3条件を設定した。これらの値は, Li, Bolt & Fu (2005)[23], Li, Bolt & Fu (2006)[24]が実際の大問形式の読解テストの分析から得た$\sigma^2_{\gamma_{id(j)}}$の値の範囲にある。これらの研究から得られた$\sigma^2_{\gamma_{id(j)}}$の最小値は0.13，最大値は2.1である。

また，テストレットの長さTについて，T=2であるテストレットが10ある条件，T=5であるテストレットが4つある条件，T=10であるテストレットが2つある条件の3条件を設定した。

なお，すべての条件において，局所依存性を持つ項目数は20，局所独立性を満たす項目数は20である。

テストレットの局所依存性の強さの3条件，テストレットの長さの3条件を組み合わせ，9つの条件におけるシミュレーションデータを生成する。それぞれの条件のシミュレーションデータを50回生成し，パラメタの推定を行った上で，それぞれのパラメタの推定誤差の大きさについて検討を行う。

4）パラメタの推定誤差の指標　それぞれのデータに対して，2つの分析方法を適用する。

1つ目の分析方法は，40項目全てに対して2PLMによる分析を行うものである。これを2PLM単一分析と呼ぶ。

2つ目の分析方法は，局所独立性の仮定が満たされる20項目については2PLMによる分析を行い，テストレットに含まれる20項目については，反応データをテストレットに関して合計し，GRMにより分析を行う。この分析を，2PLM+GRM分析と呼ぶ。

2PLM単一分析と2PLM+GRM分析のそれぞれから求めた能力パラメタの推定値と真値について，また，項目パラメタの推定値と真値について，*MD* (Mean Difference)と*RMSE* (Root Mean Square Error)を指標として求める。*MD*はパラメタの推定値の過大推定または過小推定の程度を示し，正の値である場合に過大推定の傾向があることを，負の値である場合に過小推定の傾向があることを示す。*RMSE*はパラメタの推定の誤差を示し，値が高くなるほどパラメタの推定の誤差が大きいことを示す。

*MD*と*RMSE*について，それぞれ式(7)，式(8)に示す。

$$MD(\boldsymbol{\lambda})=\frac{1}{50}\sum_{r=1}^{50}(\hat{\boldsymbol{\lambda}}_r-\boldsymbol{\lambda}_r) \tag{7}$$

$$RSME(\boldsymbol{\lambda})=\sqrt{\frac{1}{50}\sum_{r=1}^{50}(\hat{\boldsymbol{\lambda}}_r-\boldsymbol{\lambda}_r)^2} \tag{8}$$

$\boldsymbol{\lambda}$は能力パラメタまたは項目パラメタのベクトルを示す。$\hat{\boldsymbol{\lambda}}_r$は$r$回目の推定で得られた能力パラメタまたは項目パラメタの推定値のベクトルである。また，$\boldsymbol{\lambda}_r$はr回目の推定における能力パラメタまたは項目パラメタの真値のベクトルである。

本研究では，パラメタの推定値について，真値と同じ平均や分散を持つように標準化することは行っていない。これは，パラメタの推定値の誤差の大きさや方向性について，真値と比較することにより検証を行うためである。

能力パラメタの*MD*，*RMSE*については，受験者数について平均をとったものを平均*MD*，平均*RMSE*として能力パラメタの推定誤差の大きさの指標とする。

また，局所独立性を満たす項目の困難度パラメタ，識別力パラメタについては，それぞれ20項目の平均*MD*，平均*RMSE*を求め，項目パラメタの推定誤差の大きさの指標とする。

本研究では，局所依存性を持つ項目の困難度パラメ

タ，識別力パラメタの推定誤差の大きさについて，分析モデル間の比較を行っていない。2PLM+GRM 分析では，局所依存項目に対し，反応データを多値データとしてまとめた上で GRM を適用しているため，BREM や 2PLM から得られる真値との比較を行うことができないためである。したがって，局所依存項目の識別力パラメタ，困難度パラメタの推定誤差の大きさについては，2PLM 単一分析の結果のみを示す。

5）シミュレーションの手順　本研究のシミュレーションの手続きは，以下のようにまとめられる。

Step1.　能力パラメタ, 困難度パラメタ, 識別力パラメタ，$\gamma_{id(j)}$ のそれぞれの真値を乱数から生成する。$\gamma_{id(j)}$ は局所依存性の強さの条件にしたがい，$N(0,\sigma^2_{\gamma_{id(j)}})$ から生成する。

Step2.　生成されたパラメタの真値をもとに，局所独立性を満たす 20 項目は 2PLM にしたがい，局所依存性を持つ 20 項目は BREM にしたがい，正答確率行列 $\boldsymbol{K}$ を生成する。

Step3.　$U(0,1)$ から一様乱数を生成し，正答確率行列 $\boldsymbol{K}$ と等しい要素数を持つ一様乱数行列 $\boldsymbol{L}$ を生成する。

Step4.　行列 $\boldsymbol{K}$，$\boldsymbol{L}$ の各要素を比較し，受験者 i の項目 j に対する応答が，$k_{ij} \geq l_{ij}$ ならば 1，$k_{ij} < l_{ij}$ ならば 0 とおいた項目反応行列 $\boldsymbol{M}$ を生成する。このとき，項目反応行列 $\boldsymbol{M}$ の各要素の 1 は正答を, 0 は誤答を意味する。

Step5.　Step 4. で生成された項目反応行列 $\boldsymbol{M}$ のうち，テストレットとおいた 20 項目の各要素について，テストレットごとに合計し，項目反応行列 $\boldsymbol{N}$ を生成する。

Step6.　項目反応行列 $\boldsymbol{M}$ を用いて 2PLM 単一分析，項目反応行列 $\boldsymbol{N}$ を用いて 2PLM+GRM 分析を行い，それぞれの項目反応行列から能力パラメタ，項目パラメタの推定を行う。

Step7.　Step1. から Step6. までを 50 回繰り返す。

Step8.　局所依存項目の識別力パラメタについて，Ip (2010)[(22)] の方法を用いて BREM における識別力パラメタの真値を 2PLM における識別力パラメタの推定値と比較可能になるように変換する。

Step9.　50 回分の 2PLM 単一分析と 2PLM+GRM 分析から得られたパラメタの推定値と，パラメタの真値を比較し，能力パラメタ，局所独立項目の項目パラメタ，2PLM 単一分析の局所依存項目の項目パラメタのそれぞれにおいて平均 *MD*，平均 *RMSE* を算出する。

以上の過程から得られた，各条件における平均 *MD*, 平均 *RMSE* から，テストレットの長さがパラメタの推定誤差の大きさに与える影響について検討する。

3．結果

2PLM 単一分析と 2PLM+GRM 分析から得られた平均 *MD* について，図 1 に示す。以下に示す図において，局所依存性の強さの条件について，σ^2 として示している。σ^2=0.2 の場合，局所依存性の強さにおいて $\sigma^2_{\gamma_{id(j)}}$=0.2 の条件であることを示す。

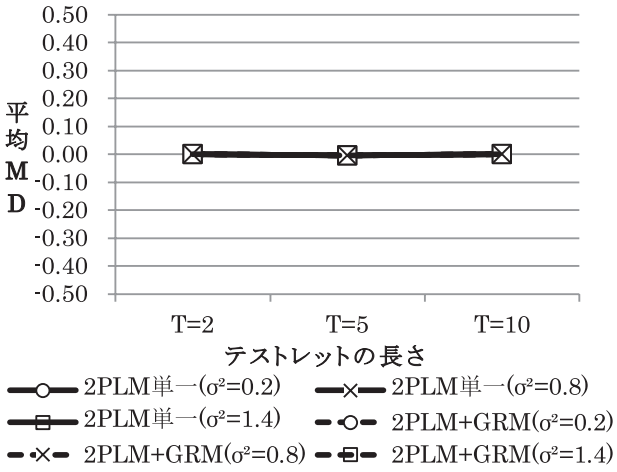

図 1　能力パラメタの平均 *MD*

図 1 より，全ての条件について能力パラメタの平均 *MD* は，ほぼ 0 であることが示された。能力パラメタの推定において，局所依存性の強さやテストレットの長さによらず，過大推定や過小推定の傾向は示されなかった。

2PLM 単一分析と 2PLM+GRM 分析から得られた能力パラメタの平均 *RMSE* について，図 2 に示す。

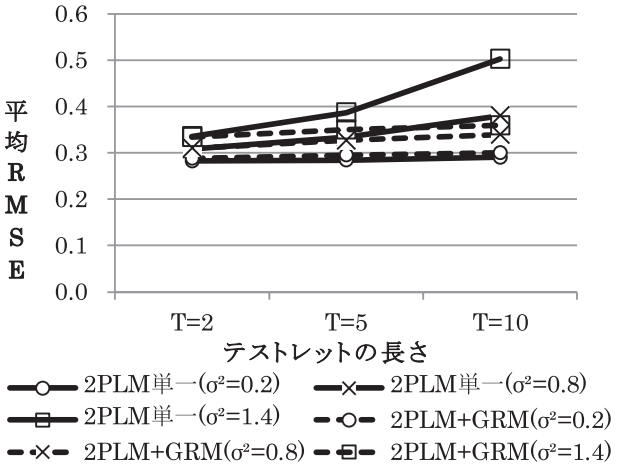

図 2　能力パラメタの平均 *RMSE*

図 2 より，局所依存性が強くなるほど平均 *RMSE* が大きくなることが示された。また，$\sigma^2_{\gamma_{id(j)}}$=0.8 以上の場合，テストレットが長くなるほど平均 *RMSE* が大きくなることが示された。特に, 2PLM 単一分析においてテストレットが長い場合に能力パラメタの平均 *RMSE* が大きくなった。

テストレットの長さが 2 項目である場合，全ての局所依存性の強さの条件において 2PLM+GRM 分析の平均 *RMSE* が 2PLM 単一分析の平均 *RMSE* とほぼ同じ値を示している。テストレットの長さが 2 項目かつ $\sigma^2_{\gamma_{id(j)}}$=0.2 である場合の平均 *RMSE* は，2PLM 単一分析において 0.282，2PLM+GRM 分析の平均 *RMSE* において 0.288 という値が得られた。また，テストレットの長さが 2 項目かつ $\sigma^2_{\gamma_{id(j)}}$=1.4 の場合の平均 *RMSE* は，2PLM 単一分析において 0.336，2PLM+GRM 分析において 0.334 という値が得られた。テストレットの長さが 2 項目の場合においては，2PLM 単一分析と 2PLM+GRM 分析の能力パラメタの推定誤差の大きさはほぼ変わらないことが示唆される。

テストレットの長さが 10 項目である場合，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合の平均 *RMSE* において，2PLM+GRM 分析から得られた値が 2PLM 単一分析から得られた値と比較して，0.144 小さい値が得られた。

テストレットが長くなるにつれて，また，局所依存性が強くなるにつれて，2PLM+GRM 分析の能力パラメタの平均 *RMSE* は 2PLM 単一分析から求めるより値が小さくなる傾向が見られた。

2PLM 単一分析と 2PLM+GRM 分析から得られた局所独立項目の識別力パラメタの平均 *MD* を図 3 に示す。

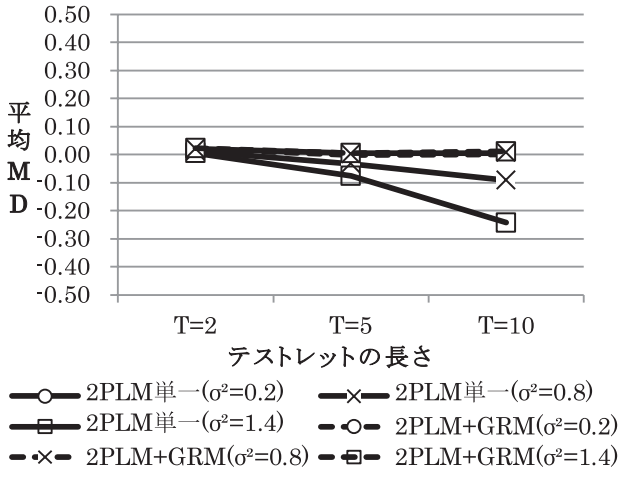

図 3　局所独立項目の識別力パラメタの平均 *MD*

図 3 より，2PLM 単一分析の場合，テストレットが長く，局所依存項目の局所依存性が強い場合に，局所独立項目の識別力が過小推定される傾向があることが示された。一方，2PLM+GRM 分析においては，テストレットが長く，局所依存項目の局所依存性が強い場合においても，局所独立項目の識別力パラメタについて過小推定される傾向は示されなかった。

なお，$\sigma^2_{\gamma_{id(j)}}$=1.4 かつテストレットの長さが 10 の場合の 2PLM 単一分析においては，局所独立項目の識別力パラメタの真値が大きくなるにつれて過小推定の程度が大きくなることが確認された。識別力パラメタの真値と識別力パラメタの *MD* の散布図について，図 4 に示した。

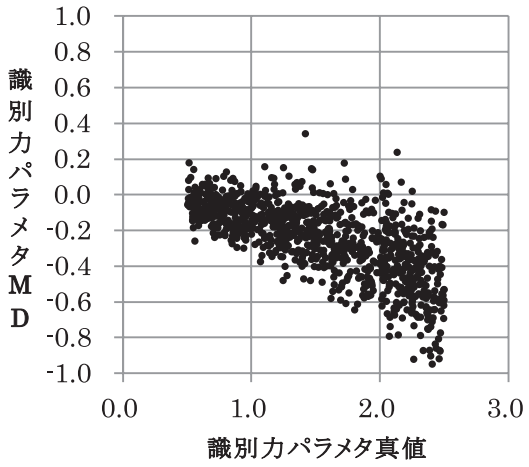

図 4　局所依存性の強さ 1.4，テストレットの長さ 10 の 2PLM 単一分析の，識別力パラメタ真値と *MD* の散布図

識別力パラメタが 1 未満のときの平均 *MD* は -0.08，1 から 2 のときの平均 *MD* は -0.22，2 以上のときの平均 *MD* は -0.42 となった。一方，2PLM+GRM 分析や，テストレットの長さが 2，5 のときの 2PLM 単一分析の場合においてはこのような傾向は見られず，識別力パラメタの真値によらず *MD* は 0 に近い値をとった。

2PLM 単一分析と 2PLM+GRM 分析から得られた局所独立項目の識別力パラメタの平均 *RMSE* を図 5 に示す。

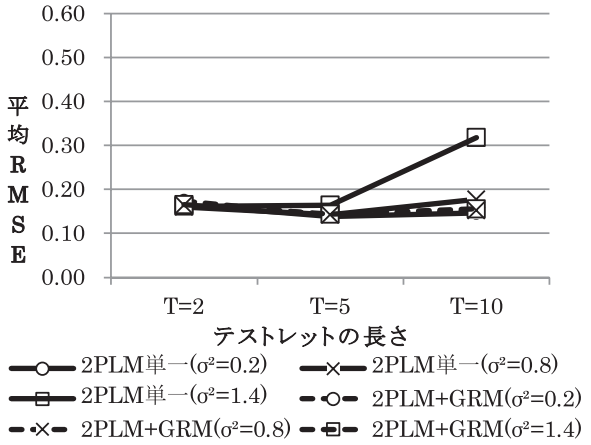

図 5　局所独立項目の識別力パラメタの平均 *RMSE*

図 5 より，局所依存性の強さの条件が $\sigma^2_{\gamma_{id(j)}}$=1.4 かつテストレットの長さが 10 の場合の 2PLM 単一分析において局所独立項目の識別力パラメタの平均 *RMSE* が大きくなることが示された。2PLM+GRM 分析と比較すると，2PLM 単一分析から得られた平均 *RMSE* は約 0.163 大きい値を示している。

一方で，局所依存性の強さの条件が $\sigma^2_{\gamma_{id(j)}}$=0.8 以下，またはテストレットの長さが 5 項目以下の条件において

は，全ての条件の平均 *RMSE* が 0.160 に近い値を取っており，2PLM+GRM 分析と 2PLM 単一分析との平均 *RMSE* の値に大きな差が見られなかった。

局所独立項目の識別力パラメタの推定誤差の大きさについてまとめる。2PLM 単一分析ではテストレットが長く，局所依存項目の局所依存性が強い場合に識別力パラメタの過小推定が確認された。また，2PLM 単一分析の場合，2PLM+GRM 分析と比較して，テストレットの長さが 10, $\sigma^2_{\gamma_{id(j)}}$=1.4 の時に平均 *RMSE* が大きい値を示した。一方で，2PLM+GRM 分析はテストレットの長さの条件によらず，局所独立項目について識別力パラメタの過小推定の傾向を示すことはなかった。また，2PLM+GRM 分析から求められる平均 *RMSE* についてはテストレットの長さに関わらず，ほぼ同じ値を示した。

2PLM 単一分析から得られた局所依存項目の識別力パラメタの平均 *MD* を図 6 に示す。

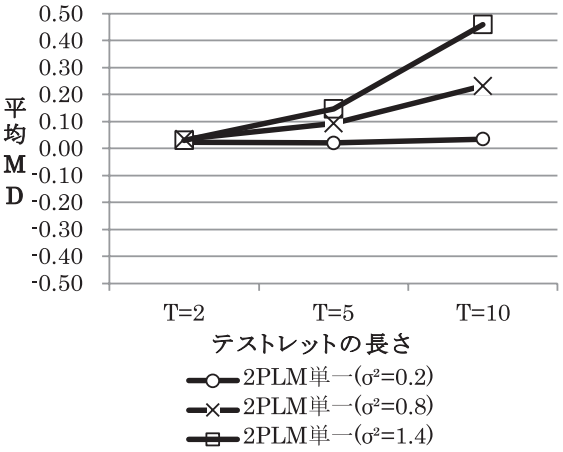

図 6　局所依存項目の識別力パラメタの平均 *MD*

図 6 より，$\sigma^2_{\gamma_{id(j)}}$=0.8 以上の条件において，テストレットが長くなるにつれて，局所依存項目の識別力パラメタの過大推定の程度が大きくなる傾向が示された。$\sigma^2_{\gamma_{id(j)}}$=0.8 の場合，テストレットの長さが 5 項目のときに平均 *MD* は 0.093, 10 項目のときに 0.231 となった。また，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合，テストレットの長さが 5 項目であるときに平均 *MD* は 0.147, 10 項目であるときに 0.460 となった。

ただし，$\sigma^2_{\gamma_{id(j)}}$=0.2 の場合は，テストレットの長さが 10 項目の条件においても識別力パラメタの過大推定の程度は他の条件と比較して特に高くならなかった。また，テストレットの長さが 2 項目である場合，局所依存性の強さが異なる場合においても平均 *MD* の値に比較的大きな差が見られず，最小で 0.020，最大で 0.035 であった。

2PLM 単一分析から得られた局所依存項目の識別力パラメタの平均 *RMSE* を図 7 に示す。

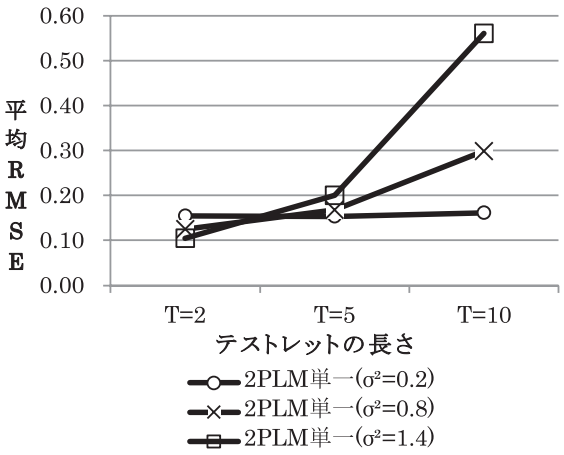

図 7　局所依存項目の識別力パラメタの平均 *RMSE*

図 7 より，$\sigma^2_{\gamma_{id(j)}}$=0.8 以上の条件において，テストレットが長くなるにつれて，局所依存項目の識別力パラメタの平均 *RMSE* が大きくなる傾向が示された。特に，テストレットの長さが 10 項目であるときに平均 *RMSE* が大きくなり，$\sigma^2_{\gamma_{id(j)}}$=0.8 の条件で平均 *RMSE* は 0.298, $\sigma^2_{\gamma_{id(j)}}$=1.4 の条件で平均 *RMSE* は 0.561 となった。

2PLM 単一分析と 2PLM+GRM 分析から得られた局所独立項目の困難度パラメタの平均 *MD* を図 8 に示す。

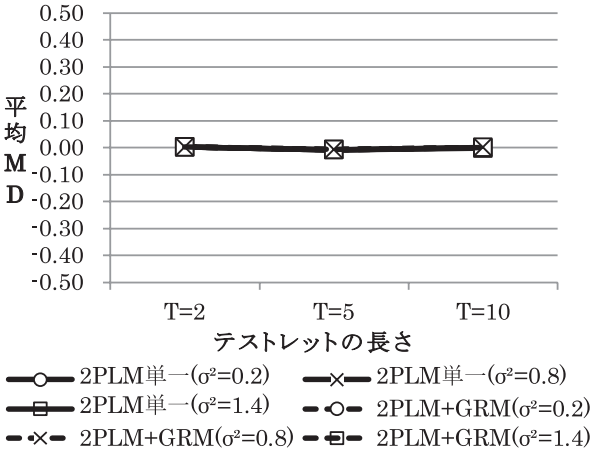

図 8　局所独立項目の困難度パラメタの平均 *MD*

図 8 より，全ての条件について平均 *MD* は，ほぼ 0 であることが示された。局所依存性の強さやテストレットの長さによらず，局所独立項目の困難度パラメタの平均的な大きさは，真値と推定値とでほぼ同じであることが示された。

ただし，$\sigma^2_{\gamma_{id(j)}}$=1.4 かつテストレットの長さが 10 の場合の 2PLM 単一分析においては，困難度パラメタの真値によって *MD* の傾向が異なることが確認された。このことについて，困難度パラメタの真値と困難度パラメタの *MD* の散布図について，図 9 に示した。

困難度パラメタが 1 以下のときの平均 *MD* は -0.15，1 以上のときの平均 *MD* は 0.18 となった。一方，2PLM+GRM

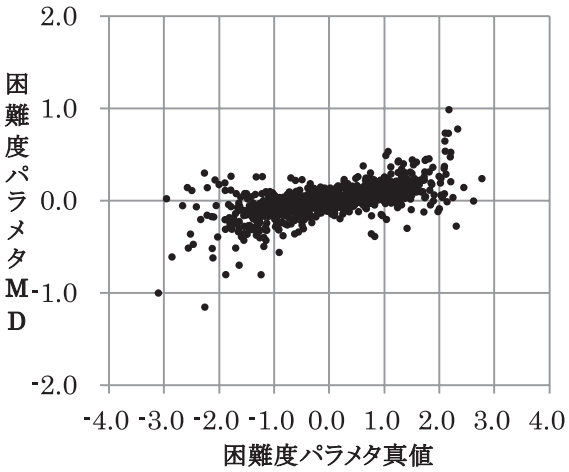

図 9　局所依存性の強さ 1.4，テストレットの長さ 10 の 2PLM 単一分析の，困難度パラメタ真値と *MD* の散布図

分析や，テストレットの長さが 2，5 のときの 2PLM 単一分析の場合においてはこのような傾向は見られず，困難度パラメタの真値によらず *MD* は 0 に近い値をとった。

2PLM 単一分析と 2PLM+GRM 分析から得られた局所独立項目の困難度パラメタの平均 *RMSE* を図 10 に示す。

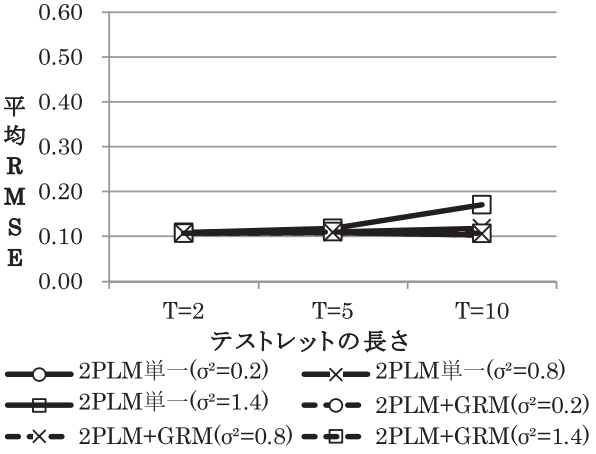

図 10　局所独立項目の困難度パラメタの平均 *RMSE*

図 10 より，テストレットが 10 項目であり，$\sigma^2_{\gamma_{id(j)}}$=1.4 である条件において 2PLM 単一分析を行った場合の局所独立項目の困難度パラメタの平均 *RMSE* が，他の条件と比べて比較的高い値を示していることが分かる。一方で，2PLM+GRM 分析を行った場合は，テストレットの長さや局所依存性の強さに関わらず，ほぼ同じ値を示した。

テストレットが 10 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 である条件の 2PLM+GRM 分析と 2PLM 単一分析の平均 *RMSE* の差は，0.064 であった。

ただし，局所独立項目の困難度パラメタの平均 *RMSE* は，能力パラメタや識別力パラメタにおける平均 *RMSE* の値と比較して全体的に低い値である。2PLM 単一分析と 2PLM+GRM 分析の平均 *RMSE* の差についても，他のパラメタの結果と比べて小さい値を示した。

2PLM 単一分析から得られた局所依存項目の困難度パラメタの平均 *MD* を図 11 に示す。

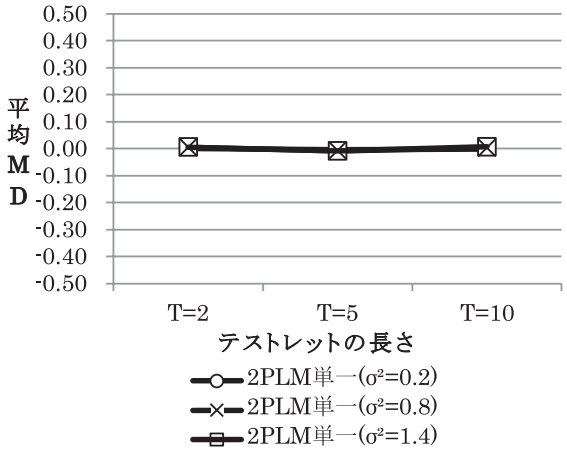

図 11　局所依存項目の困難度パラメタの平均 *MD*

図 11 より，全ての条件において平均 *MD* の値がほぼ 0 であることから，局所依存性の強さやテストレットの長さによらず，局所依存項目の困難度パラメタの平均 *MD* について，過大推定や過小推定の傾向は示されなかったと考えられる。

2PLM 単一分析から得られた局所依存項目の困難度パラメタの平均 *RMSE* を図 12 に示す。

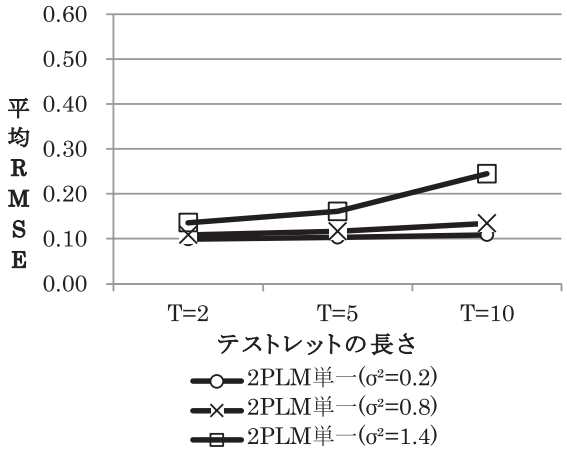

図 12　局所依存項目の困難度パラメタの平均 *RMSE*

図 12 より，テストレットが長いほど 2PLM 単一分析における局所依存項目の困難度パラメタの平均 *RMSE* が高くなる傾向が見られる。特に $\sigma^2_{\gamma_{id(j)}}$=1.4 の場合は平均 *RMSE* が大きくなり，テストレットの長さが 10 項目である場合は 0.245 と，他の条件と比較して大きい値を示した。

局所依存項目の困難度パラメタの平均 *RMSE* は，局所独立項目の困難度パラメタの平均 *RMSE* と比較して大きい値を示している。ただし，能力パラメタや識別力パラ

メタにおける平均 *RMSE* の値と比較すると全体的に低い値である。

4．考察

1） 先行研究との比較　DeMars (2006)[12] や登藤 (2012a)[13] では，局所依存性を考慮するモデルと考慮しないモデルとで能力パラメタの推定誤差が大きく変わらないことが報告されていた。しかし，本研究では，テストレットの長さや局所依存性の強さによっては，モデル間で能力パラメタの推定誤差において比較的大きな差が見られるという結果が得られた。

DeMars (2006)[12] では，受験者数が 2000 人のデータ，テストレットの長さが 5 項目であるデータから能力パラメタの平均 *RMSE* を求めた上で，局所依存性を考慮するモデルと考慮しないモデルのどちらにおいても，能力パラメタの推定誤差の大きさが同等であると結論付けている。分析モデル間の平均 *RMSE* の差は，最大で 0.02 であった。また，登藤 (2012a)[13] では，受験者数 1000 人のデータで，5 項目の長さのテストレットが 4 つあり，$\sigma^2_{\gamma_{id(j)}}$=1.4 である条件を検討しているが，GRM，BREM，2PLM の間に能力パラメタの平均 *RMSE* に 0.05 以上の差が見られず，分析モデル間に大きな差が見られなかったと結論付けている。一方，本研究においては，テストレットの長さが 10 項目かつ $\sigma^2_{\gamma_{id(j)}}$=1.4 である条件において，2PLM+GRM 分析は 2PLM 単一分析と比較して平均 *RMSE* に 0.144 の差が示され，能力パラメタの推定誤差に比較的大きな差が見られた。

本研究において能力パラメタの平均 *RMSE* において，2 値データと多値データに 0.05 より大きな差が見られたのは，テストレットの長さが 10，$\sigma^2_{\gamma_{id(j)}}$=1.4 の条件のみであった。この条件において，DeMars (2006)[12] や登藤 (2012a)[13] で報告されている分析モデル間の能力パラメタの平均 *RMSE* の差より大きな値を示した。先行研究と本研究から得られた，能力パラメタの平均 *RMSE* について，表 1 に示す。表 1 より，テストレットが 5 項目である場合は本研究，先行研究とともに平均 *RMSE* において分析モデル間の差が小さいが，テストレットが 10 項目である場合に比較的大きな差が見られることが示される。以上のことから，テストレットが 5 項目より長く，比較的強い局所依存性が予想される場合は，局所依存性を考慮するモデルを用いる方が，用いない場合より，能力パラメタにおいてより高い推定精度が得られると考えられる。また，DeMars (2006)[12] と同様，条件によっては局所依存性を考慮するモデルを用いる方が，用いない場合よりも能力パラメタの推定誤差が大きくなる場合があることが示された。本研究では σ^2=0.2，テストレットの長さが 2 の条件において，2PLM+GRM 分析の能力パラメタの推定誤差が 2PLM 単一分析の推定誤差より，およそ 0.01 上回っている。ただし，この差は DeMars (2006)[12] がほぼ同等であると解釈した，0.02 の差を下回っている。今回のシミュレーションにおいて，局所依存性を考慮するモデルを用いた時に，用いない場合と比較して能力パラメタの推定誤差が特に大きくなる条件は無かったと考えられる。

項目パラメタの推定誤差の大きさについて，登藤 (2012b)[14] と比較する。登藤 (2012b)[14] では受験者数が 300 人，5 項目の長さのテストレットが 4 つあり，$\sigma^2_{\gamma_{id(j)}}$=1.4 である条件を検討している。この条件の場合，2PLM による分析を行った場合において局所依存項目の識別力パラメタに約 0.20 の過大推定があることを示した。一方，本研究では，テストレットの長さが 5 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 の条件において，局所依存項目の識別力パラメタについて約 0.15 の過大推定があった。

本研究の条件では，テスト全体の項目数が 40 項目，受験者数が 1000 人と，項目数，受験者数が登藤 (2012b)[14] より多いことから，やや低い平均 *MD* が得られたと考えられる。また，本研究ではテストレットの長さが 10 項目であるときの平均 *MD* は，$\sigma^2_{\gamma_{id(j)}}$=0.8 の条件で 0.231，$\sigma^2_{\gamma_{id(j)}}$=1.4 の条件で 0.460 という値が得られた。これらの値は，登藤 (2012b)[14] の示した，5 項目からなるテストレッ

表 1　先行研究と本研究から得られた能力パラメタの平均 *RMSE* の条件間の比較

	受験者数，項目数	テストレットの長さ	局所依存性を無視した場合の平均 *RMSE*	GRM を含めて分析した場合の平均 *RMSE*
本研究(σ^2＝1.4)	1000, 40	10	0.50	0.36
本研究(σ^2＝1.4)	1000, 40	5	0.39	0.35
本研究(σ^2＝0.2)	1000, 40	2	0.28	0.29
DeMars (2006)[12]	2000, 50	5	0.14	0.14
DeMars (2006)[12]	2000, 25	5	0.23	0.25
登藤 (2012a)[13]	1000, 20	5	0.53～0.55	0.53～0.55

トを条件としたときのおよそ 0.20 の識別力パラメタの過大推定より大きい。したがって，テストレットが長い場合において，識別力パラメタがより大きく過大推定されることが示唆される。

また，登藤 (2012b)[14] では 2PLM による分析における局所依存項目の識別力パラメタの平均 *RMSE* と困難度パラメタの平均 *RMSE* はそれぞれともに，最大でおよそ 0.25 の値を示している。本研究ではテストレットの長さが 10 項目の場合に項目パラメタの平均 *RMSE* の値が 0.25 を上回る条件があった。識別力パラメタの平均 *RMSE* について，局所依存性の強さが $\sigma^2_{\gamma_{id(j)}}$=0.8 の条件で 0.298，$\sigma^2_{\gamma_{id(j)}}$=1.4 の条件で 0.561 という値が得られた。ただし，困難度パラメタの平均 *RMSE* については，$\sigma^2_{\gamma_{id(j)}}$=1.4 の条件において 0.245 という値が得られ，登藤 (2012b)[14] と同等の大きさであった。

局所依存性を持つ項目に対して 2PLM を適用した場合について先行研究と比較すると，テストレットの長さが 10 項目かつ $\sigma^2_{\gamma_{id(j)}}$=0.8 以上の条件において，識別力パラメタの推定誤差は特に大きくなると考えられる。ただし，困難度パラメタの推定誤差の大きさは先行研究と同等の値である。識別力パラメタと比較すると，困難度パラメタはテストレットが長い場合においても，推定誤差が大きくならないと考えられる。

また，Tuerlinckx & De Boeck (2001)[11] は，データに局所依存性が含まれる場合に局所依存性を考慮しないモデルを適用すると，局所依存項目の識別力パラメタは過大推定され，局所独立項目の識別力パラメタは過小推定されることが報告されているが，本研究ではその知見を支持する結果が得られた。また，テストレットが長く，局所依存性が強くなるにつれて，局所独立項目の識別力パラメタがより大きく過小推定され，局所依存項目の識別力パラメタがより大きく過大推定されることが示された。

また，これまで，局所独立項目の項目パラメタについては局所依存性を考慮する場合と考慮しない場合との間の推定誤差の比較は検討されていなかった。本研究では，テストレットの長さが 10 項目以上かつ $\sigma^2_{\gamma_{id(j)}}$=1.4 の場合において，局所依存性を考慮する場合に，局所依存性を考慮しない場合と比較して，局所独立項目の識別力パラメタの過小推定が見られず，項目パラメタの推定誤差がより小さくなることが示された。

2）総合考察　本研究の目的は，テストレットの長さの条件によって，局所依存性を考慮する場合と考慮しない場合との間で，項目パラメタや能力パラメタの推定誤差の大きさに差が見られるかを検討することであった。本研究の結果から，テストレットの長さによって，局所依存性を考慮する場合としない場合との間で，能力パラメタ，項目パラメタの推定誤差の大きさに差が見られた。

テストレットの長さが 10 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合において，局所依存性を考慮するモデルは考慮しないモデルと比較して，能力パラメタの平均 *RMSE* において 0.144 低い値が得られた。しかし，それ以外の条件は能力パラメタの平均 *RMSE* について，分析方法の間の差は 0.05 以下であることが示され，大きな差が見られなかった。能力パラメタの推定誤差を小さくするという目的から局所依存性を考慮するモデルを適用するのは，テストレットの長さが 5 項目を上回り，局所依存性が強い場合において有効であることが示唆される。

局所独立項目の識別力パラメタの平均 *MD* については，テストレットの長い場合や局所依存性が強い場合においても，考慮しない場合と比較して 0 に近い値が得られた。局所依存性が強くなるにつれ，また，テストレットが長くなるにつれ，局所依存性を考慮しない場合は局所独立項目の識別力パラメタが過小推定されるが，局所依存性を考慮する場合にはその傾向が見られなかった。

特にテストレットの長さが 10 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合において，局所独立項目の識別力パラメタの真値が高くなるにつれて，識別力パラメタの過小推定の程度が大きくなるという傾向が見られた。この傾向は，テストレットの長さが 2，5 である場合や，局所依存性を考慮するモデルを用いた場合では見られなかった。

また，局所独立項目の識別力パラメタの平均 *RMSE* については，テストレットの長さが 10 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合において，局所依存性を考慮する場合，考慮しない場合よりも 0.162 小さい値が得られた。

一方，局所独立項目の困難度パラメタの平均 *RMSE* については，局所依存性を考慮する場合と考慮しない場合の差が 0.064 と，比較的小さい値が得られた。局所依存性を考慮する場合において，局所独立項目の識別力パラメタの推定誤差は小さくなるが，困難度パラメタの推定誤差は比較的小さくならないことが示唆された。ただし，困難度パラメタの *MD* については，テストレットの長さが 10 項目，$\sigma^2_{\gamma_{id(j)}}$=1.4 の場合の 2PLM 単一分析においてのみ，困難度パラメタの真値が 0 より大きい場合は過大推定，0 より小さい場合は過小推定が起こることが確認された。

以上のことをまとめると，一定の局所依存性の強さが見込まれ，かつテストレットの長さが 10 項目程度である場合には局所依存性を考慮するモデルを用いるほうが IRT の分析において，能力パラメタや項目パラメタについて，より高い推定精度が得られると考えられる。また，局所依存性が弱い場合や，テストレットの長さが短い場合は，局所依存性を考慮するモデルと，考慮しないモデルとの間で，能力パラメタや項目パラメタの推定精度は同程度であることが示唆された。

3）今後の課題　今後の課題として，シミュレーションにおいて検討する条件について，より詳しく検討するこ

とが挙げられる。

まず，データの分析モデルについてである。本研究では局所依存性を考慮するモデルとしてGRMを用いたが，BREMをはじめとした他のモデルを用いた分析を用いることが考えられる。より多くの分析モデルを適用することにより，分析モデルが異なる場合においても今回の研究結果と同様の結果が得られるか，あるいは，よりパラメタの推定誤差の小さいモデルはないかということについて検討することができる。特に，BREMはデータ発生モデルとして用いたため，分析モデルとして扱うことが考えられる。データ発生モデルと同じモデルを分析モデルとし，パラメタの推定誤差の大きさを求めることで，データ発生モデルと同じ分析モデルのパラメタの推定誤差と比較して，他のモデルはどの程度の推定誤差の大きさが得られるかについて検討することができる。

また，本研究で検討したテストレットの長さ，局所依存性の強さの条件について，より多くの条件を設定した上で検討することが考えられる。例えば，本研究ではテストレットの長さが10項目である場合において局所依存性を考慮しないモデルのパラメタの推定誤差がより大きくなるという結果が得られたが，5項目から10項目の間のテストレットの長さの条件を設けた上で検討することが考えられる。また，今回はテストレットの長さが2項目である場合には局所依存性が強い場合においても，局所依存性を考慮するモデルと考慮しないモデルとの間にパラメタの推定誤差の大きさに差が見られなかった。しかし，極端に高い局所依存性が考えられる場合においても同様の結果が得られるかについては検討の余地があると考えられる。今回設定した局所依存性の強さは実際のテストデータから得られた知見をもとにしており，テストの内容は言語，読解，分析推論であった。しかし，Yen (1993)[(3)]は数学テストにおいて，同じデータに対して似た計算を行う場合や，ある項目の答えを次の答えに用いる場合に特に強い局所依存性があったことを示している。そのような場合に，BREMによる分析を行うと$\sigma^2_{\gamma_{id(j)}}$の値はどの程度の値を示すのか，2項目間に極端に高い局所依存性が見られた場合においても，局所依存性を考慮しないモデルと考慮しないモデルとの間にパラメタの推定誤差の大きさに差が見られないかどうか，項目パラメタの過大推定，過小推定の程度が小さいかについて検討することが考えられる。また，シミュレーションの条件をより多く設定することにより，本研究で得られた結果が系統だったものになっているかどうかを確認することができる。例えば，本研究ではテストレットが長くなるほど能力パラメタや項目パラメタの推定精度が悪化することが示唆されているが，テストレットの長さが10項目を超える場合においても同様に推定精度が悪化するかどうかについては今後の課題として残されている。

また，テスト全体の項目数，テストレットの長さ，局所依存項目の割合のうち，本研究ではテスト全体の項目数と局所依存項目との割合を統制し，テストレットの長さを変化させる場合のみを検討した。この他，テスト全体の項目数を変化させるシミュレーションや，局所依存項目の割合を変化させるシミュレーションも考えることができる。これらを検討することで，どの要因がより大きく能力パラメタや項目パラメタの推定精度に影響するのかが明らかになると考えられる。

最後に，本研究で得られた結果が，シミュレーションによって得られたものであるという課題が残されている。実際のテストデータでは，シミュレーションで生成したデータよりモデルの適合度が低いことが考えられる。また，大問ごとの問題形式への慣れや疲労効果などの，シミュレーションで想定しなかった要因がパラメタの推定精度に影響する可能性がある。本研究から得られた結果が，実際のテストデータにおいても同様であるかどうかについては，更なる検討が必要であると考えられる。

―謝　辞―

本論文は平成24年度岡山大学大学院教育学研究科修士論文を加筆修正したものです。

ベネッセコーポレーションの加藤健太郎様，堀一輝様には，貴重なご意見，ご指摘をいただきました。心より感謝の意を表します。

東京大学大学院の登藤直弥様には，本研究の着想の元となる考えを示していただきました。深く感謝申し上げます。

―引用文献―

(1) 芝　祐順（編）『項目反応理論―基礎と応用―』東京大学出版会，1991

(2) 豊田秀樹『項目反応理論 [入門編]―テストと測定の科学―』朝倉書店，2002

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

(4) 荒井清佳，前川眞一「日本の公的な大規模試験に見られる特徴―標準化の観点から―」『日本テスト学会誌』1, pp. 81-92, 2005

(5) 石塚智一，中畝菜穂子，内田照久，前川眞一「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, pp. 1-24, 2001

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

(7) Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometorika Monograph Supplement*, Vol. 34, pp. 100-114, 1969
(8) Bradlow, E. T., Wainer, H., & Wang, X. H. A bayesian random effects model for testlets. *Psychometrika*, Vol. 37, pp. 29-51, 1999
(9) 登藤直弥「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, pp. 17-28, 2010
(10) Chen, C., & Wang, W. Effect of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement*, Vol. 31, pp. 388-411, 2007
(11) Tuerlinckx, F., & De Boeck, P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, Vol. 6, pp. 181-195, 2001
(12) DeMars, C. E. Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, Vol. 43, pp. 145-168, 2006
(13) 登藤直弥「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, pp. 85-100, 2012a
(14) 登藤直弥「項目反応間の局所依存性が項目母数の推定に与える影響―項目母数の比較可能性を確保した上での検討―」『行動計量学』39, pp. 81-91, 2012b
(15) Wainer, H., & Wang, X. Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, Vol. 37, pp. 203-220, 2000
(16) 中央教育審議会「新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け，主体的に考える力を育成する大学へ～」2013, http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2012/10/04/1325048_1.pdf（2013年1月6日閲覧）
(17) Harwell, M., Stone , C. A., Hsu, T. C., & Kirisci, L. Monte Carlo studies in item response theory. *Applied Psychological Measurement*, Vol. 20, pp. 101-125, 1996
(18) Zu, J., & Liu, J. Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, Vol. 47, pp. 395-412, 2010
(19) Muraki, E., & Bock, R. D. PARSCALE. [computer software]. Chicago: Scientific Software International, 1997
(20) Thissen, D. MULTILOG. [computer software]. Chicago: Scientific Software International, 1991
(21) Cai, L., Thissen, D., & du Toit, S. H. C. IRTPRO for Windows. [computer software]. Lincolnwood, IL: Scientific Software International, 2011
(22) Ip, E. H. Interpretation of the three parameter testlet response model and information function. *Applied Psychological Measurement*. Vol. 34, pp. 467-482, 2010
(23) Li, Y., Bolt, D.M., & Fu, J. A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*. Vol. 29, pp. 340-356, 2005
(24) Li, Y., Bolt, D.M., & Fu, J. A comparison of alternative models for testlets. *Applied Psychological Measurement*, Vol. 30, pp. 3-21, 2006