日本オペレーションズ・リサーチ学会和文論文誌
Transactions of the Operations Research Society of Japan
Vol. 58, 2015, pp. 122–147

項目反応理論を用いたテスト運用への切り替えコスト軽減の試み
―多数の潜在特性尺度の同時等化法を利用して―

豊田 秀樹
早稲田大学

岩間 徳兼
独立行政法人国際交流基金

中村 彩子
元株式会社イー・コミュニケーションズ

齋藤 康寛
株式会社イー・コミュニケーションズ

(受理 2014 年 8 月 5 日; 再受理 2015 年 8 月 17 日)

和文概要　古典的テスト理論によるテストから項目反応理論によるテストへの切り替えにおいては，テスト項目を集めた項目プールを作るための予備調査の実施に大きなコストがかかる．本論文ではある特性を測定するために作成された古典的テスト理論による複数の個別テストおよびそのデータがある状況において，それらを活用することでコストを抑えて大きな項目プールを作る方法を示す．項目プールを作成するのに必要な等化係数の有効な推定方法を提案し，実データの分析例とシミュレーション研究の結果から提案手法について確認を行った．実データの分析では，提案した方法によって一度の調査から大きな項目プールを作ることに成功した．また，シミュレーションデータの分析では，提案した推定法が統計的に適切な性質を有していることが分かり，実データの推定結果の妥当性が認められた．

キーワード: 統計, 項目反応理論，同時等化，周辺最尤推定法

1. 問題と目的

項目反応理論（Item Response Theory: 以下 IRT，[2, 4, 6]）によるテストは，受験者の能力分布やテストに含まれる項目の難易度が異なっても同じ尺度上で受験者の能力を比較できる等，その有用性が認められている．TOEFL や日本語能力試験，IT パスポート試験等，受験者が毎年何万人，何十万人といるような大規模な試験や公的な試験は IRT を用いたテストへの切り替えが進んでいる．また，医学，歯学教育場面においては臨床実習前の学生の総合的知識の確認のために IRT を利用した共用試験が実施されており，薬学，看護学，獣医学へもその動きは広がっている．しかし，テスティング・カンパニーが単独で実施するような中小規模のテストにおいては，その有用性は認識されてはいるものの，IRT を用いたテストへの切り替えは進んでいないのが現状である．その要因の一つとして，IRT への切り替えコストが膨大かつ予測しにくい点があげられる．

IRT への切り替えの際のコストには「ソフトウェア」,「人」,「データ」の 3 つの要素が考えられる．1 つ目の「ソフトウェア」は，推定を行うための統計ソフトウェアの導入にかかるコストのことである．以前は高価な市販の統計ソフトウェアが主流であったが，最近では「R」等のフリーのソフトウェアがあり，安価に導入することも可能となってきている．2 つ目の「人」は，IRT の理論を理解し，実用に向けてテストを計画・実施できる人材の確保・育成にかかるコストである．大規模なテストでは，テスト運営団体が独自の研究機関を持ち，内部で全て運用している場合もある．しかし，理論の習得にはある程度の期間が必要であり，中小規模のテスト運営団体が内部に研究機関を設け人材を確保することは困難であ

る．現実的な解決策としては，テストを専門に扱い，コンサルティングから運営までのトータルサポートを提供するような機関に業務を委託する方法がある．3 つ目の「データ」はテストを実施する前の準備に必要なデータ収集にかかるコストのことである．IRT でテスト運営をする場合の多くは，項目の特徴を表す母数 (項目母数) が知られた項目からなる大きな集合 (項目プール) が事前に必要であり，そのために予備テストを行い，項目母数を推定することが求められる．予備テストを行うためには受験者や会場の確保が必要だが，多くのデータを集めようとすればそれだけコストが増大する．例えば，1000 問の項目プールを作成するために 1 フォーム 100 問の予備テストを複数作成して共通項目法で等化する場合，共通項目を 25 問設けるとすると 13 グループ必要となり，各グループの受験者数を 300 人とした場合，延べ 3,900 人の受験者が必要となる．事前テスト実施のための会場費用や受験者への謝礼等で，一人あたり 5,000 円の費用がかかるとすると，実施費用は 1950 万円かかり，事前テストの実施だけで 1 問あたり 19,500 円かかると試算される．事前の予備テストは必要最低限にしたいというのが実際である．しかし，事前にどの程度データを収集すればいいのかを示した知見はほとんどない．

本論文では，上述の 3 つのコストの要素のうち「データ」部分に焦点を当て，そのコストの増大を抑えたうえで IRT テストへの容易な切り替えを実現する方法を検討する．古典的テスト理論に基づいて実施された過去のテストデータがある状況を想定し，予備テストのための効率的なテストの構成法を示し，さらに，収集データから大規模な項目プールを作成するために用いられる等化係数を推定するのに有効な方法を提案する．そして，古典的テスト理論に基づいて長年実施し，受験データを蓄積してきたあるテストにおいて，手法を適用した事例を提示するとともに，提案した推定手法の妥当性についてもシミュレーション研究によって確認を行う．

提案手法の利用によって，IRT への切り替え以前に得られたテストのデータを有効に活用し，コストを抑えながら簡単な手続きで大規模な項目プールを作成することが可能となる．

2. 方法

2.1. 項目および尺度値の等化と等化係数

IRT を用いたテスト運用においては，項目プールの拡充や，テスト結果の比較を目的として等化という手続きが採られる．等化は異なる機会に得られたテストの結果を同一尺度上に乗せる手続きのことである．IRT においては任意に原点と単位を決めることができる潜在特性を扱っているので，異なる集団のもとで得られた複数の尺度値や項目母数を比較可能にするためにはそれらの原点と単位を揃えるための等化が必要となる．

これまで様々な等化法が提案されているが（たとえば，[3, 5, 11, 15]），IRT の文脈において成立する項目母数や潜在的特性値についての線形関係を利用したものが代表的である．ロジスティック型の項目特性曲線や正規累積型の項目特性曲線を利用した 1 母数モデルでは，潜在特性についての尺度 θ 上で位置付けられる全項目共通の識別力母数 a および困難度母数 b_j（j は項目を表す添え字）と尺度 θ^* 上で位置付けられる a^*, b_j^* に関して，以下の (2.1) 式から (2.3) 式の関係が成り立つ．

$$\theta^* = \alpha\theta + \beta \tag{2.1}$$

$$a^* = \alpha^{-1}a \tag{2.2}$$

$$b_j^* = \alpha b_j + \beta \tag{2.3}$$

この関係は尺度の変換や等化に用いられ，等化の文脈において傾き α と切片 β を合わせて等化係数と呼ぶ．

なお，1 母数モデルでは，a^*, a が特定の同じ値（たとえば 1）となるように母数化することも可能である [14] ．その場合には $\alpha(=1)$ を無視することができるので

$$\theta^* = \theta + \beta \tag{2.4}$$

$$b_j^* = b_j + \beta \tag{2.5}$$

の関係のみが成り立ち，等化においても β（以下，等化切片と呼ぶ）だけに着目すればよい．以降ではそのような状況を想定して話を進める．

2.2. パイロットテストの構成

ある 1 つの特性を測定するために過去に古典的テスト理論によって実施された，受験者や実施日時が同一でない T 個のテストが存在する状況を考える．まずは，IRT の観点から個別に各テストの項目母数を推定する．ただし，これらは T 個の比較可能でない尺度であり，個別に実施されたテストから巨大な項目プールを作成するためにはこれらを 1 つの尺度に等化する必要がある．そのためのデータ収集を 1 回のテスト (以下，パイロットテスト) の実施により行う．

パイロットテストは，T 個の個別テストからそれぞれ $J_t(t = 1, 2, \ldots, T)$ 個の項目を選んで集めて構成した合計項目数 $J(= \sum_{t=1}^{T} J_t)$ のテストである（図 1）．これを I 人の被験者に受験してもらい，T 個の個別テストについて既に得られた各項目の母数とパイロットテストの反応パタンから，パイロットテストに各テストを等化するための係数を求める．具体的に，パイロットテストと T 個の個別テストにおいて，それぞれ含まれる項目の母数の間にはテスト t の場合の (2.5) 式を意味する

$$b_{tj}^* = b_{tj} + \beta_t \tag{2.6}$$

の関係式が成り立っているとする．ここで，j はテスト t に含まれる j 番目の項目を表す添え字である．j は t に依存しており，1 から J_t までの値をとる．b_{tj}^* はパイロットテストの尺度上で，b_{tj} はそれぞれの個別テストの尺度上で位置づけられる困難度を表す母数である．また，β_t は t 番目の個別テストの項目母数をパイロットテストの尺度上に位置付けるための等化切片である．このとき，多数の潜在尺度を同時に等化するための $\beta_t(t = 1, \ldots, T)$ を推定するための方法を以下で示す．

2.3. 等化係数の推定方法

等化係数の推定と多数のテストの等化

等化係数を推定するための方法は，複数のテストに関する共通の項目の情報を利用する共通項目計画と複数のテストに共通した被験者の情報を利用する共通被験者計画とに大別される．共通被験者計画を利用した等化係数の最尤推定については，これまで様々な方法が提案されている（たとえば，[8–10, 12]）．これらの方法は同時最尤推定法か周辺最尤推定法か，観測変数と推定量のどちらに基づいて尤度構成を行っているか，共通被験者と事前の項目母数推定時の被験者の母集団の関係の捉え方などに違いがあるものの，どの方法も 2 つのテストの間の等化を扱い，基準となる片方のテストの尺度上にもう一方のテストを位置づけるという点では共通している．すでに IRT によるテストの運用が行われている状況では，計画的に共通被験者を募り回答データを収集すれば 2 つのテスト間の等化で項目プールを拡

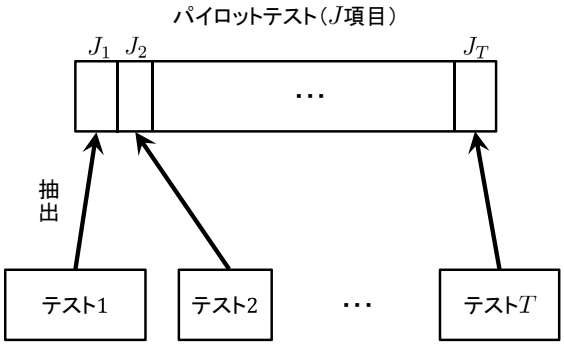

図 1: パイロットテストの構成

充していくことができ，問題はないように思われる．しかしながら，本論文で扱っているような，ある潜在特性を測定する多数の個別下位テストがすでに存在する状況においては，それらを同時に等化できるのならば，パイロットテストおよび等化を一度だけ実施すれば済むので IRT によるテスト運用へとスムーズに切り替えることが可能となる．

複数テストを同時に等化することについては，豊田 [13] の方法，多母集団 IRT モデル [1]，欠損値のある IRT モデルなどの利用が考えられる．しかしながら，これらの方法は本論文で扱う課題に対してあまり有効とはいえない．豊田 [13] では，等化係数の推定に先立ってテストごとの共通受験者の θ の推定値が必要となるので，本論文の適用例で取り上げるようなテストの数が多い場合には実用性が高くない．また，多母集団 IRT モデルではテストごとのデータの独立性などのモデル上の仮定が，欠測値のある IRT モデルでは被験者の欠測情報の捕捉が必要であり，これらのアプローチは過去の受験データが既に得られている状況への適用が難しい．さらに，テストの数や項目数が多い場合には推定時間が非常にかかるとともに非収束となる可能性も高い．IRT によるテストへの切り替えは最近のトレンドであり，その際の敷居の低さが求められている中でそれらの方法はあまり有効とは言えない．

他方，実際に複数のテストを等化しており，本研究と関連のある応用研究として，吉村・荘島・杉野・野澤・清水・齋藤・根岸・岡部・サイモン [16] がある．この研究は本研究で示すテストデザインと同様のデザインを採用しているものの，等化のアプローチに違いが見られる．吉村ら [16] では，英語学力の年度変化の考察を目的として，15 年度分のセンター試験の英語テストにおいて初年度のテストに対して残り 14 年度のテストを個別に等化して共通尺度を構成している．この場合，それほど多くのテストを扱っていないので個別に等化することに大きな問題はないが，本研究では項目プールの構築を目的として多数のテストを等化しなければならない状況を想定しており，個別の等化は実用的ではない．

そこで，上記の問題を解決でき，IRT によるテストへの切り替えが容易でかつ有効な方法として，多数の潜在特性尺度の同時等化を実現するための周辺最尤推定法に基づく等化係数の推定を提案する．

多数のテストを同時に等化するための β の周辺最尤推定

T 個のテストを同時に等化し，すべての項目母数をパイロットテストの尺度上に位置付けるためには，T 個の β を推定することが求められる．まず，被験者 i のテスト t に含まれる項目 j に対する反応を u_{itj} と表すことにする．u_{itj} は当該項目に正答ならば 1，誤答ならば 0 をとるものとする．この時，その項目に対する尺度値 θ^* に基づく反応確率を，1 母数ロジ

スティック型の項目反応関数を利用して

$$P(u_{itj} \mid \theta_i^*, b_{tj}^*) = \frac{\exp\{-D(\theta_i^* - b_{tj}^*)(1-u_{itj})\}}{1+\exp\{-D(\theta_i^* - b_{tj}^*)\}} \tag{2.7}$$

と表現する．ここで，b_{tj}^* は困難度を表す母数である．D は尺度因子であり，本論文では 1.7 とする．また，パイロットテストの尺度値が $\theta^*(\theta^* \sim N(0,1))$ であると仮定する．

被験者 i のテスト t に対する反応を $\boldsymbol{u}_{it} = [u_{it1}, \ldots, u_{itj}, \ldots, u_{itJ_t}]'$ のようにベクトル形式で表すと，$\boldsymbol{u}_{it}$ が得られる確率は，局所独立の仮定のもとで

$$P(\boldsymbol{u}_{it} \mid \theta_i^*, \boldsymbol{b}_t^*) = \prod_{j=1}^{J_t} P(u_{itj} \mid \theta_i^*, b_{tj}^*) \tag{2.8}$$

として得られる．ただし，$\boldsymbol{b}_t^* = [b_{t1}^*, \ldots, b_{tJ_t}^*]'$ と定める．さらに，被験者 i のパイロットテスト全体に対する反応ベクトルを $\boldsymbol{u}_i = [\boldsymbol{u}_{i1}', \ldots, \boldsymbol{u}_{it}', \ldots, \boldsymbol{u}_{iT}']'$ のように，T 個の各テストに対する反応ベクトルをつなぎ合わせた形で表現すると，$\boldsymbol{u}_i$ が得られる確率は

$$P(\boldsymbol{u}_i \mid \theta_i^*, \boldsymbol{B}^*) = \prod_{t=1}^{T} P(\boldsymbol{u}_{it} \mid \theta_i^*, \boldsymbol{b}_t^*) \tag{2.9}$$

と表される．ここで，$\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_T^*]$ である．

(2.6) 式を考慮すると，(2.9) 式は等化前の各テストの項目母数および等化切片を用いて

$$\begin{aligned} P(\boldsymbol{u}_i \mid \theta_i^*, \boldsymbol{B}^*) &= P(\boldsymbol{u}_{11} \mid \theta_i^*, (\boldsymbol{b}_1 + \beta_1 \boldsymbol{1})) \\ &\quad \times P(\boldsymbol{u}_{12} \mid \theta_i^*, (\boldsymbol{b}_2 + \beta_2 \boldsymbol{1})) \\ &\quad \times \cdots \\ &\quad \times P(\boldsymbol{u}_{1T} \mid \theta_i^*, (\boldsymbol{b}_T + \beta_T \boldsymbol{1})) \\ &= P(\boldsymbol{u}_i \mid \theta_i^*, \boldsymbol{B}, \boldsymbol{\beta}) \end{aligned} \tag{2.10}$$

と書き下せる．ただし，$\boldsymbol{b}_t = [b_{t1}, \ldots, b_{tJ_t}]'$, $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_T]$ とする．また，等化切片を集めたベクトルを $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_T]'$ と表す．なお，$\boldsymbol{\beta}$ の推定に際しては，等化先のパイロットテストの尺度において不定性を回避するための制約が必要であり，先述のように尺度値の平均が 0，標準偏差が 1 であるという仮定を設けた．

(2.10) 式を $\boldsymbol{u}_i, \boldsymbol{B}$ が与えられたときの $\boldsymbol{\beta}$ についての尤度関数 $L(\boldsymbol{u}_i \mid \theta_i^*, \boldsymbol{B}, \boldsymbol{\beta})$ と見なした上で，局外母数 θ_i^* が標準正規分布に従うと仮定してその密度関数 $g(\theta_i^*)$ を用いて周辺化すると

$$L(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta}) = \int L(\boldsymbol{u}_i \mid \theta_i^*, \boldsymbol{B}, \boldsymbol{\beta}) g(\theta_i^*) d\theta_i^* \tag{2.11}$$

となる．これを被験者の人数 I に関して拡張すれば，I 人の被験者の全項目に対する反応パタン行列 $\boldsymbol{U}$ が与えられたときの周辺尤度関数は

$$L(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta}) = \prod_{i=1}^{I} L(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta}) \tag{2.12}$$

として得られ，これを最大化することによって $\boldsymbol{\beta}$ の周辺最尤推定値を得ることができる．ただし，実際場面では $L(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})$ の代わりに，区分求積法を用いた近似式

$$L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta}) = \sum_{m=1}^{M} L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta}) h(x_m) \tag{2.13}$$

を利用し

$$L^*(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta}) = \prod_{i=1}^{I} L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta}) \tag{2.14}$$

を最大化することが多い．ここで x_m は分点の位置，$h(x_m)$ はそれに対応する重み係数である．本研究では標準正規密度関数 $g(x)$ に基づく区分求積法を利用し，等間隔に設定した分点 x_m に対応する重み係数 $h(x_m) = g(x_m)(x_{m+1} - x_m)$ によって積分を近似した．

(2.14) 式を最大化するにあたっての周辺対数尤度関数の勾配ベクトルは

$$\nabla_{\boldsymbol{\beta}} = \sum_{i=1}^{I} \sum_{m=1}^{M} Dh(x_m) \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})} \boldsymbol{r} \tag{2.15}$$

として導かれる（付録）．ただし，$\boldsymbol{r}$ は以下のようである．

$$\boldsymbol{r} = [r_1, \ldots, r_t, \ldots, r_T]' \tag{2.16}$$

$$r_t = \sum_{j=1}^{J_t} \{P(u_{itj} = 1 \mid x_m, b_{tj}, \beta_t) - u_{itj}\} \tag{2.17}$$

標本誤差による $\boldsymbol{\beta}$ の推定量の平均的なばらつきである標準誤差の推定値は収束点におけるヘッセ行列の逆行列の対角要素の平方根で計算される．なお，実用場面では，尺度値推定値の標準誤差の評価時と同じく，項目母数の真値を推定値に置き換えて評価をすることになる．

多数のテストを個別に等化するための β_t の周辺最尤推定

つづいて，T 個のテストの結果をそれぞれ，パイロットテストに等化するため β_t の推定法について説明する．この方法は計算量の観点から，前述の同時等化の簡便法に位置づけられる．また，同時等化の際の初期値として利用することも可能である．

テスト t のみに注目すると，パイロットテストでの反応確率は，等化前の母数と等化切片を用いて

$$P(\boldsymbol{u}_{it} \mid \theta_i^*, \boldsymbol{b}_t^*) = \prod_{j=1}^{J_t} P(u_{itj} \mid \theta_i^*, b_{tj}, \beta_t) \tag{2.18}$$

と表現できる．$\boldsymbol{b}_t$ が既知であるとしたもとで，テスト t に対する反応ベクトル $\boldsymbol{u}_{it}$ および反応行列 $\boldsymbol{U}_t$ が与えられれば

$$L_t^*(\boldsymbol{u}_t \mid \boldsymbol{b}_t, \beta_t) = \sum_{m=1}^{M} P(\boldsymbol{u}_{it} \mid x_m, \boldsymbol{b}_t, \beta_t) h(x_m) \tag{2.19}$$

$$L_t^*(\boldsymbol{U}_t \mid \boldsymbol{b}_t, \beta_t) = \prod_{i=1}^{I} L_t^*(\boldsymbol{u}_{it} \mid \boldsymbol{b}_t, \beta_t) \tag{2.20}$$

が導かれる．(2.20) 式を最大化することで β_t の推定値が得られ，これをすべての個別テストについて繰り返して全体の推定値を得る．また，標準誤差の推定値は先と同様に計算することができる．

本研究では，R 言語 (version2.15.2) により各推定法を実行するためのコードを記述した．最適化計算には optim 関数を採用し，上で導出した勾配ベクトルを引数に指定するとともに，周辺対数尤度の最大化法として準ニュートン法を指定した．また，ヘッセ行列は optim 関数より数値的に返される．

3. 適用事例

3.1. 方法

期間 2011 年 12 月 20 日および 2012 年 2 月 6 日

受験者 2012 年 2 月実施の医師国家試験を受験予定の予備校生 110 名

調査方法 A 社は国家試験用対策試験を毎年実施している．国家試験は全ての科目が混在した試験となっているが，対策試験は 11 科目，各科目 2 種の 22 種類の試験に分かれており，それぞれ実施日が異なる．まず，これらの過去受験済みの 54 試験 (以下個別試験) から項目母数を個別に推定し，54 個の比較可能でない尺度を作成した ($T = 54$，全項目数は 4531)．各個別試験の受験者数が 100 名程度であるので，1 母数ロジスティックモデルを採用した．これらの尺度を単一の尺度に等化するために，各個別試験から 3 から 7 項目ずつ選び，合計 300 項目からなるパイロットテストを作成した．各個別試験からパイロットテストに含める項目を選ぶ際には，当該国家試験における知識の集積度の高さから内容的な偏りを考慮した．専門家に依頼し，その分野において公表されている細目表からバランスよく選んでもらうこととした．

[14] によると，2 つのテストの等化においては共通項目数 5 つ以上を目安としている．共通項目が 3 項目というのはその目安を超えていないが，受験負荷の観点からやむを得ず 2 試験の共通項目を 3 項目とした．パイロットテストは PC 上で行う CBT(Computer Based Testing) 方式のテストであり，受験者の都合の良い方の日時に受験してもらった．受験機会の両方を経験した者はいないので 2 回の受験者を併合した集団を等化の際に基準となる集団として扱うこととし，受験機会の 2 カ月の間の試験準備の効果については考えなくてもよい．

3.2. 項目分析

まず，パイロットテストの受験データについて α 係数を算出したところ，0.960 であった．続いて点双列相関係数を算出した．等化については，共通項目数が多い方が等化の結果が安定し，点双列相関係数が低い項目を共通項目として使用すると等化に悪影響を及ぼすことが知られている．そこで等化に及ぼす影響のバランスを考え，パイロットテストの結果から点双列相関係数が 0.15 未満であり，かつ当該項目よりも点双列相関係数が高い項目が同一個別試験内に 5 項目以上ある場合のみ，当該項目を削除項目とした．削除されたのは 11 項目であった．個別試験との共通項目数は，3 項目が 2 試験，5 項目が 32 試験，6 項目が 17 試験，7 項目が 3 試験であった (個別試験との共通項目数については表 1 中の No. の列の括弧内を参照のこと)．各受験者の 289 項目の反応パタンと個別試験の過去の受験データから得た項目母数より，個別試験の項目をパイロットテストに等化するための等化切片を求めた．

3.3. 結果と考察

同時等化のための推定法によって得られた各個別試験の等化切片と標準誤差推定値を表 1 に示す．標準誤差推定値を見ると，最も数値が小さいのは No.42 と No.47 で 0.105，続いて No.13 で 0.106 である．この 3 試験はパイロット試験との共通項目数が 7 項目である．最も数値が大きいのは No.54 で 0.123，続いて No.51 で 0.122 である．この 2 試験は，パイロットテストとの共通項目数が 3 項目である．共通項目数が少ないと標準誤差推定値は大きくなることがわかる．標準誤差推定値 0.123 という値は，偏差値に換算すると約 1.23 であり，共通項目数が 3 項目でもその程度の誤差で推定することができれば許容範囲であると思われる．同時等化法は同時に等化するテスト数が増えるほど 1 度に行う計算量が膨大になるが，できるだけ多くのテストを同時に等化した方が巨大な項目プールを作成するための等化作業の

表 1: 等化切片 $\boldsymbol{\beta}$ の推定値と標準誤差

No.	推定値	s.e.	No.	推定値	s.e.	No.	推定値	s.e.
1(5)	.199	.110	19(6)	.299	.107	37(5)	.400	.110
2(5)	-.188	.112	20(5)	.155	.111	38(5)	.402	.109
3(5)	.441	.110	21(5)	.200	.109	39(5)	.089	.111
4(5)	-.191	.111	22(5)	.466	.110	40(6)	.115	.108
5(5)	.094	.109	23(5)	.113	.110	41(5)	.175	.110
6(5)	.335	.110	24(5)	.039	.110	42(7)	-.240	.105
7(5)	-.063	.110	25(5)	-.198	.113	43(5)	.260	.109
8(5)	-.308	.112	26(5)	-.100	.110	44(6)	-.161	.109
9(5)	-.265	.111	27(5)	.260	.109	45(6)	-.105	.109
10(5)	-.155	.110	28(5)	.353	.109	46(6)	.003	.109
11(5)	.259	.110	29(6)	-.023	.108	47(7)	.178	.105
12(5)	.019	.110	30(6)	.118	.107	48(5)	.513	.110
13(7)	.292	.106	31(6)	.040	.107	49(6)	-.130	.110
14(6)	.176	.108	32(6)	.041	.109	50(5)	.346	.113
15(6)	-.108	.108	33(5)	.053	.112	51(3)	.392	.122
16(6)	.088	.109	34(6)	.545	.108	52(6)	-.644	.111
17(5)	.649	.109	35(6)	-.027	.108	53(5)	.029	.112
18(6)	-.083	.107	36(5)	.071	.110	54(3)	.025	.123

負担が軽減され，かつ等化の精度も保たれる有効な方法であることがわかる．これらの等化切片から各テストの等化後の項目母数を算出し，最終的に 4531 項目からなる項目プールが作成できた．

以上より，54 試験という多数の等化されていない尺度を，1 回のパイロットテストの反応パタンから周辺最尤推定法を用いて全テストの等化切片を同時に推定することで，大きな負荷なく古典的テスト理論で運用されていた試験を IRT による試験に切り替えることが可能であることが示された．

4. シミュレーション研究 1

シミュレーションによって発生させたデータを同時等化のための推定法を用いて分析し，推定法の性質について考察する．本シミュレーションでは適用事例の条件となるべく近い状況を設定し，先の分析結果の妥当性を確認することを目的とする．

4.1. セッティング

適用事例でのデータに合わせて，テスト数は 54，各テストに含まれる項目数は 3 から 7(各テストの項目数については表 2 中の No. の列の括弧内を参照のこと)，被験者数は 110 とした．そして，項目母数 b_{tj} の真の値は実データの分析によって得られた推定値とした．等化切片の真の値も同様である．また，等化切片を推定するときの初期値には mean/sigma 法 [7] による推定値を利用した．この状況でデータ行列の発生と推定を 300 回繰り返した．なお，区分求積法における区分点の数は 60 個とし，範囲は $[-3.5, 3.5]$ とした．

4.2. データ発生手続き

シミュレーションデータは，T を 54，I を 110，J_t を 3 から 7 として，以下のような手続きによって発生させた．なお，本シュミレーションでは分析モデルに合わせて (2.7) 式の 1 母数モデルを採用した．

(1) T 個のテストの等化前の項目母数 $\boldsymbol{b}_t$ を設定する．
(2) $\boldsymbol{\beta}$ の真値を設定する．

表 2: シミュレーション研究 1 の結果

No.	$\bar{b}_j$	真値	平均	RMSE	$s.\bar{e}.$	No.	$\bar{b}_j$	真値	平均	RMSE	$s.\bar{e}.$	No.	$\bar{b}_j$	真値	平均	RMSE	$s.\bar{e}.$
1(5)	-.38	.199	.204	.119	.115	19(6)	-.36	.299	.301	.111	.112	37(5)	-.32	.400	.402	.113	.115
2(5)	-.34	-.188	-.180	.117	.116	20(5)	-.21	.155	.156	.111	.116	38(5)	-.45	.402	.410	.109	.114
3(5)	-.17	.441	.445	.116	.115	21(5)	-.41	.200	.206	.118	.114	39(5)	-.38	.089	.098	.116	.115
4(5)	-.33	-.191	-.187	.110	.116	22(5)	-.45	.466	.478	.115	.115	40(6)	-.28	.115	.115	.114	.112
5(5)	-.22	.094	.097	.111	.114	23(5)	-.42	.113	.121	.116	.115	41(5)	-.20	.175	.185	.114	.114
6(5)	-.45	.335	.344	.111	.114	24(5)	-.18	.039	.049	.117	.114	42(7)	-.01	-.240	-.234	.110	.109
7(5)	-.31	-.063	-.065	.118	.115	25(5)	-.57	-.198	-.200	.116	.117	43(5)	-.23	.260	.267	.117	.114
8(5)	-.35	-.308	-.307	.118	.116	26(5)	-.20	-.100	-.093	.115	.114	44(6)	-.20	-.161	-.158	.106	.113
9(5)	-.06	-.265	-.258	.119	.115	27(5)	-.16	.260	.265	.113	.114	45(6)	-.40	-.105	-.100	.116	.112
10(5)	-.06	-.155	-.147	.117	.115	28(5)	-.43	.353	.358	.121	.114	46(6)	-.20	.003	.005	.115	.112
11(5)	-.34	.259	.266	.112	.115	29(6)	-.43	-.023	-.018	.114	.112	47(7)	-.29	.178	.180	.113	.109
12(5)	-.36	.019	.017	.116	.115	30(6)	-.22	.118	.124	.109	.111	48(5)	-.42	.513	.515	.117	.115
13(7)	-.30	.292	.294	.111	.109	31(6)	-.23	.040	.049	.111	.112	49(6)	-.46	-.130	-.126	.112	.113
14(6)	-.37	.176	.182	.112	.112	32(6)	-.22	.041	.045	.117	.112	50(5)	-.52	.346	.348	.114	.116
15(6)	-.24	-.108	-.105	.116	.112	33(5)	-.37	.053	.061	.122	.116	51(3)	-.38	.392	.394	.132	.127
16(6)	-.27	.088	.088	.109	.112	34(6)	-.22	.545	.551	.112	.112	52(6)	-.19	-.644	-.638	.112	.114
17(5)	-.47	.649	.653	.117	.114	35(6)	-.32	-.027	-.013	.111	.112	53(5)	-.21	.029	.034	.116	.116
18(6)	.02	-.083	-.082	.116	.111	36(5)	-.39	.071	.072	.117	.115	54(3)	-.73	.025	.029	.135	.128

(3) $t = 1, \ldots, T$ について $\boldsymbol{b}_t^* = \boldsymbol{b}_t + \beta_t \mathbf{1}$ の変換を施して，テスト t の等化後の項目母数を計算する．

(4) $\mu = 0, \sigma = 1$ の正規分布から被験者集団に属する被験者の尺度値 θ_i^* を 110 個発生させる．

(5) テスト $t(t = 1, \ldots, T)$ について，$P(u_{itj} = 1 \mid b_{tj}^*, \theta_i^*)$ を計算する．

(6) (0,1] の一様分布から乱数 r_{itj} を発生させ，$r_{itj} < P(u_{itj} = 1 \mid b_{tj}^*, \theta_i^*)$ なら $u_{itj} = 1$，そうでなければ $u_{itj} = 0$ とする．

(7) 手続き (5)，(6) を各被験者と項目に関して行い，テスト t についての反応行列 $\boldsymbol{U}_t$ を得る．そして，$\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T]$ によって反応行列を得る．

4.3. 結果と考察

300 回の繰り返しのうち，適切に収束していないと思われる推定値は認められず，等化切片に関して得られた 300 個の推定値について

1) 推定値の平均
2) 真値からの平均 2 乗誤差平方根 (RMSE)
3) 標準誤差推定値の平均 $(s.\bar{e}.)$

を算出し，その結果を表 2 に示した．表 2 に含まれる $\bar{b}_j$ の列には各テストにおける項目母数 b_{tj} の平均を掲載した．

分析結果を見ると，提案した手法による推定の適切性と適用事例での結果の妥当性が示されている．推定値の平均についてはどの等化切片も真値に近い値である．また，RMSE については，今回の設定では標準誤差推定値の平均とほぼ同じ値を示しており，標準誤差の推定にバイアスが無く，標準誤差推定値の参照にも問題がないことを示唆している．テスト間で困難度の平均は異なっているが，推定値や標準誤差推定値のバイアスは認められなかった．

等化切片の推定精度を個別に検討すると，テストに含まれる項目数が 7 つの場合 (No.13, 42, 47) は推定精度が相対的に高く，項目数が 3 つの場合 (No.51, 54) には，推定精度が相対的に低くなっていることが認められる．この傾向は適用事例での分析結果とも整合があり，結果の正当性を支持する結果と言える．

5. シミュレーション研究 2

本シミュレーション研究においては，1 母数モデルに関して，同時等化のための周辺最尤推定法の性質を複数の側面から確認することを目的とする．

5.1. セッティング

検討する要因とその水準として，

1) テスト数 T(2 水準; [i]30，[ii]60)
2) 各テストに含まれる項目数 J_t(3 水準; [i]3，[ii]5, [iii]7)
3) 被験者数 I(3 水準; [i]50，[ii]100，[iii]200)

を取り上げた (全 18 条件)．なお，項目母数 (等化前困難度) の真値は平均 0，標準偏差 0.8 の正規分布から乱数として発生させて定めた．また，等化切片の真値は平均 0，標準偏差 0.5 の正規分布から発生させて定めた．各条件で繰り返し数を 100 回とし，100 個のデータを同時等化と個別等化の 2 つの場合で推定した．区分求積法における区分点数と範囲はそれぞれ 60 個，$[-3.5, 3.5]$ とした．そして，得られた推定値に関して，シミュレーション 1 と同様の 3 つの指標を算出した．

5.2. 結果と考察

すべての条件において適切に収束していないと思われる推定値は認められず，100 個の推定値から計算されたシミュレーションの結果を図 2 から図 13 にまとめた．各図においては上段から下段に向けて項目数が 3, 5, 7 の場合の結果を示している．また，右列の RMSE に関する図において，バツの打点は s.e. の平均，白丸の打点は RMSE を表している．なお，横破線は RMSE の平均である．全体的傾向としては，18 の条件すべてにおいて推定値の平均は真値に近かった．また，RMSE と標準誤差の平均値の値も近く，今回の結果からも標準誤差の値を推測に用いることの妥当性が示された．

次に，シミュレーション結果についてテスト数，項目数，被験者数の 3 つの観点から検討する．第一にテスト数の影響はほとんど見られなかった．テスト数の増加によって RMSE の減少が確認できる条件もいくつかあるが，それらはシミュレーションデータの変動の範囲内である可能性が高い．テスト数の増加は推定母数の増加を意味するため，必ずしも推定精度の向上にはつながらなかったと考えられる．第二に，項目数については，水準間のテスト数の違いはわずかであったが推定結果にはその影響が現れていた．被験者数が多い場合には顕著ではないものの，全ての条件において項目数が増えるほど推定精度が高くなる傾向が認められた．最後に，被験者数が推定結果に与える影響は非常に明確であり，被験者数が増加するほど推定精度は高くなっている．被験者数が 4 倍になった場合には RMSE の値は 1/2 になっており，理論的にも適切な推定結果を得られることが示されている．

同時推定と個別推定による推定結果の違いとしては，多くの場合に同時推定のほうが若干推定精度が高いことが挙げられる．多くのテストを同時に等化することで，推定に利用する情報が増えて推定精度がわずかばかり向上したのだと思われる．

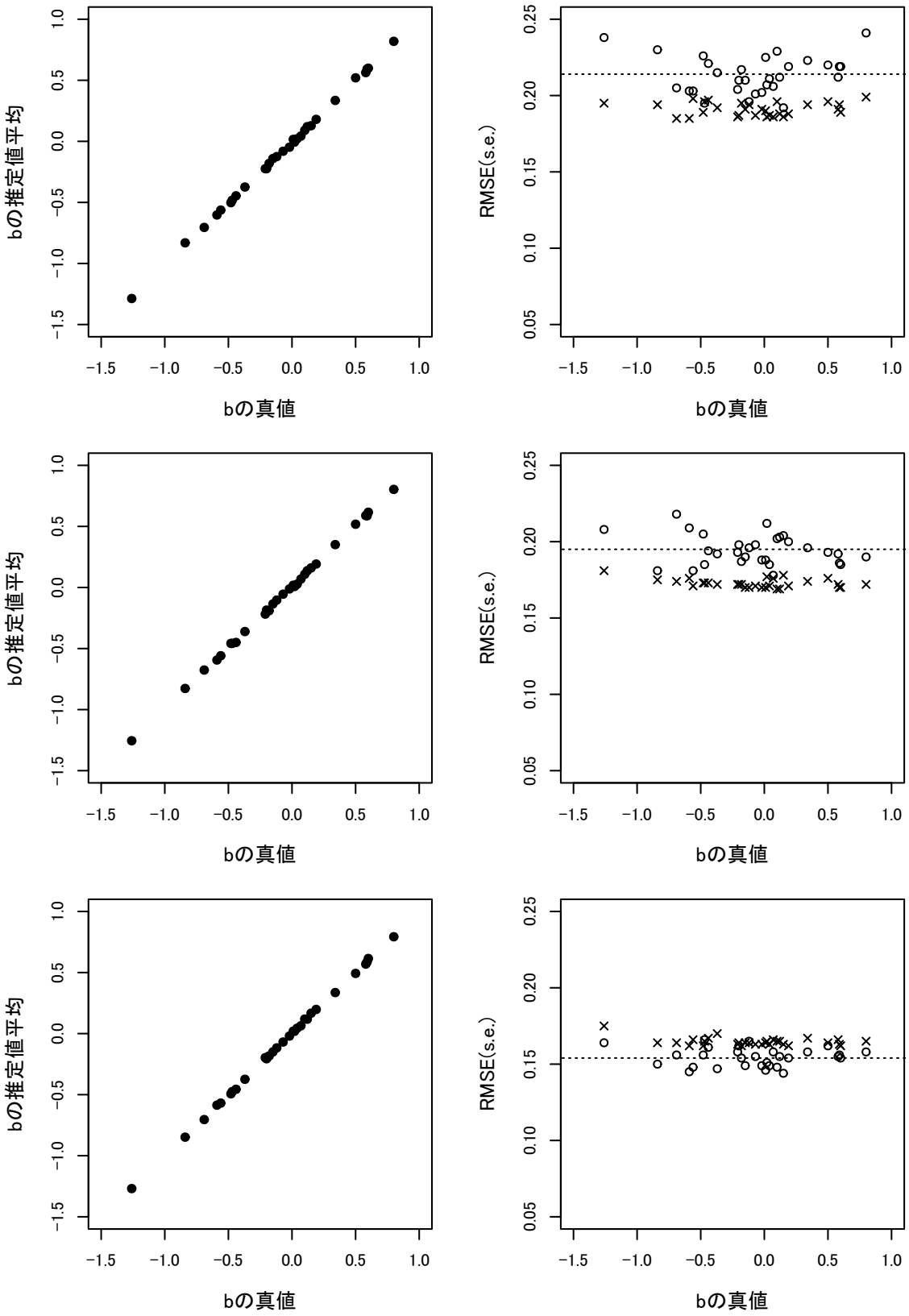

図 2: 被験者数 50，テスト数 30，項目数 [上段から 3, 5, 7]，同時推定の結果

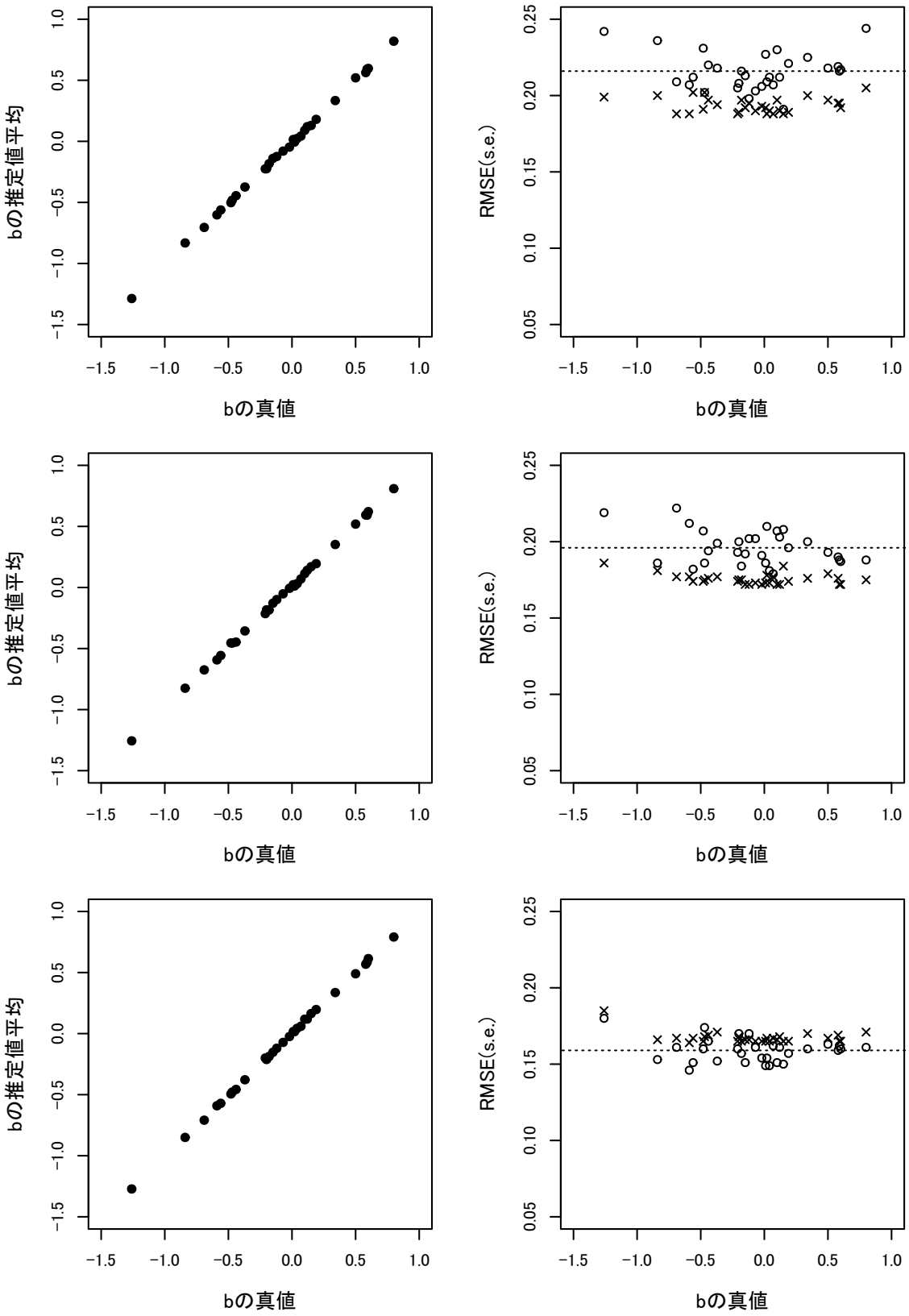

図 3: 被験者数 50，テスト数 30，項目数 [上段から 3, 5, 7]，個別推定の結果

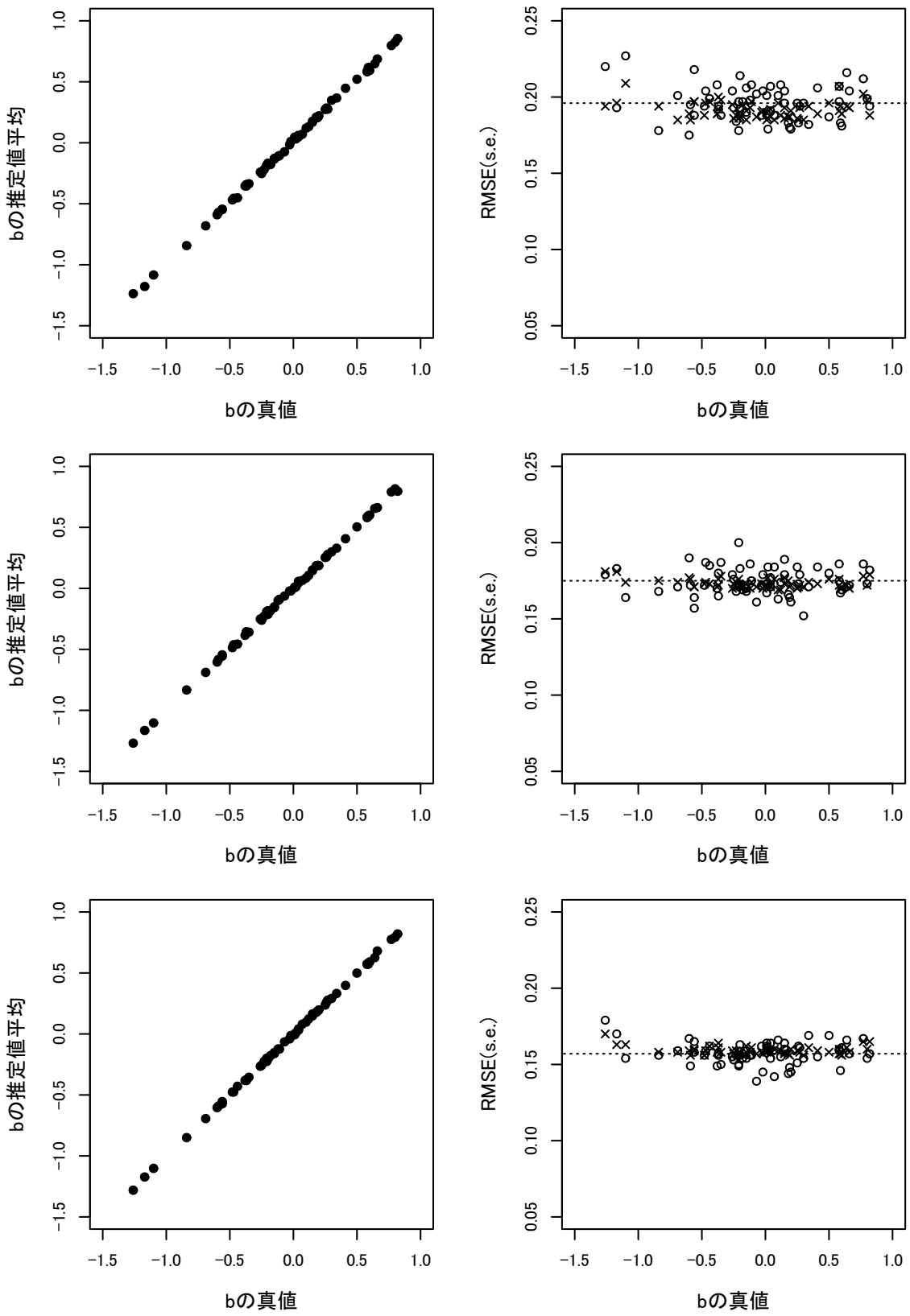

図 4: 被験者数 50，テスト数 60，項目数 [上段から 3, 5, 7]，同時推定の結果

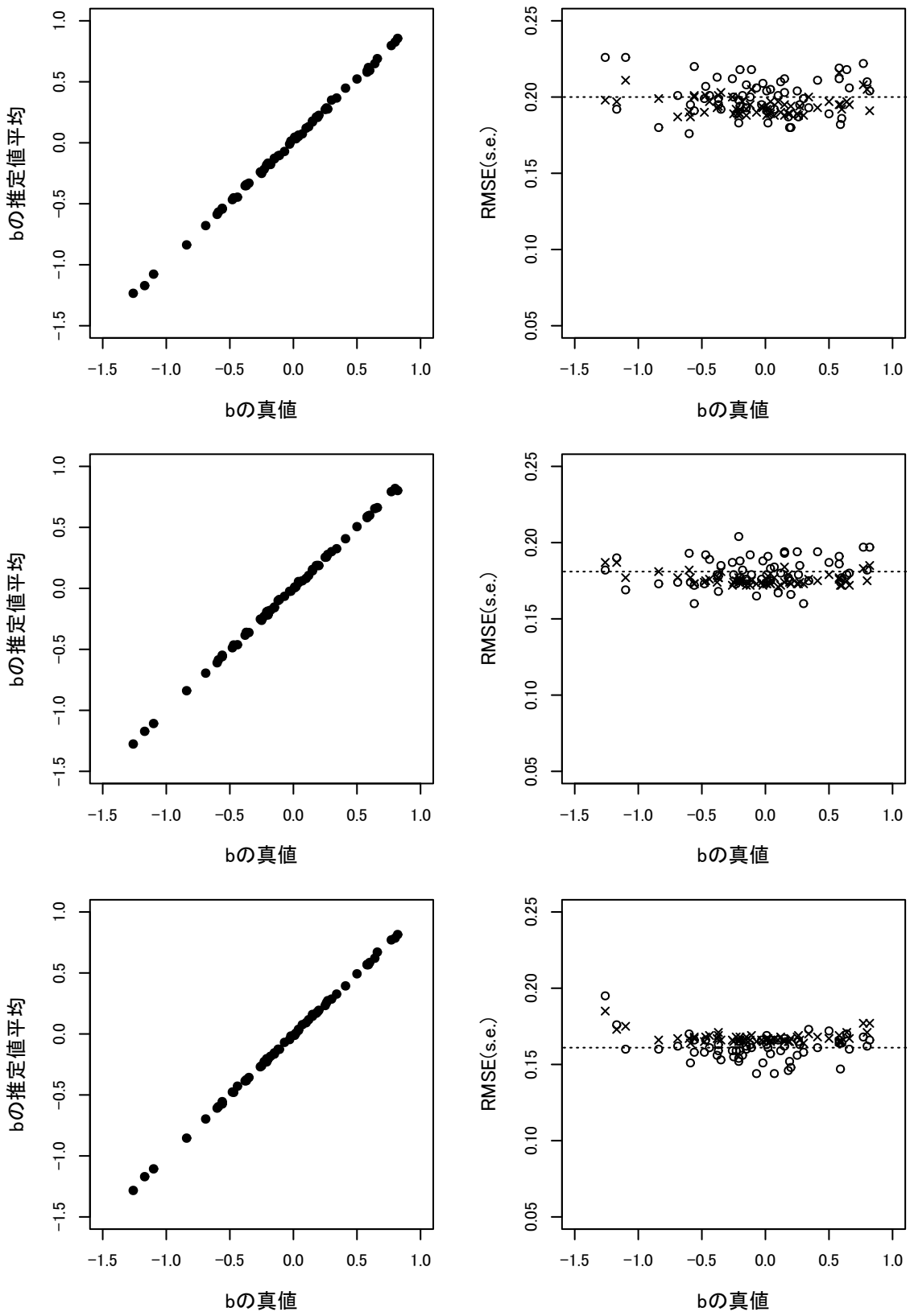

図 5: 被験者数 50，テスト数 60，項目数 [上段から 3, 5, 7]，個別推定の結果

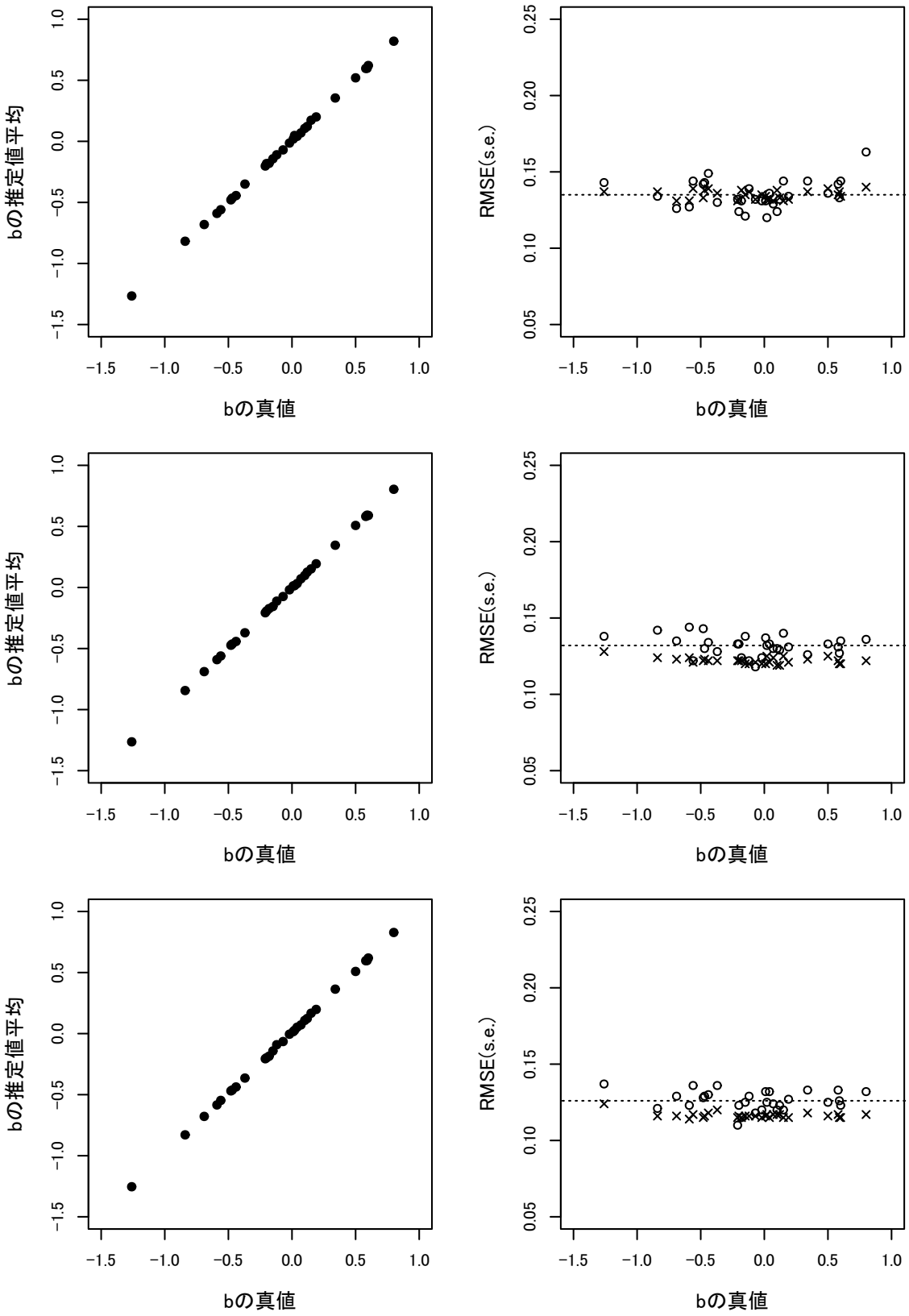

図 6: 被験者数 100，テスト数 30，項目数 [上段から 3, 5, 7]，同時推定の結果

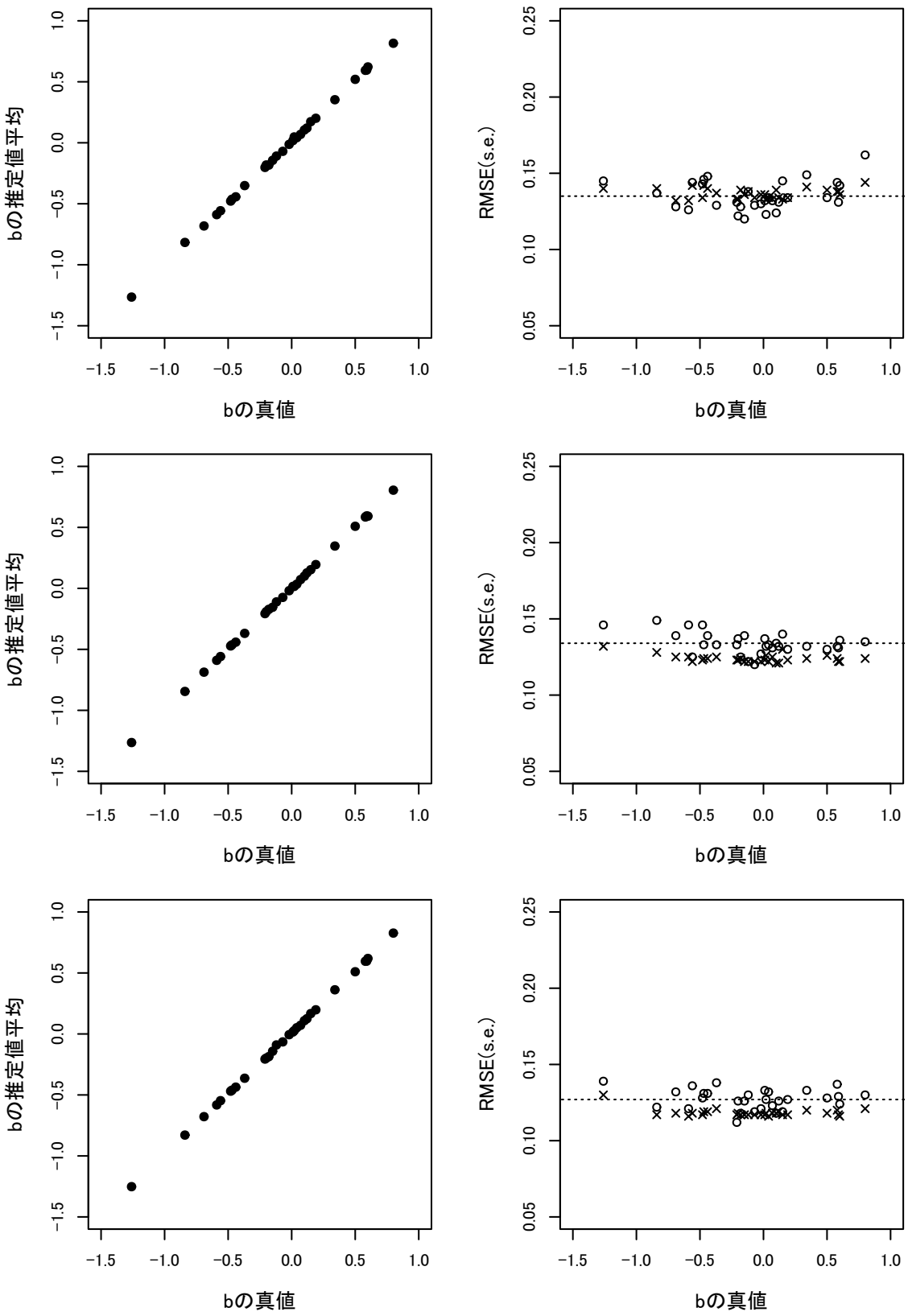

図 7: 被験者数 100，テスト数 30，項目数 [上段から 3, 5, 7]，個別推定の結果

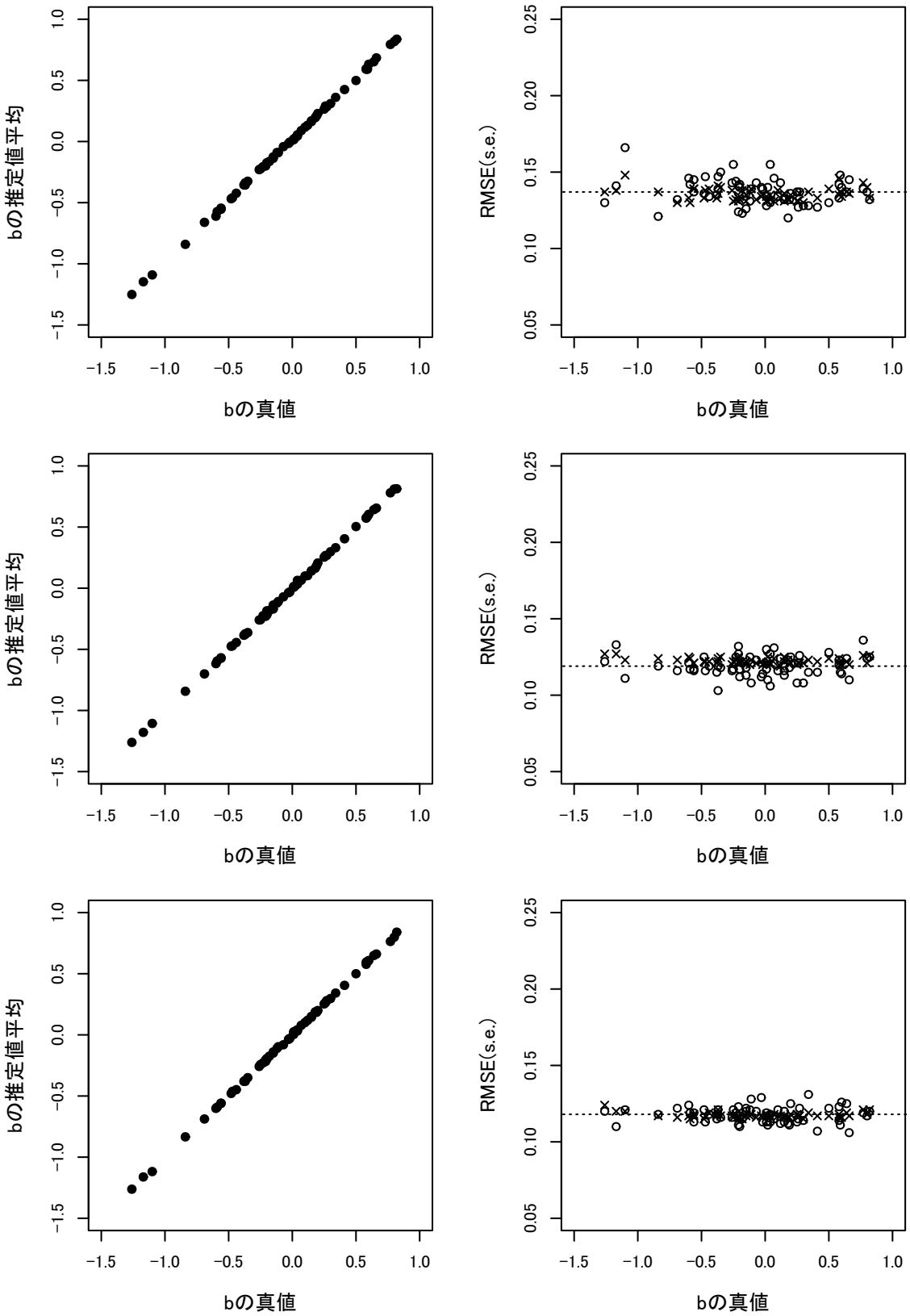

図 8: 被験者数 100，テスト数 60，項目数 [上段から 3, 5, 7]，同時推定の結果

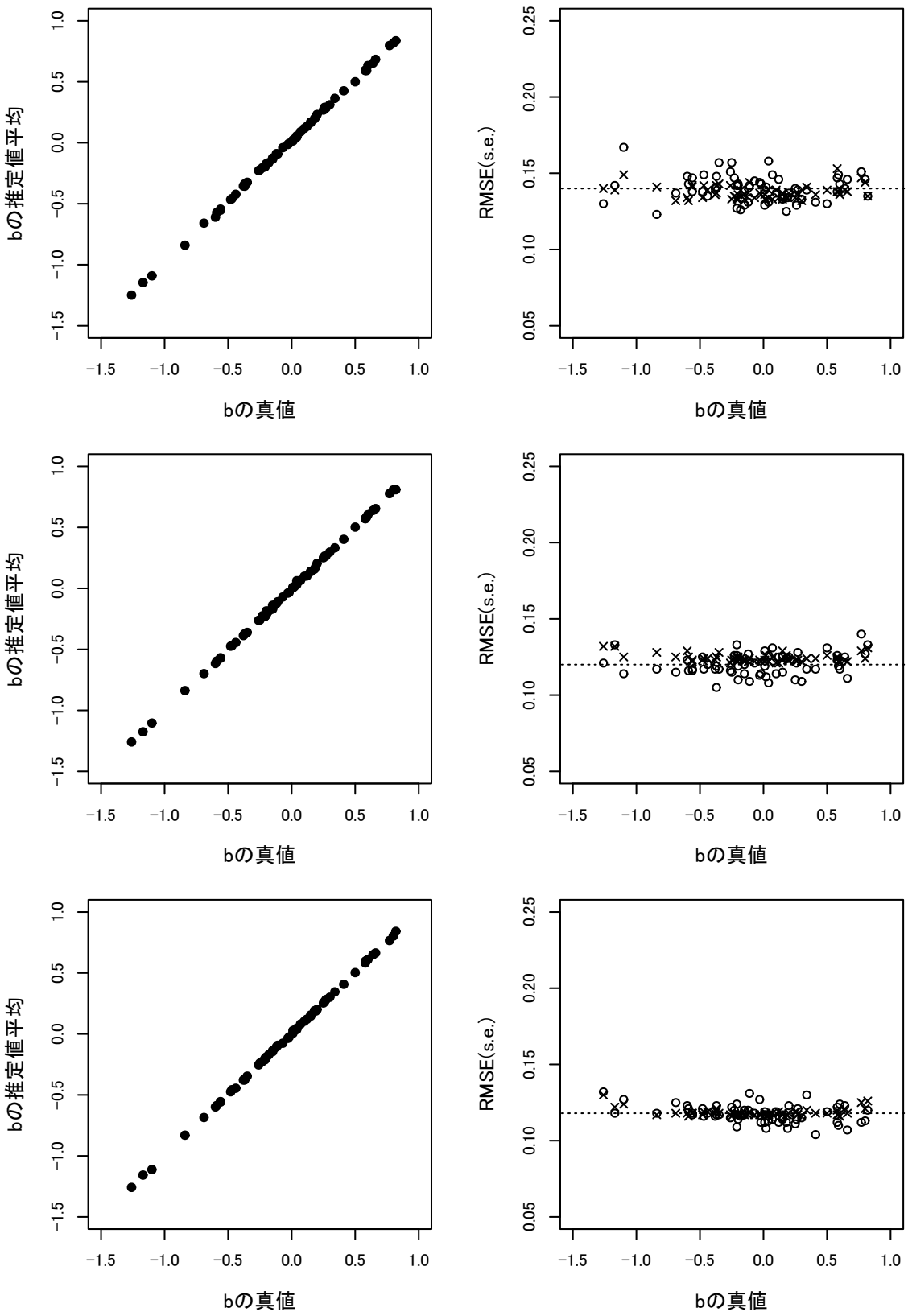

図 9: 被験者数 100，テスト数 60，項目数 [上段から 3, 5, 7]，個別推定の結果

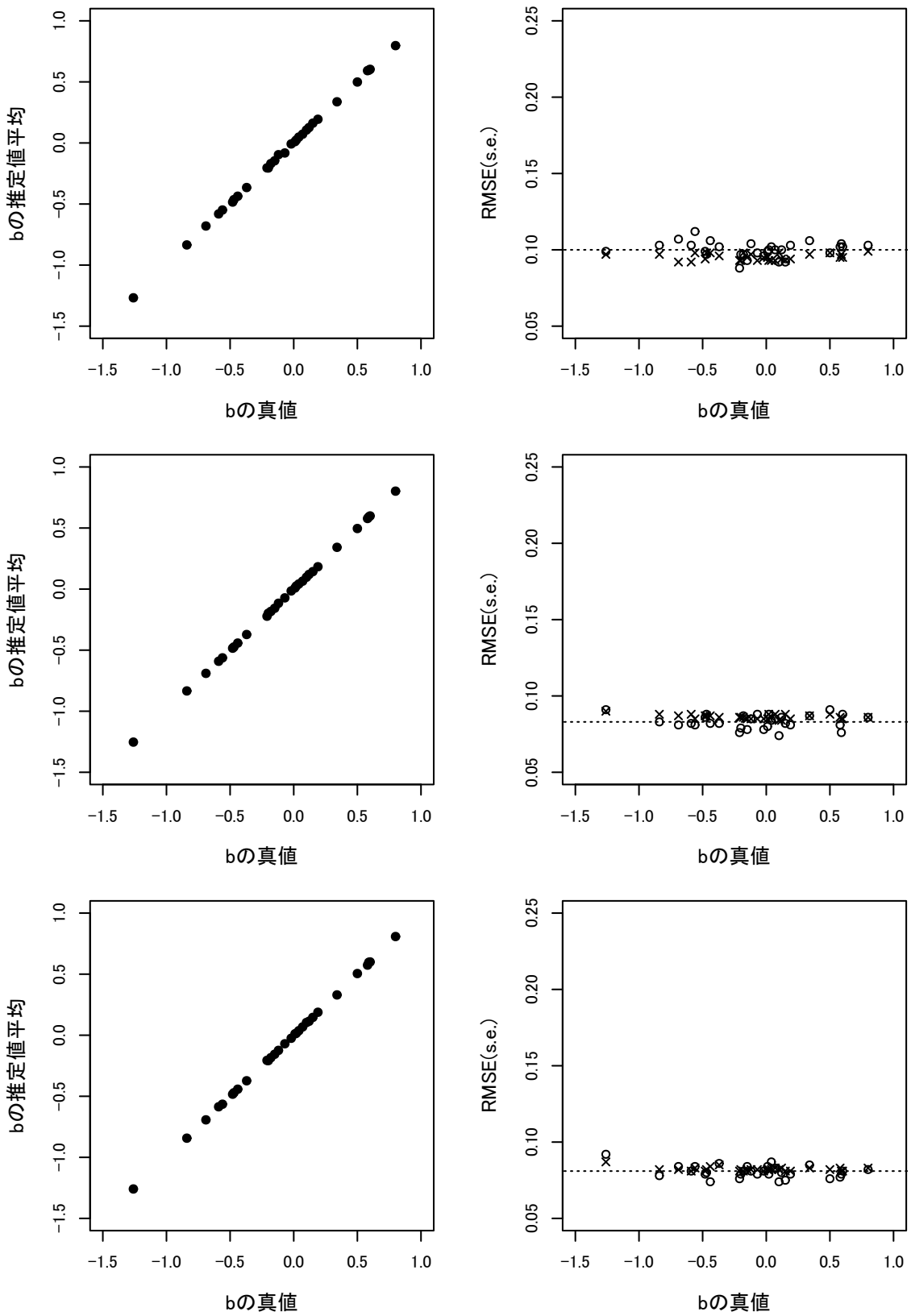

図 10: 被験者数 200，テスト数 30，項目数 [上段から 3, 5, 7]，同時推定の結果

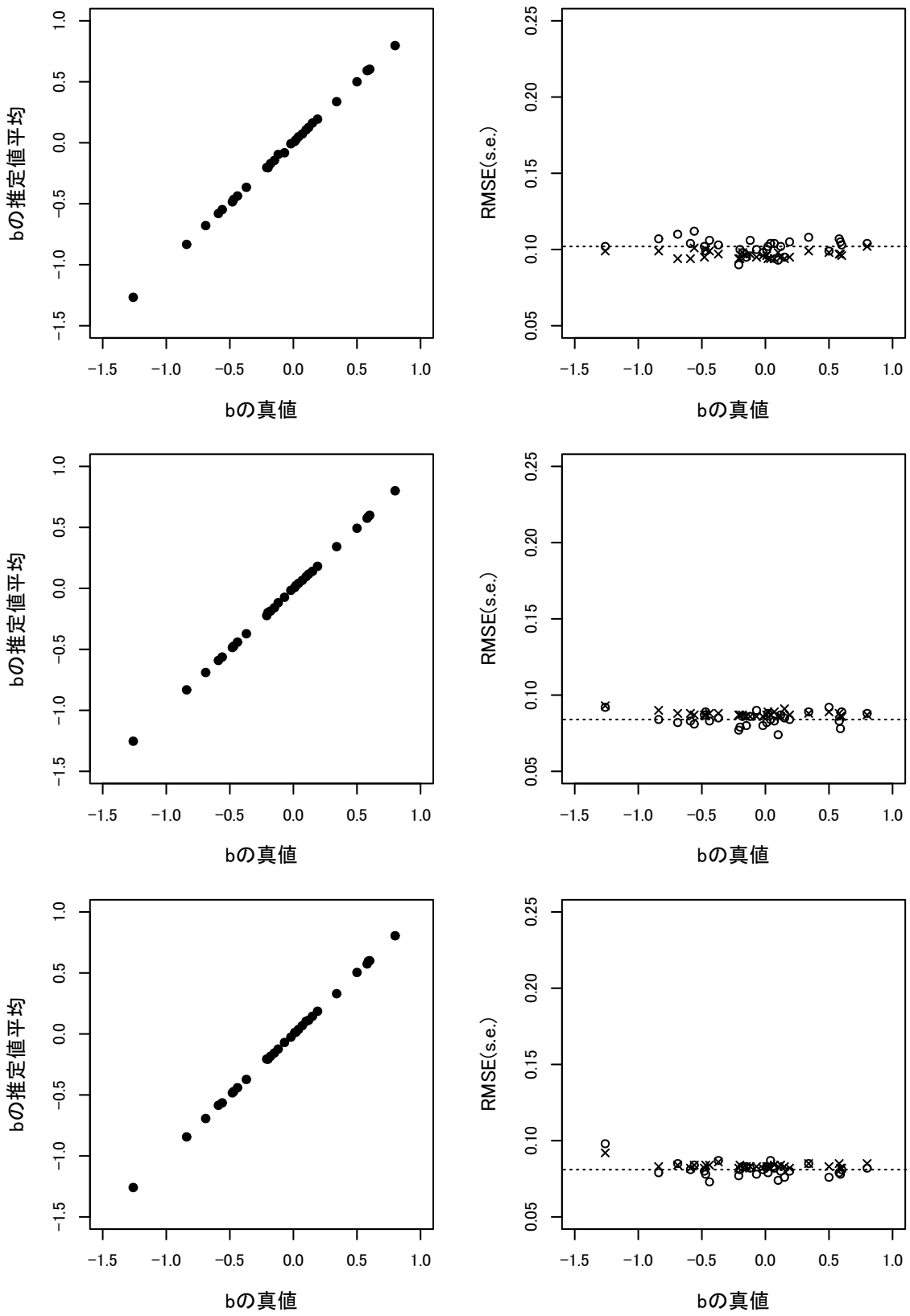

図 11: 被験者数 200，テスト数 30，項目数 [上段から 3, 5, 7]，個別推定の結果

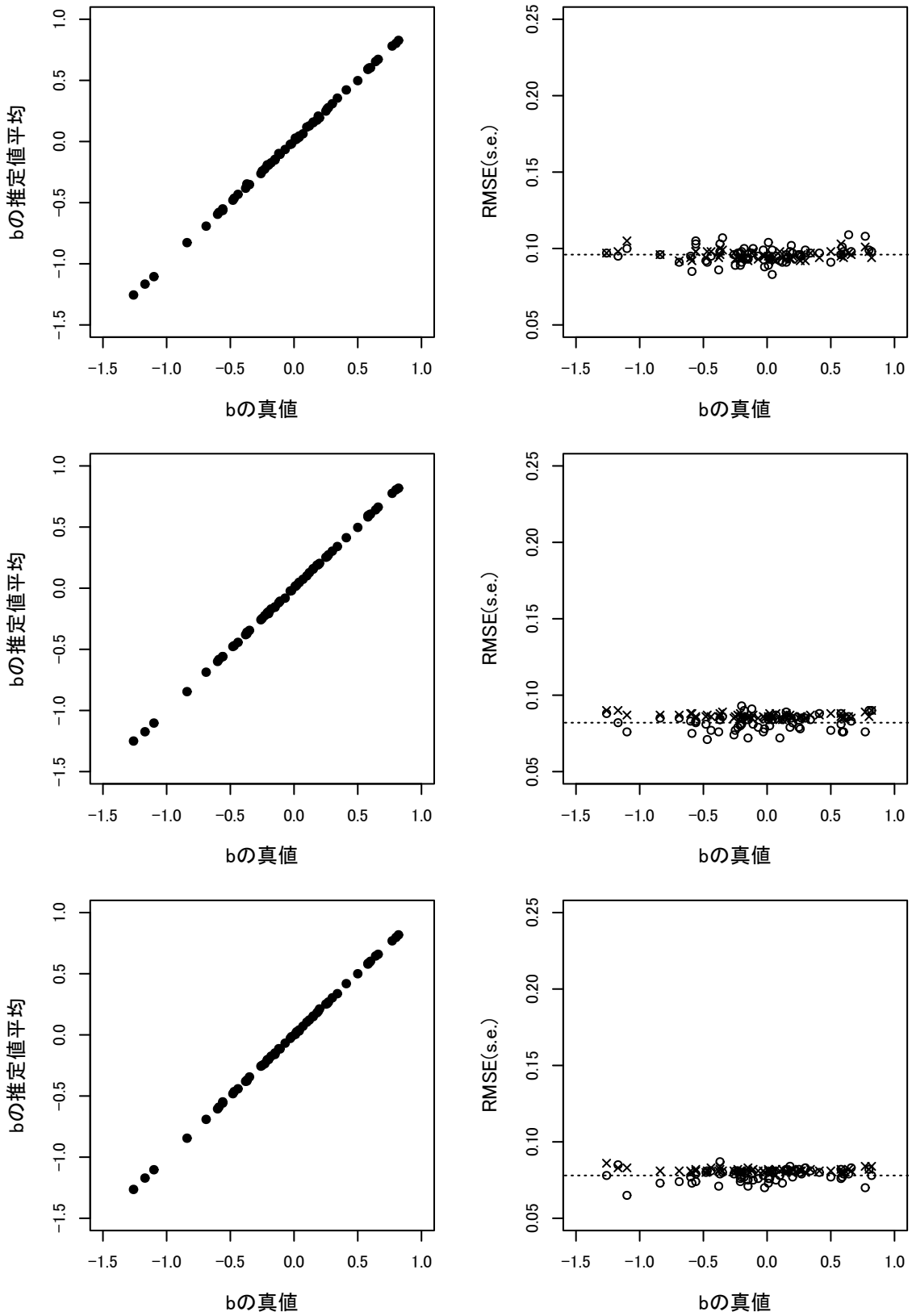

図 12: 被験者数 200，テスト数 60，項目数 [上段から 3, 5, 7]，同時推定の結果

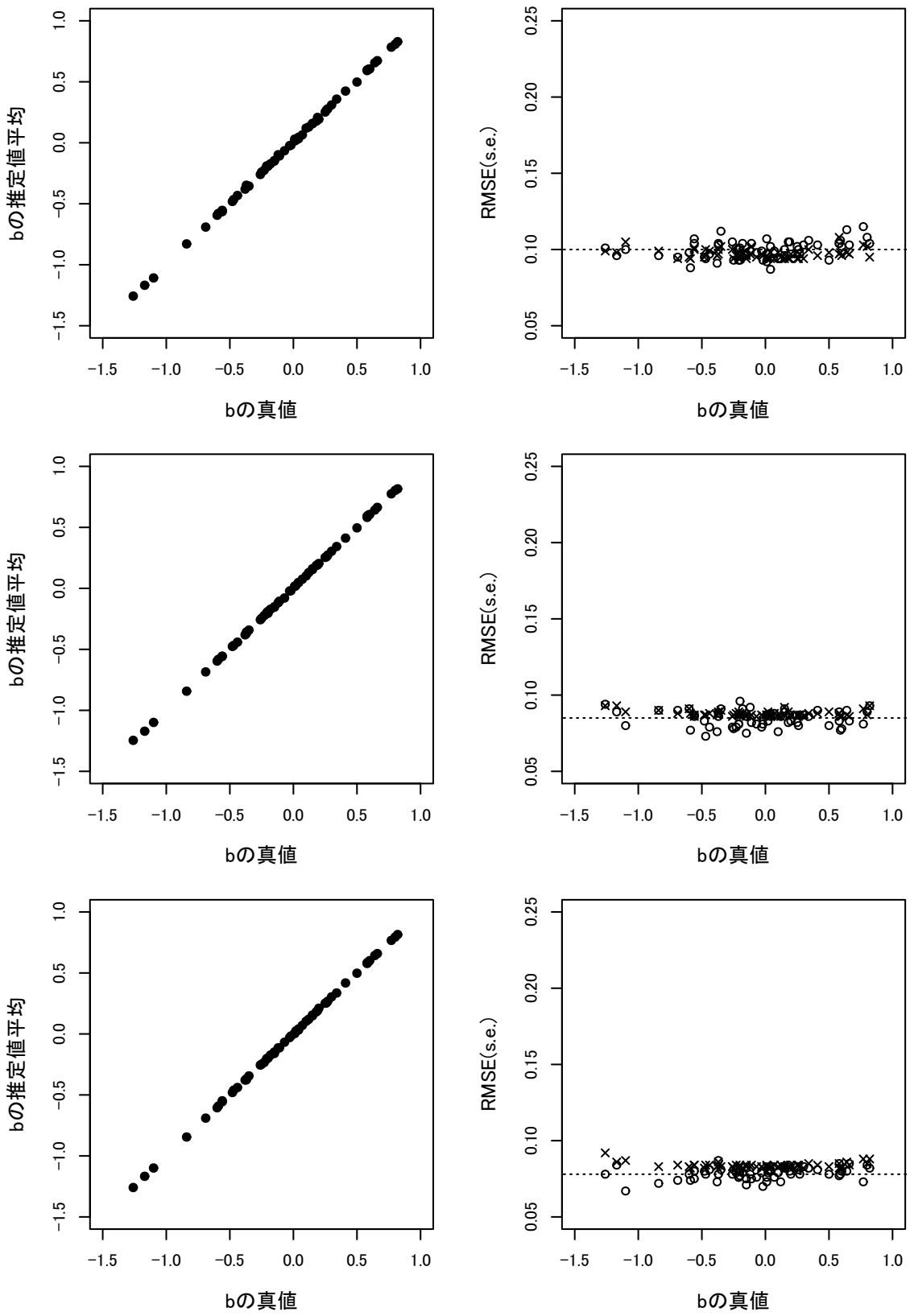

図 13: 被験者数 200，テスト数 60，項目数 [上段から 3, 5, 7]，個別推定の結果

6. まとめ

適用事例，シミュレーション研究 1，シミュレーション研究 2 を通じて，提案した多数の潜在特性尺度の同時等化法とそのための等化切片の推定法が有用であることが示された．

適用事例とシミュレーション研究 1 の結果から，本手法は実用に十分耐えうると判断され，推定結果が妥当であることも示された．また，シミュレーション研究 2 の結果からは，同時等化のための周辺最尤推定法の性質が統計的な観点からみても望ましいものであることが示唆された．シミュレーション研究では特性値 θ の推定までは検討していないが，本研究では等化において切片だけを扱っているから，θ の推定における等化による誤差は切片の標準誤差として考えればよい．シミュレーション研究 1 での各等化係数の標準誤差は 0.1 程度であったが，たとえば 54 のテストから偏りなくランダムに 54 項目を選んでテストを構成すればその影響は 0.014（$= 0.1/\sqrt{54}$）ほどになるから実用的には大きな問題はないと考えられる．本研究での提案手法により，多数のテストを効率よく等化することができ，IRT を用いたテスト運用への切り替えコストの軽減が十分に期待できるだろう．

参考文献

[1] D.R. Bock and M.F. Zimowski: Multiple group IRT. In W.J. van der Linden and R.K. Hambleton (eds.): *Handbook of Modern Item Response Theory* (Springer-Verlag, New York, NY, 1997).

[2] R.K. Hambleton, H. Swaminathan, and H.J. Rogers: *Fundamentals of Item Response Theory* (Sage, Newbury Park, CA, 1991).

[3] P.W. Holland and N.J. Dorans: Linking and equating. In R.L. Brennan (ed.): *Educational Measurement, Fourth Edition* (Greenwood, Westport, CT, 2006).

[4] 池田央: 現代テスト理論 (朝倉書店, 1994).

[5] M.J. Kolen and R.L. Brennan: *Test Equating, Scaling, and Linking: Methods and Practices, Second Edition* (Springer, New York, NY, 2004).

[6] F.M. Lord: *Applications of Item Response Theory to Practical Testing Problems* (Lawrence Erlbaum Associates, Nahwah, NJ, 1980).

[7] G.L. Marco: Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, **14** (1977), 139–160.

[8] 野口裕之: 被験者の反応パタンを利用した潜在特性尺度等化法. 教育心理学研究, **34** (1986), 315–323.

[9] 野口裕之: 共通被験者デザインにおける等化係数の周辺最尤法による推定. 名古屋大学教育学部紀要, **37** (1990), 191–198.

[10] H. Ogasawara: Marginal maximum likelihood estimation of item response theory (IRT) equating coefficients for the common-examinee design. *Japanese Psychological Research*, **43** (2001), 72–82.

[11] N.S. Petersen, M.J. Kolen, and H.D. Hoover: Scaling, norming, and equating. In R.L. Linn (ed.): *Educational Measurement, Third Edition* (Macmillan, New York, NY, 1989).

[12] 豊田秀樹: 被験者の推定尺度値とテスト情報関数を利用した潜在特性尺度の等化法. 教育心理学研究, **34** (1986), 163–167.

[13] 豊田秀樹: テスト情報関数を利用した潜在特性尺度の同時等化法. 東京大学教育学部紀要, **27** (1987), 293–295.

[14] 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, 2012).

[15] M. von Davier and A.A. von Davier: A general model for IRT scale linking and scale transformations. In A.A. von Davier (ed.): *Handbook of Modern Item Response Theory* (Springer, New York, NY, 2010).

[16] 吉村宰, 荘島宏二郎, 杉野直樹, 野澤健, 清水裕子, 齋藤栄二, 根岸雅史, 岡部純子, サイモンフレイザー: 大学入試センター試験既出問題を利用した共通受験者計画による英語学力の経年変化の調査. 日本テスト学会誌, **1** (2005), 51–58.

付録：$\boldsymbol{\beta}$ の推定における周辺対数尤度関数の勾配ベクトルの導出

周辺対数尤度関数の勾配ベクトルの導出にあたり，まずは β_t についての周辺対数尤度関数の偏導関数を求める．

$$\frac{\partial \log L^*(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta})}{\partial \beta_t} = \sum_{i=1}^{I} \frac{\partial}{\partial \beta_t} \log L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta}) \tag{6.1}$$

$$= \sum_{i=1}^{I} w^{-1} \frac{\partial}{\partial \beta_t} \left[\sum_{m=1}^{M} L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta}) h(x_m) \right] \tag{6.2}$$

$$= \sum_{i=1}^{I} w^{-1} \frac{\partial}{\partial \beta_t} \left[\sum_{m=1}^{M} \prod_{t=1}^{T} \prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t) h(x_m) \right] \tag{6.3}$$

ただし，$w = L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})$ とする．(6.3) 式の偏導関数に関して

$$\frac{\partial}{\partial \beta_t} \left[\sum_{m=1}^{M} \prod_{t=1}^{T} \prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t) h(x_m) \right] = \sum_{m=1}^{M} h(x_m) \frac{\partial}{\partial \beta_t} \prod_{t=1}^{T} \prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t) \tag{6.4}$$

$$= \sum_{m=1}^{M} \left[h(x_m) \left\{ \frac{\partial}{\partial \beta_t} \prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t) \right\} \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{\prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t)} \right] \tag{6.5}$$

$$= \sum_{m=1}^{M} \left[h(x_m) \left\{ \prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t) \sum_{j=1}^{J_t} \frac{\partial}{\partial \beta_t} \log L(u_{itj} \mid x_m, b_{tj}, \beta_t) \right\} \times \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{\prod_{j=1}^{J_t} L(u_{itj} \mid x_m, b_{tj}, \beta_t)} \right] \tag{6.6}$$

が導かれる．これを (6.3) 式に代入して整理することで

$$\frac{\partial \log L^*(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta})}{\partial \beta_t} = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{j=1}^{J_t} w^{-1} L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta}) \frac{\partial}{\partial \beta_t} \log L(u_{itj} \mid x_m, b_{tj}, \beta_t) h(x_m) \tag{6.7}$$

となる．さらに

$$\frac{\partial}{\partial \beta_t} \log L(u_{itj} \mid x_m, b_{tj}, \beta_t) = D \left\{ P(u_{itj} = 1 \mid x_m, b_{tj}, \beta_t) - u_{itj} \right\} \tag{6.8}$$

なのでこれを (6.7) 式に代入して整理すれば

$$\frac{\partial \log L^*(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta})}{\partial \beta_t}$$

$$= \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{j=1}^{J_t} w^{-1} L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta}) D \left\{ P(u_{itj} = 1 \mid x_m, b_{tj}, \beta_t) - u_{itj} \right\} h(x_m) \tag{6.9}$$

$$= \sum_{i=1}^{I} \sum_{m=1}^{M} Dh(x_m) \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})} \sum_{j=1}^{J_t} \left\{ P(u_{itj} = 1 \mid x_m, b_{tj}, \beta_t) - u_{itj} \right\} \tag{6.10}$$

$$= \sum_{i=1}^{I} \sum_{m=1}^{M} Dh(x_m) \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})} r_t \tag{6.11}$$

が導かれる．ここで，$r_t = \sum_{j=1}^{J_t} \{ P(u_{itj} = 1 \mid x_m, b_{tj}, \beta_t) - u_{itj} \}$ とする．

以上の結果から，周辺対数尤度関数の勾配ベクトルは以下のように得られる．

$$\nabla_{\boldsymbol{\beta}} = \frac{\partial \log L^*(\boldsymbol{U} \mid \boldsymbol{B}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{I} \sum_{m=1}^{M} Dh(x_m) \frac{L(\boldsymbol{u}_i \mid x_m, \boldsymbol{B}, \boldsymbol{\beta})}{L^*(\boldsymbol{u}_i \mid \boldsymbol{B}, \boldsymbol{\beta})} \boldsymbol{r} \tag{6.12}$$

$$\boldsymbol{r} = [r_1, \ldots, r_t, \ldots, r_T]' \tag{6.13}$$

豊田秀樹
早稲田大学
文学学術院
〒 162-8644 東京都新宿区戸山 1-24-1
E-mail: toyoda@waseda.jp

ABSTRACT

EFFORT TO CUT COSTS FOR SWITCHING TO TESTS BASED ON ITEM RESPONSE THEORY —UTILIZING SIMULTANEOUS EQUATING METHOD FOR A NUMBER OF LATENT SCALES—

Hideki Toyoda
Waseda University

Norikazu Iwama
The Japan Foundation

Ayako Nakamura
former e-communications, Inc.

Yasuhiro Saito
e-communications, Inc.

When switching from tests based on classical test theory to ones based on item response theory, it costs a lot to do surveys for item pool construction. In this article, we suppose the circumstances where we have many kinds of tests based on classical test theory that measure one latent trait and already have the data of the tests. Then, we show the way of cutting the costs for switching to tests based on item response theory and constructing a large item pool under such circumstances. We show how to compose a test and propose methods of estimating equating coefficients, which are necessary for item pool construction. We examined our method through its application to real data and simulation data. In the real data analysis, we succeeded to obtain a large item pool by a one-time survey. We also confirmed the statistical propriety of our estimation method and the validity of the result of the real data analysis in the simulation.