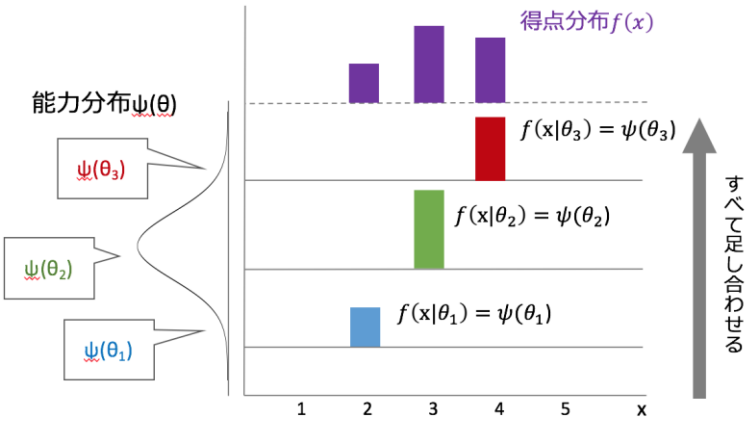

図２　IRT true score の度数分布生成イメージ

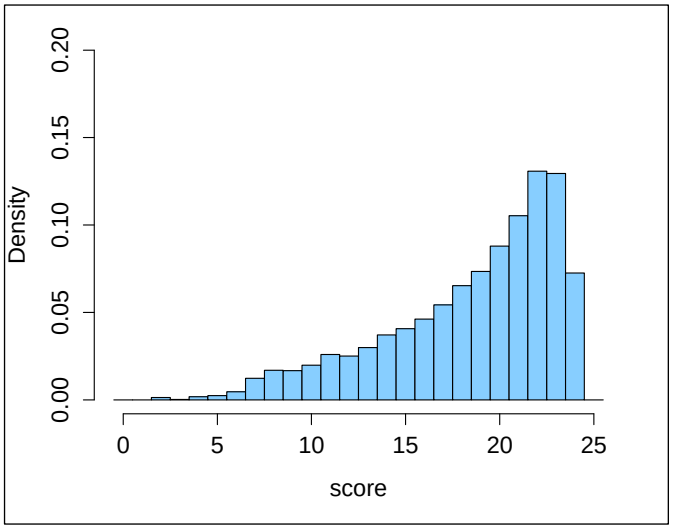

図３　IRT true score の相対度数分布の例

2.2.2　IRT observed score とその分布の生成

IRT true score が能力 θ_i の受検者に対しただ一つ求められる期待テスト得点である一方で，IRT observed score は，全ての項目反応パターンを考慮し，能力 θ_i の受検者がとりうる得点の確率分布として求められる。IRT observed score は，計算アルゴリズムとして Lord & Wingersky（1984）の Recursion Formula を用いることで求められる。

能力 θ_i の受検者が，項目 j に正答する確率を $P_j(\theta_i)$ としたとき，誤答となる確率 $Q_j(\theta_i)$ は $1-P_j(\theta_i)$ と表される。Lord & Wingersky の Recursion Formula では，正答確率 $P_j(\theta_i)$ と誤答確率 $Q_j(\theta_i)$ を用いて，そのテストに含まれる項目に対するすべての反応パターンを考慮することで，能力 θ_i の受検者がとりうる得点の確率分布を求めることができる。

今，項目数２のテストにおいて，能力 θ_i の受検者が正答数得点 x をとる確率を考える。項目数が２のとき，得点 x がとり得る値は０，１，２のいずれかである。x=0 となる場合は，項目１と項目２の両

方が誤答である。能力 θ_i の受検者が項目１と項目２に正答する確率をそれぞれ P_1, P_2 とすると，x =0 となる確率は $(1\text{-}P_1)(1\text{-}P_2)$ である。項目１と２の両方に正答し x=2 となる場合，その確率は P_1P_2 である。また，x=1 となる場合は，項目１が正答で項目２が誤答の場合か，項目１が誤答で項目２が正答の場合である。したがって，x=1 となる確率は $P_1(1\text{-}P_2)+(1\text{-}P_1)P_2$ と表せる。つまり，項目数２のテストにおいて能力 θ_i の受検者がとりうる得点の確率分布 $f_2(x|\theta_i)$ は，

$$f_2(x|\theta_i)=\begin{cases} f_2(0|\theta_i)=(1-P_1)(1-P_2), \\ f_2(1|\theta_i)=P_1(1-P_2)+(1-P_1)P_2, \\ f_2(2|\theta_i)=P_1P_2, \end{cases} \qquad (5)$$

となる。次に，項目数３の場合を考える。項目数が３のとき，得点 x がとり得る値は，０,１,２,３である。よって，項目数３のテストにおいて能力 θ_i の受検者がとりうる得点の確率分布 $f_3(x|\theta_i)$ は，（2）式を用いると，

$f_3(0|\theta_i)=f_2(0|\theta_i)(1-P_3)$,

$$f_3(x|\theta_i)=\begin{cases} f_3(1|\theta_i)=f_2(1|\theta_i)(1-P_3)+f_2(0|\theta_i)P_3, \\ f_3(2|\theta_i)=f_2(2|\theta_i)(1-P_3)+f_2(2|\theta_i)P_3, \\ f_3(3|\theta_i)=f_2(2|\theta_i)P_3, \end{cases} \qquad (6)$$

と表せる。これを一般化して考えると，項目数 r のテストにおける任意の能力 θ_i の受検者がとりうる得点の確率分布は，

$$f_r(x|\theta_i)=\begin{cases} f_{r-1}(x|\theta_i)(1-P_r), & (x=0), \\ f_{r-1}(x|\theta_i)(1-P_r)+f_{r-1}(x-1|\theta_i)P_r, & (0<x<r), \\ f_{r-1}(x-1|\theta_i)P_r, & (x=r), \end{cases} \qquad (7)$$

と表すことができる。（7）式が，Lord & Wingersky（1984）の Recursion Formula の一般式である。

次に，IRT observed score の度数分布（IRT observed score distribution，意訳；復元得点分布）を生成することを考える。受検者集団全体の得点分布を求めるには，各々が異なる能力 θ をもつ受検者全員について，それぞれがとりうる得点の確率分布を求め，最終的に足し合わせる必要がある。この操作を「周辺化」といい，その理論的な手続きは以下のように行う。簡単のため，以下の得点分布 $f_r(x|\theta_i)$ を，項目数 r を省略して $f(x|\theta_i)$ と表現する。

この場合の周辺化は，能力分布を考慮して行う。まず，能力 θ_i の受検者がとりうる得点の確率分布 $f(x|\theta_i)$ に $\psi(\theta_i)$ を掛け合わせることで，受検者全体に含まれる能力 θ_i の受検者集団の得点分布 $f(x|\theta_i)\psi(\theta_i)$ を求める。次に，受検者全体の得点分布を求めるために，それぞれの能力ごとの得点分布 $f(x|\theta_i)\psi(\theta_i)$ を累積する。したがって，受検者全体の得点分布は，

$$f(x) = \int_{\theta} f(x|\theta)\,\psi(\theta)\,d\theta, \qquad (8)$$

という関数で表すことができる。この操作により，能力 θ が積分消去される。(8) 式の場合，能力 θ の分布は連続分布として考えられている。しかし，実際には，受検者数には限界があるため，能力はとびとびの値をとる。したがって，実際に計算するときには能力 θ の分布を離散分布として扱う。この場合，受検者全体の得点分布$f(x)$は以下のように表される。

$$f(x) = \sum_{i} f(x|\theta_i)\,\psi(\theta_i), \qquad (9)$$

最後に，テストごとに受検者数が異なる場合でも比較可能にするために，(6) 式の分布を受検者数で割ることで復元相対度数分布を求める。これを数式で表すと以下の通りとなる。

$$f(x) = \frac{1}{N}\sum_{i} f(x|\theta_i), \qquad (10$$

周辺化をイメージ図で表したものが，図４の通りである。また，true score の得点分布のヒストグラムを描画した際と同様に，A 県の数学データを使用して observed score のヒストグラムを描画したところ，図５の通りとなった。

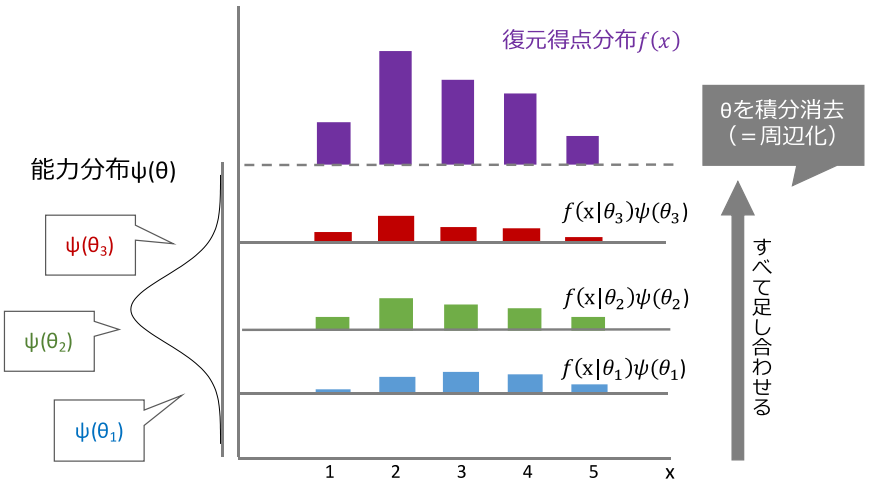

図４　周辺化のイメージ図

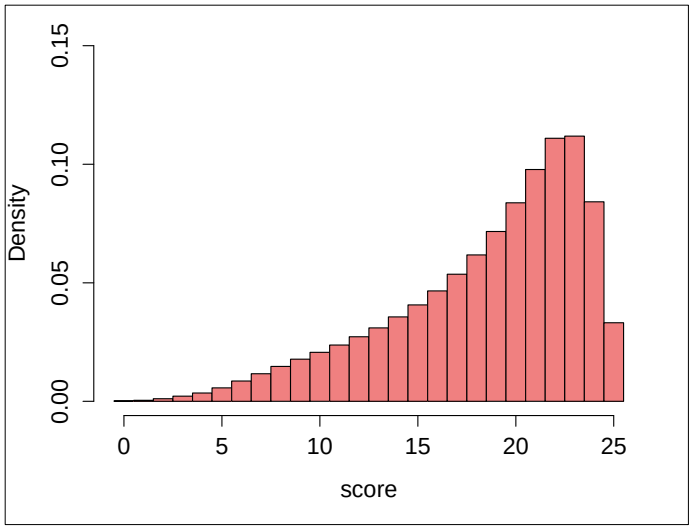

図 5　復元得点分布の例

2.3　復元得点分布の利用

最初にも述べたが，本研究で復元した得点の度数分布は，問題項目・受検者集団がともに異なる年度間の調査結果を等パーセンタイル法により対応づけする際に利用することを目的としている。そのため，復元された度数分布は，素得点の分布すなわち現実的な分布により近いものとなることが望ましい。

IRT モデルを介して得点分布を生成する際には，推定した母数を用いて求められる能力分布を（2）式や（6）式における能力分布$\psi(\theta)$として使用する。推定した能力分布を用いて求められた復元得点分布は，正答数得点（素得点）の分布に比べて標準偏差が大きく，系統的な歪みが生じると言われている（Han et al., 1997）。この分布の歪みは，実際の能力分布の代わりに推定した能力分布を使ったことに起因すると考えられている。よって，復元得点分布は素得点の分布とは必ずしも一致しないと言える。一方で，復元得点分布は，受検者のすべての項目反応パターンを考慮して生成するため，期待テスト得点を用いて生成した IRT true score の度数分布に比べ，より現実的な分布に近づくと予想される。本研究では，A 県の数学データを用いて IRT true score の分布と復元得点分布，素得点の分布をシミュレーション比較することで，復元得点分布が素得点の分布，すなわち現実的な分布にどれくらい近い分布となるかを検証した。なお，能力分布$\psi(\theta)$には，MAP 推定法により推定した能力母数を用いた。

三つの分布について累積相対度数分布を生成し，重ねて描画したところ，図 6 の通りとなった。

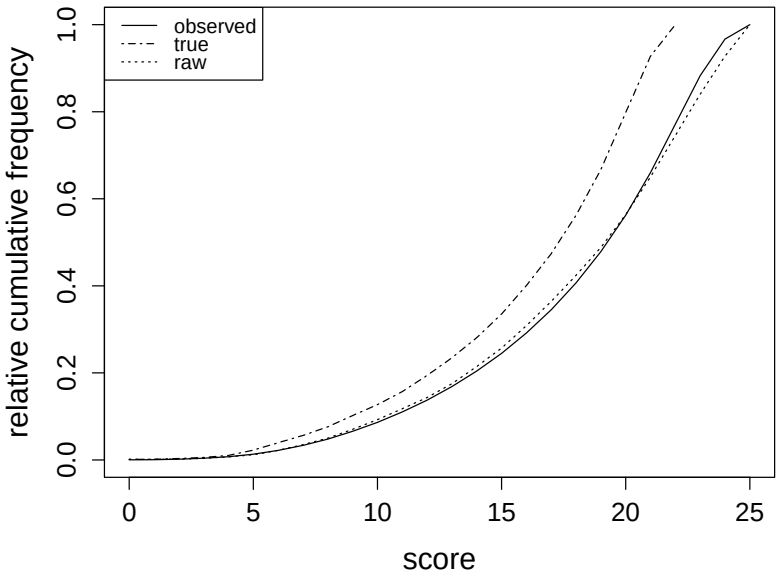

図 6　累積相対度数分布による比較

図 6 において，復元得点分布と素得点の分布は，理論的に言われているとおり完全には一致せず，若干のずれが生じている。しかし，そのずれはわずかであり，分布自体はほぼ同じ形状を示すことが明らかになった。さらに，true score の分布は素得点分布から大幅にずれていることがわかる。また，能力分布として MLE 推定法や EAP 推定法などの別の方法で推定した能力母数を使用した場合や標準正規分布を仮定した場合，別年度の A 県における数学データを用いて同様の比較を行った場合にもこの傾向がみられた。したがって，より現実的な度数分布に近い分布を生成することが求められる場合には，復元得点分布の利用が適していると言える。

2.4　R における計算アルゴリズム

本研究では，以上の手続きを，R を用いて行う。R における計算アルゴリズムは以下の通りである。

① 推定した能力母数により能力分布を仮定する，

② 推定した項目識別力母数aと項目困難度母数b，①で仮定した能力分布を Recursion Formula を表す関数に代入し，受検者それぞれがとりうる得点の確率分布を求め，行列を作成する，（図 7）

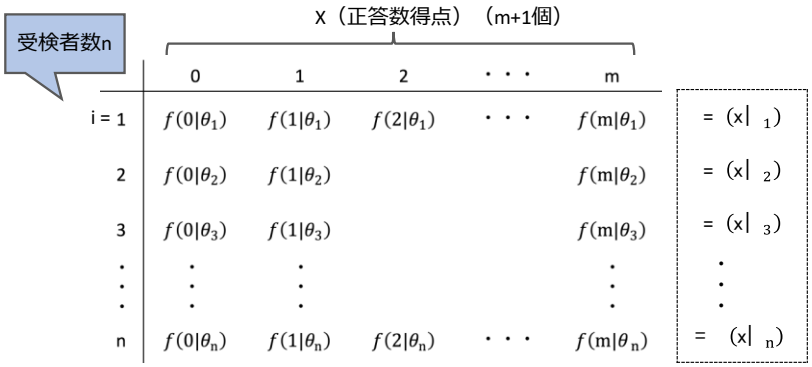

図 7　R において受検者全員分の得点分布を求める過程のイメージ図

③　②で作成した行列の転置行列に対し，行成分に受検者の数だけ１をもつベクトルを掛け合わせることで得点ごとの度数が縦一列に並んだ行列を作成する，（図８）

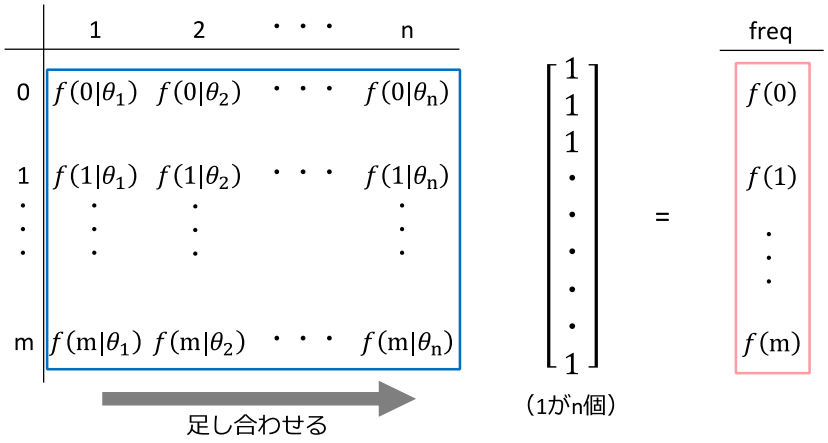

図８　R における周辺化の計算のイメージ図

④　③で作成した行列に正答数得点が縦一列に並んだ行列を結合させ，１列目の成分に得点，２列目の成分に度数をもつ行列を作成する，

⑤　④の行列をもとに，復元得点分布のヒストグラムを近似的に描画する。細かい手順は，以下の通りである，

・④の行列の２列目の成分（度数）の小数点以下を四捨五入する。

・上の操作で整数値になった度数を④の行列の１列目の成分（得点）とともにベクトルに展開する。

・ヒストグラムの関数を使って描画する。

⑥　度数分布，相対度数分布，累積分布，累積相対度数分布をデータフレーム形式で出力する。

3. フォンノイマン棄却法を用いた推算値（Plausivle Values）算出のアルゴリズム

3.1 推算値

大規模学力調査における下位領域ごとの集団比較について考える。通常, 全米学力調査（National Assessment of Educational Progress, NAEP）をはじめとした国際的な学力調査では,試験のデザインに重複テスト分冊法が採用されているため，個人別の能力を測定することを目的とした試験に比べると各受検者が解答する項目数は少ない。さらに下位領域ごとの比較を目的とすると利用できる項目数はさらにしぼられることになる。そのため，個々の EAP（expected a posteriori）推定値や MLE（Maximum Likelihood Estimation）推定値から集団の能力分布を推定する方法の場合，分散の過大評価・過小評価が生じ，集団間の正確な分散の比較ができない。すなわち，各受検者が解答する項目数が少なく各受検者の能力値から集団統計量を算出する方法の場合, 正確に分散を評価できない。このことから，各集団の能力分布の平均値に統計的に有意な差があるかを判断することができないという問題などが生じる。

そこで本研究では, 推算値（plausible values）を用いた下位領域ごとの集団比較について検討する。推算値とは多重補完法（Multiple Imputation）に関する Rubin（1987）の理論的基礎をもとに，Mislevy ら（1991）により大規模アセスメントに適用された手法であり，現在 PISA や TIMSS をはじめとした国際的な学力テストにおいても採用されている。推算値は受検者の能力母数 θ の事後分布からの無作為に取り出した複数の能力値である。推算値を使う利点として，個人の能力の不確実性を考慮することができ，また項目数が少ない場合でも，集団の能力分布の分散を正確に推定できるとともに，能力分布のパーセンタイルも正確に推定することができる点が指摘できる。

3.2 von Neumann の棄却法（rejection method）

本研究では事後分布からの無作為抽出をおこなうにあたり von Neumann の棄却法（rejection method）を利用する。棄却法は特定の領域において乱数を発生させ採択域内であればその値を分布に従う乱数としてとり，それ以外の領域に発生していた場合棄却し任意の個数の乱数を得るまでそのサブルーチンを繰り返す手法である。この方法を利用することで一様乱数を用いて受検者の能力母数 θ の事後分布に従う標本を生成することができる。受検者の能力母数 θ の事後分布$\mathbf{h(\theta|x)}$は,

$$\mathbf{h(\theta|x)} = \frac{\boldsymbol{f(x|\theta)g(\theta)}}{\int \boldsymbol{f(x|\theta)g(\theta)d\theta}} \tag{1}$$

で表される。このとき受検者の項目反応パターンを x，能力母数を θ， θ が所与のときの項目反応パターン x の条件付き確率密度関数を f(x|θ)としている。また，事前分布は通常，正規分布 g（ θ ）~N（μ ,σ^2）を仮定している。

一様乱数を発生させる領域を設定する際，受検者の能力母数 θ の尺度値である EAP 推定値や事後分布の最大確率密度が既知である必要がある。そのため，棄却法を行うにあたって受検者の能力値に関する推定事後分布を求める必要がある。本研究では効率よく棄却サンプリングを行うため，一様乱数を発生させる領域を，横軸にあたる θ に関しては$[\theta_{EAP}$-4.75，θ_{EAP}+4.75]，縦軸にあたる確率密度を０から最大事後確率に 1.0001 をかけた値として設定する。

棄却法による推算値の算出の一連の過程を整理すると，

①θ_{EAP}±4.75 で一様乱数を発生させ θ の仮の値とする
②０から最大事後確率に 1.0001 をかけた値までの間で一様乱数を発生させる
③発生した乱数が採択域である分布関数に①の値を代入したものよりも小さければ推算値として採用し，それ以外の場合棄却する
④任意の数（本研究では各受検者に対し 10 個）の推算値が得られるまで繰り返す

となる。棄却法の概念図は図 1 で示したようになる。このとき赤で示した点が採択された乱数で，青で示した点が棄却された乱数を示している。

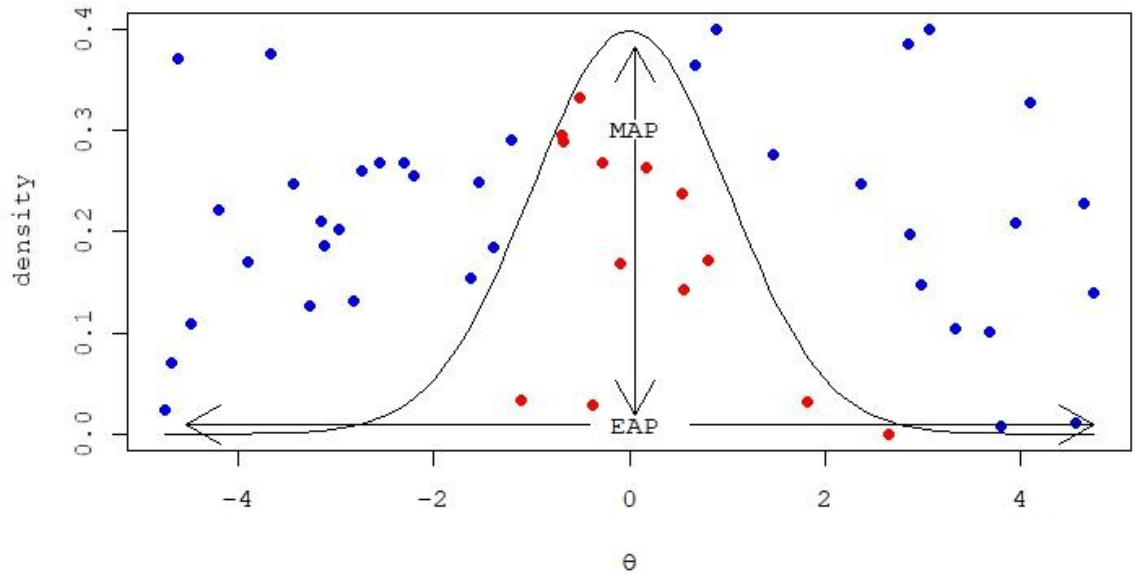

図 2 棄却法の概念図

3.3 推定事後分布

3.3.1 EAP 推定値と周辺分布

上述したように，推算値を得るために推定事後分布を求める必要がある。はじめに EAP 推定値と周辺分布を求める。EAP 推定値は θ の事後分布の期待値として，

$$\boldsymbol{\theta_{EAP}} = \int_{-\infty}^{\infty} \boldsymbol{\theta \mathrm{h}(\theta|\mathrm{x})} \boldsymbol{d\theta} \tag{2}$$

と定義される。ただしこの積分計算を解析的に行うことはできないため, 数値計算により近似的に解を求める。具体的には$L(X_i)$を$f(\mathrm{x}|\theta)$の対数尤度関数とし，ノード（分点）を $\mathrm{X_i}$,ノード数を n として，エルミート・ガウス求積法（Hermite-Gauss quadrature）を用いて算出する.

村木（2011）は分点について等間隔で区切るよりもエルミート・ガウス求積法を採用したほうが効率的に精度の高い値が得られる利点があることが報告している。本研究では分点数を 32 として行った。ガウス型公式は

$$\int_b^a \boldsymbol{w(x)f(x)dx} = \sum_{i=1}^{N} \boldsymbol{A_i f(x_i)} \tag{3}$$

であり，エルミート・ガウス求積法における区間［a,b］は［-は,∞］であり，重み関数$w(x)$は

$$\boldsymbol{w(x) = exp(-x^2)} \tag{4}$$

とあらわされ，重みA_iは,

$$\boldsymbol{A_i} = \frac{\boldsymbol{2^{n+1} n! \sqrt{\pi}}}{\boldsymbol{[H_n'(x_i)]^2}} \tag{5}$$

となる。求積点X_iにおける事後分布を表す重みG_iは

$$G_i = \frac{\boldsymbol{L(X_i)A_i}}{\sum_{i=1}^{n} \boldsymbol{L(X_i)\, A_i}}, \qquad \boldsymbol{(\mathrm{i} = 1, 2, \ldots \mathrm{n})} \tag{6}$$

となる。このとき$\sum_{i=1}^{n} \boldsymbol{L(X_i)\, A_i}$は周辺分布である。以上より EAP 推定値は,

$$\boldsymbol{\theta_{EAP}} = \sum_{i=1}^{N} \boldsymbol{X_i G_i} \tag{7}$$

で計算できる。このとき得られた EAP 推定値を基準に±4.75 したものを θ に関して発生させる乱数の上限と下限とする。

3.3.2 MAP 推定値

次に MAP 推定値を求める。本研究では加藤（2014）を参考に算出した。
ただし計算するにあたりに解析的に求めることが困難であることから適当な初期値から解が収束するまで繰り返し推定値を更新する数値計算により近似的に解を求める。数値計算の手法として，本研究では Newton-Raphson 法を修正した Fisher のスコアアルゴリズムおよび反復回数が 100 回を超えるものに関しては二分法を利用することで MAP 推定値を得る。

得られた θ_{MAP} に 1.0001 をかけたものを採択域である分布関数の確率密度における上限とする。このとき下限は 0 とする.

以上より推算値を算出するにあたり必要な乱数の発生領域の設定ができた。この発生領域に一様乱数を発生させ棄却法のアルゴリズムを利用することで推算値を受検者 n 人×10 個得る。

3.4　推算値の利用

得られた推算値の利用について考える。前述したように推算値は，受検者の能力母数 θ の事後分布から得られる無作為標本である。また，各受検者の θ に関しての事後分布を集めた分布は，その集団の能力分布からの指定分布を与える。そのため，受検者集団の推算値の組はその集団の能力分布からの無作為標本とみなすことができる。このため，推算値は集団統計量の不遍推定値が得られるとされる。

推算値を用いた集団の平均や分散などの統計的特性の推定は関心のある統計量を推算値ごとに計算し,それらの統計量を平均することで算出する.また K 組の推算値により算出した統計量の補間分散は Little&Rubin（2002）より，

$$\widehat{V_{IMP}} = \left(1 + \frac{1}{K}\right)\left[\frac{1}{K-1}\sum_{i}(M_{PVi} - \overline{M_{PV}})^2\right] + \frac{1}{K}\sum_{i}\widehat{V}(M_{PVi}) \qquad (9)$$

となる。このときM_{PVi}は集団 i における統計量，$\overline{M_{PV}}$は統計量の集団についての平均値，$\widehat{V}(M_{PVi})$は集団 i における統計量の誤差分散の推定値である。（9）式の正の平方根が補間の標準誤差となる。

推算値は上述してきたように集団を対象とした報告に適している。しかし，その一方で個人の能力推定には不向きであることに注意する必要がある。推算値はランダムサンプリングにより算出された値であることから，同じ反応パターンをもつ受検者が複数人いても，結果として推定される能力推定値が異なってしまう。したがって，推算値はあくまでも集団統計量について同じデータに関する二次的な分析に利用すべきであることに注意する必要がある。

3.5　アルゴリズムの概略

最後に本研究の計算に用いた R の計算アルゴリズムは，

①ガウス・エルミート求積の分点と重みを設定する
②EAP 推定値と周辺分布の計算を行う
③受検者能力分布の最大確率密度の計算を行う
④棄却法を利用して推算値を各受検者に対し 10 個得る⑤各集団の推算値の組ごとに求めた集団統計量を算出する

となる。

4. 対応づけによる本体調査の年度間比較の実際

本章ではここまでに開発した諸手法を組み合わせて，実際に平成 25 年度と平成 28 年度の全国学力・学習状況調査における各都道府県別の学力分布がいかなる変動を見せるかの実例を示す。ただし，都道府県はすべて匿名化しかつ提示順もランダム化している。また，本研究の目的はあくまでもノウハウ開発を主眼としているため，結果の解釈に当たっては，例えば，復元得点分布の復元精度や対応づけの精度などの検討は，その指標自体の理論的な検討も含めて今後の問題であることには注意が必要である。

4.1 分析に利用したデータ

対応表の作成をはじめとして，分析には，小学校，中学校とも，平成 25 年度および平成 28 年度の全国学力・学習状況調査の本体調査，ならびに，経年変化分析調査から得られたデータを文部科学省から貸与を受け，それを利用した。具体的に分析対象とした人数の内訳は表 4.1 に示す通りであった。例えば，平成 28 年度の本体調査・小学校国語では児童 1,045,726 名分のデータを対象としたが，そのうち経年変化分析調査に含まれる児童は 10,967 名ということが分かる（表 4.1）。

表 4.1 分析対象とした人数内訳

実施年度	調査種類	教科	学年	
			小学6年生	中学3年生
平成25年度	本体調査（全数）	国語	1,130,296	1,088,997
		算数/数学	1,130,730	1,089,359
	経年変化分析調査（抽出）	国語	5,896	10,781
		算数/数学	5,881	11,605
平成28年度	本体調査（全数）	国語	1,045,726	1,042,719
		算数/数学	1,046,363	1,042,929
	経年変化分析調査（抽出）	国語	10,967	26,531
		算数/数学	10,753	25,942

表 4.2 全テストの項目数

	小学校				中学校			
	国語		算数		国語		数学	
	本体	経年	本体	経年	本体	経年	本体	経年
H25年度	28	28	32	32	41	41	52	39
H28年度	25	39	29	52	42	52	51	65
共通項目数	0	24	0	28	0	38	0	36
非共通項目数	53	19	61	28	83	17	103	32

また，表 4.1 にはすべての調査における項目数を示した。なお，本研究では総体としての学力変動をとらえることを最重要の目的としたため，いわゆる A 問題，B 問題の区別は行っていない。この問題は本質的には次章で扱う下位領域ごとに細分化された学力変動をいかにとらえるかの議論に収れんできるものである。表 4.1 もそのことを反映し基本的には調査単位ごとの項目数となっていることに注意が必要である。

例えば，小学校国語においては，本体調査において平成 25 年度は 28 項目，平成 28 年度は 25 項目が実施されていることが分かる。したがって最終的にもとめるべき対応表には平成 25 年度の正答数得点（範囲 0 から 28）を平成 28 年度の正答数得点（範囲 0 から 25）に換算できる機能が必要となる。また，本体調査における共通項目とは平成 25 年度と平成 28 年度のテスト間に共通して含まれる項目のことであるが，そのような項目は存在しないので該当項目数は 0 となる。また，非共通項目とは共通項目を除いた項目のため，本体調査においては平成 25 年度と平成 28 年度の項目数の合計となる。例えば，小学校国語においてはその項目数は 53 となり，これらが最終的に経年変化分析調査項目とともに共通尺度上に定位，表現されることとなる。

一方，同じ小学校国語であっても，経年変化分析調査では平成 25 年度は 28 項目，平成 28 年度は 39 項目が利用され，その共通項目数は 24 項目，それ以外の項目（非共通項目）が両年度併せて 19 項目となっている。経年変化分析調査内で共通項目 24 個の情報を使って IRT 等化が行われており，すでに項目母数は推定されている。以下で詳述するように，この報告書では，この項目群をいわばターゲットにして，本体調査の項目を共通受検者集団の情報を介して対応づけることになる。他の学年，年度，教科についても同様である。

4.2　データ収集デザイン

また，併せて，文部科学省から経年変化分析調査の実施対象校が特定可能な資料の提供を受け，それに基づき，本体調査データと経年変化分析調査データとの共通受検者集団を特定した。共通受検者集団と各年度，各調査のデータ構造は図 4.1 に示すとおりである。

まず平成 25 年度データに着目すると，本体調査ブロックの全数データのうちに経年変化分析調査も受けた集団が存在する。それを図 4.1 では濃いめの網掛けで表現している。実際には標本抽出をしているが，わかりやすいように図中では標本抽出された集団をまとめて表現した。この集団は平成 25 年度経年変化分析調査の 2 分冊を受けている。どの分冊を受けるかは学校ごとに無作為で振り分けているため，平成 25 年度はデータ収集デザインとしては，同一母集団から抽出された統計的には等質な二つの標本集団が別々の冊子を受ける，等価グループ・デザイン（Equivalent Groups Design）となっている。

一方，平成 28 年度は本体調査ブロック側での抽出手続きは同じであるが，経年変化分析調査ブロック側での実施方式が BIB デザイン（Balance Incomplete Block Design：BIBD：釣合型不完備ブ

ロックデザイン）に基づく重複テスト分冊方式となっているため，使用された 13 分冊それぞれに学校単位で無作為に割り当てられていることに注意が必要である。

さらに平成 25 年度と平成 28 年度の経年変化分析調査に着目すると，項目には両年度とも実施された項目が含まれている。これを共通項目とする。実際には各分冊ともそれらの項目が含まれているが，描画すると複雑になるため，わかりやすいように，図 4.1 では項目欄のみにその様子を描いている。

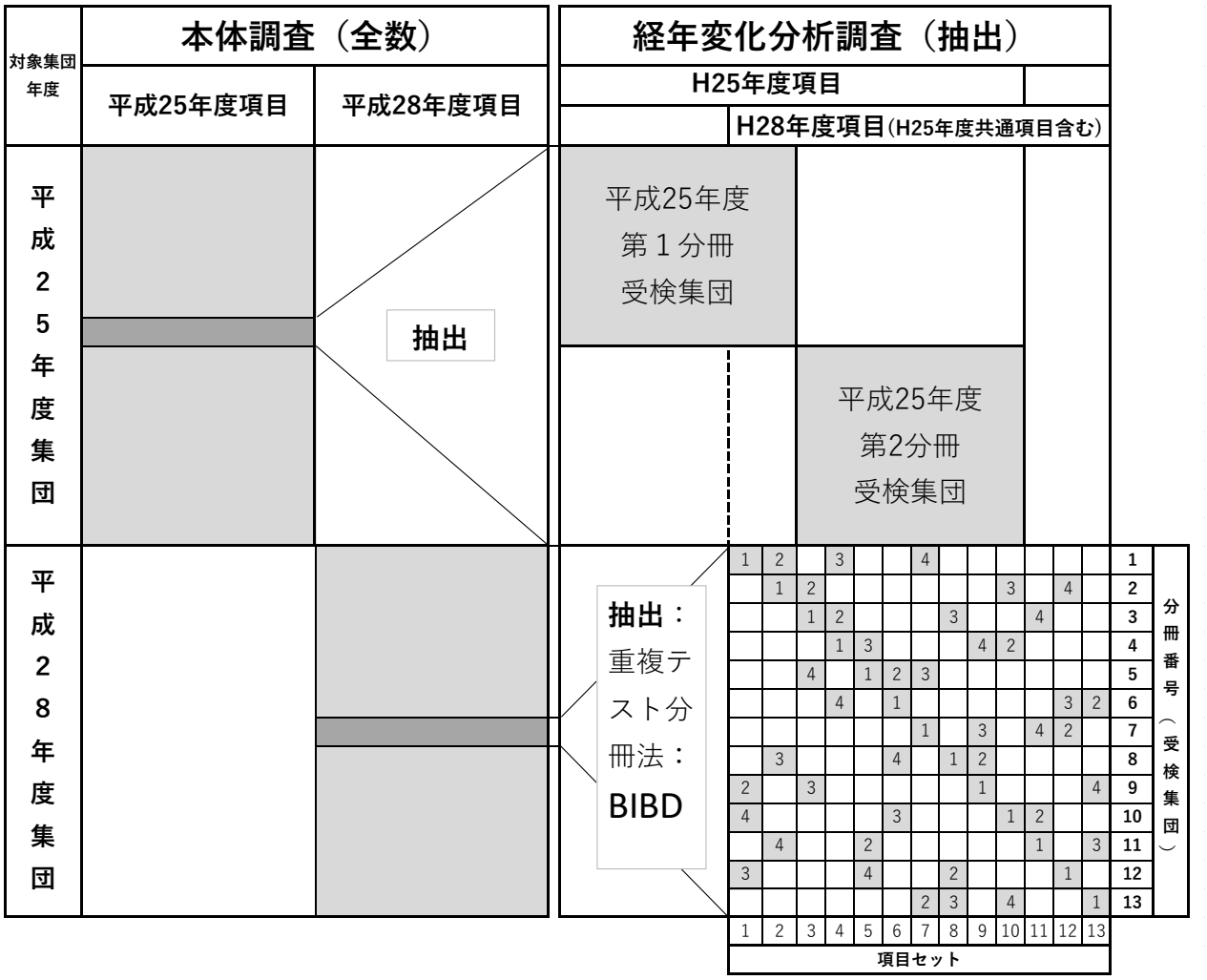

図 4.1　データ収集デザインの概略[3]

[3] BIBD の表についてグレーに塗りつぶされているマスは，それに対応する列番号の項目セットが，対応する分冊番号に含まれることを示す。なお，マス目内の番号は，その分冊内での項目セットの順番を示す。

4.3　対応表の作成手順

異なる二つのテストを等化するとき，①個別のテストごとに母数を推定し，等化係数を求める方法，②異なるテストを単一のデータ行列として扱い，一度にすべての母数を求める方法，③一方のテストの項目母数を固定して，もう一方のテストの項目母数を推定する方法，の三つが一般的に用いられる。この方法を順に，個別尺度調整（個別推定）法，同時尺度調整（同時推定）法，項目固定法（FCIP）と呼ぶ。以上は，データ行列の中に項目側か受検者側に直接含まれている共通の情報を利用することで等化を実行する方法である。そのほかにも，④一つの受検者集団をランダムに２分割し，そのランダムに分割した受検者集団を等価な集団とみなすことで等化を実行する等価グループ・デザインによる等化も存在するがここでは扱わない。

本研究では独立尺度調整法に基づく等化を実行した。等化係数の推定方法には共通項目法を Stocking-Lord の方法，共通受検者法を熊谷・野口の方法（2012）を採用した。Stocking-Lord の方法（1983）に基づく等化係数推定値には貸与を受けた推定値を使用し，熊谷・野口の方法（2012）による等化係数の推定には，Easy Estimation（Ver. 2.0.4）[4]の POP オプションを使用した。本節ではデータの読み込みから等化までの実行手順を述べる。

4.3.1　利用したデータとプログラム

利用したデータは，大きくは

・本体調査の CSV データ

・経年変化調査の CSV データ

の 2 種類であった。また，分析やプログラム開発に用いたソフトウェアは，

・R（Version 3.4.3, 64bit）

・EmEditor Professional（Version 17.4.2 ,64-bit）（データ確認用）

・熊谷（2009）Easy Estimation（Version 2.0.4）

であった。

なお，この分析過程におけるすべての項目母数の推定には，経年変化調査に参加した受検者のみのデータを用いて行った。したがって，本体調査の項目母数も表 4.3 に示す経年変化調査参加数に準じたサイズの本体調査からのサンプリング・データを用いている。項目困難度母数の推定精度に関しては，実データに基づくこれまでのシミュレーション研究（柴山他，2014）により，5000 名程度以上のサンプルサイズがあれば問題無いとの結論を得ている。

表 4.3　項目母数の推定のためのサンプリング数

	小学校	中学校
平成25年度	11,777	22,386
平成28年度	21,720	52,473

[4] なお，執筆時点では Easy Estimation（Ver. 2.0.6）がリリースされたが，このバージョンの変更点はいずれも能力パラメタの点推定値に関する基準であり，今回の分析に関わるプログラム部分の変更点はない。

4.3.2 等化実行前の準備

まず，貸与を受けた CSV データを R にて読み込み，項目反応データ以外の部分を削除する。PC のスペックにも依存するが，まずこのデータを通常の read.csv 関数などで読み込むことはデータ容量のせいで非常に困難である。そこで R の data.table パッケージ（Dowle & Srinivasan, 2017）を使用した。このパッケージは数ギガのデータファイルであっても数十秒で，破損なくデータを読み込むことを可能とするものである。

項目反応パターン以外の情報を含む部分（データ行列の列に該当）を削除した上で，反応データを Easy Estimation で推定可能な形式にコーディングした。具体的には CSV 形式のデータのカンマ「,」を取り除いたテキストデータに変換し，さらに誤答反応を 0，正答反応を 1，欠測値を N に置き換える手続きを行った。R ではファイルを書き出す際に文字列の区切り方や欠測値を任意の文字列に置き換えをおこなうことができるため，カンマを取り除き，欠測値を N に置き換える作業は容易であった。しかし誤答反応をすべて 0 に整える手続きはやや複雑であった。もともとのデータでは無回答が 0，誤答が 2，というカテゴリーに分類されており，今回は，この二つのカテゴリーを誤答として扱うこととした。R には文字列を置き換えるための関数が用意されているが，今回扱うデータは非常に大きいため置き換えに時間がかかることはおろか，最悪置き換えできない可能性もあった。したがって，Easy Estimation で読み込み可能なテキストデータとして書き出した後に，Easy Estimation のコーディング機能（Data scoring）を利用して，すべての誤答を 0 に置き換えるほうがより確実である。詳細なコーディング方法については Easy Estimation 付属のマニュアルを参照されたい。

4.3.3 項目分析および項目母数の推定

Easy Estimation で分析をするためのデータ形式の変換作業後，各学年・各教科・各年度で項目分析と IRT 項目母数の推定を行う。そのためには，合計で 8 種類のデータを別々に分析する必要がある。

まず IRT モデルによる分析を実行する前に，構成概念の一次元性の確認や，各項目の正答率，点双列相関係数などを確認した。一次元性の確認には固有値の減衰状況を視覚的に確認するスクリープロット法を用いた。ただし，H25 年度の経年変化分析データは調査 A と調査 B の間に共通情報がないため相関係数を推定することができない。したがってこのデータの一次元性の確認は調査 A と調査 B を独立に行った。先述したように，平成 25 年度はデータ収集デザインとしては，同一母集団から抽出された統計的には等質な二つの標本集団が別々の冊子を受ける，等価グループ・デザイン（Equivalent Groups Design）となっていることから，この方法は妥当である。

構成概念の一次元性の確認と古典的テスト理論に基づく項目分析を経て，一次元 IRT モデルに当てはめることに問題がないことを確認し，IRT 項目母数の推定を行った。なお，尺度定数 D は 1.702 を使用している。項目母数の推定値については基準となる経年変化分析調査の項目母数の推定値が非公開のため本報告書でもその値は公開していない。

4.3.4 共通項目法による異なる年度の経年変化調査の等化

すべての項目母数を H28 年度の経年変化調査の尺度上に等化するためには，年度ごとに本体調査と

経年変化調査を等化する前に，異なる年度の経年変化調査を等化する必要がある。まずは H28 年度と H25 年度で共通する項目を特定し，H25 の共通項目母数をすべて H28 の母数に置き換えた。その後，H25 のみの項目（H25 非共通項目）に対して等化係数を用いて線形変換を行った。具体的には，Stocking-Lord の方法によって得られている等化係数の傾きを A，切片を B とすると，

$$a^*_{H25} = \frac{a_{H25}}{A} \ , \qquad (1)$$

$$b^*_{H25} = Ab_{H25} + K \ , \qquad (2)$$

というように計算することで，未等化な H25 年度の項目母数a_{H25}, b_{H25}を，H28 年度の共通尺度上における H25 年度項目母数a^*_{H25}, b^*_{H25}に変換した。H28 のみの項目（H28 非共通項目）は何も変換する必要はない。

4.3.5　共通受検者法による本体調査と経年変化調査の等化（対応づけ）

この手順では項目母数および項目反応パターンを用いて母集団能力分布の平均と標準偏差を推定し，その値を等化係数の推定値とする手法である。なおここでは具体的な方法に注目して等化という用語を引きつづき使用するが，本体調査と経年変化調査とは互いに異なる設計仕様をもっているため，分類概念からいえば，厳密には等化（equating）ではなく，等化よりも条件の緩い対応づけ（linking）とよぶべき手続きである。

具体的な計算手順を説明する前に，まずは熊谷・野口（2012）もとに，この手法の説明をおこなう。IRT の母数には尺度の単位と原点に不定性があるため，一定のルールに基づいて変換することが可能である。例えば A を傾き，K を切片とおくと，

$$\theta^* = A\theta + \mathrm{K} \ , \qquad (3)$$

$$a^* = \frac{a}{A} \ , \qquad (4)$$

$$b^* = Ab + K \ , \qquad (5)$$

のように線形変換をすることができる。θ^*とθが異なるテスト項目この傾きと切片のことを等化係数と呼ぶ。この等化係数の推定方法には先に述べた Stocking-Lord の方法以外にも，いくつかの手法が存在する。そのなかでも Marco（1977）の Mean&Sigma 法は共通受検者の能力値の平均と標準偏差を用いて，

$$A = \frac{\sigma_{\theta T}}{\sigma_{\theta F}} \ , \qquad (6)$$

$$K = \mu_{\theta T} - A\mu_{\theta F} \ , \qquad (7)$$

と係数を求める手法であり，計算が非常に簡便であるという特徴を持っている。なお，式中の T と F は，二つの集団で別々に収集されたデータに基づいて構成された尺度 T と F を意味している。T は等化先の尺度を，F は等化元の尺度を表しており，この表記は加藤ら（2014）に準拠したものである。本研究では等化先 T が経年変化分析調査であり，等化元 F が本体調査である。

今回は熊谷・野口（2012）の方法を採用した。これは，平均と標準偏差を計算する際に，直接受検者集団の母集団分布を推定し，等化係数の推定時の誤差分散を小さくする方法である。Easy Estimation を用いれば非常に容易に推定することができる上に，従来の Mean & Sigma 法よりも安定した等化係数の推定が行うことができる。

続いて，具体的な計算手順を説明に移る。まずは H25 年度の等化について考える。すでに経年変化調査の項目母数，すなわち等化先の尺度上の項目母数は H28 年度の尺度上に等化済みである。この等化済み項目母数と項目反応データを用いて等化先の推定母集団平均$\mu_{\theta T}$と標準偏差$\sigma_{\theta T}$を計算する。計算には Easy Estimation の「Estimation examinee parameters」のオプションのひとつである「POP」使用し，初期値である一様分布の範囲は－4 から 4 を指定し，求積点の数は 31 を指定した。

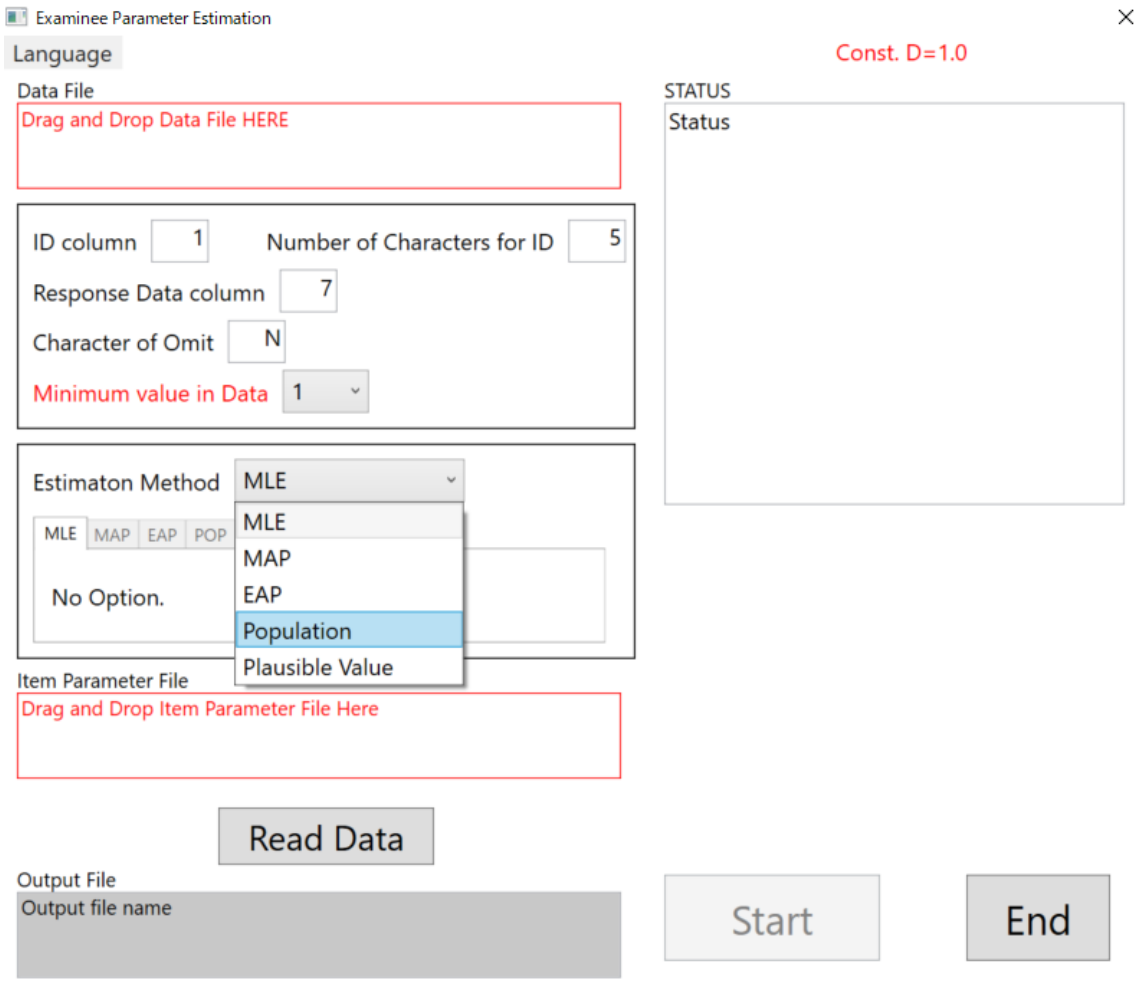

図 4.2　EasyEstimation による母集団分布推定時のコンソール画面

出力された CSV ファイル「項目反応データのファイル名＋ThetaPOP.csv」の 3 行目に平均が，4 行目に標準偏差の推定結果が書き込まれている（Figure2）。それ以降の theta と prob は受検者集団の能

力値とそれに対応する確率値であるが，直接的に等化係数の推定には関係しない。同様に H25 年度の本体調査項目の母数と項目反応パターンを用いて，等化元の推定母集団平均$\mu_{\theta F}$と標準偏差$\sigma_{\theta F}$も計算し，(6)（7）の式に代入すれば H25 年度の本体調査と経年変化調査間の等化係数の推定値が求まる。

推定された等化係数を用いて（4）(5）式の計算をおこなえば，本体調査項目を経年変化調査項目の尺度上に等化できる。すなわち，

$$a_F^* = \frac{a_F}{A} \ , \qquad (8)$$

$$b_F^* = Ab_F + K \ , \qquad (9)$$

を求めた。このとき等化先の H25 年度経年変化調査項目は，すでに H28 年度経年変化調査項目の尺度上に等化済みであるため，式（8）(9）の変換を経たa_F^*, b_F^*も，H28 年度経年変化調査項目と同一尺度上に等化されたことになる。H28 年度の本体調査と経年変化調査についてもこれまでに説明したことと同じ操作を行った。最終的に四つの独立した学力テストのそれぞれの項目群が，一つの共通尺度上に等化されたことになる。

図 4.3 「POP」のアウトプットファイル

4.3.5 IRT 母数による復元得点分布の生成

これまでの手続きで，異なる四つのテストの全項目を一つの共通尺度上に位置づけることができた。次はこの母数を用いて復元得点分布を求めた。今回得たい復元得点分布は「H25 年度の受検者」が「H28 年度のテスト項目」を受検したと仮定した場合の正答数得点の予測分布である。そのためには先に推定した H28 年度の尺度上の本体調査項目と H28 年度の尺度上の H25 受検者の能力母数が必要となる。

まず Easy Estimation の「Estimation examinee parameters」の「EAP」オプションを用いて受検者の能力値の EAP 推定値を得る。このとき MLE 推定値や MAP 推定値を採用してもよいが，MLE は尤度関数が発散してしまう場合に推定値を得られない可能性があり，MAP 推定値は反復計算の際に初期値によっては適切に推定できない可能性があるため，区分求積法により確実に推定値を計算できる EAP 推定法を採用した。推定のための項目母数は等化済みの H25 年度本体調査項目であり，項目反応パターンは本体調査受検者全数を使った。

EAP 推定値が求めた後，R にて読み込みを行った。そして等化済みの H25 年度本体調査項目母数と H28 本体調査項目母数も読み込み，あらかじめ作成しておいた復元得点分布生成用の関数 score.dist と dist.data 関数により復元得点分布を求めた。計算時間は，PC の機能にも依存するが，100 万人程度の受検者で項目数が 30 程度であれば２，３分程度かかった。なお，score.dist 関数は最終的に復元得点分布の累積度数分布のカーブを出力し，dist.data 関数は得点とその復元度数を戻り値として与える。どちらの関数も引数に能力母数，項目母数が必要である。

4.3.6　対応表の計算

復元得点分布で求めたものは「H25 年度の受検者」が「H28 年度のテスト項目」を受検した場合の正答数得点であった。この得点分布（復元得点分布）を目標分布として H25 年度の素点の度数分布（実データ分布）を等パーセンタイル等化することで目的とする対応表が得られる。等パーセンタイル等化のための関数は epe 関数である。この関数は引数にテスト X とテスト Y という，二つの異なるテストの素点を必要とし，戻り値としてテスト X のパーセンタイルランク，テスト X の素点，その素点と等パーセンタイルランクのテスト Y への変換得点（２種類の定義式によるそれぞれの得点）を与える。なお，引数の型はベクトルである。したがって，H25 年度のテスト得点を H28 年度のテスト得点へと等パーセンタイル等化したい場合には X に H25 年度の素得点を，Y には「H25 年度の受検者」が「H28 年度のテスト項目」を受検した場合の正答数得点を与えてやればよい。なお，復元得点分布の度数分布は予測分布であるため実際には小数点以下の情報が含まれているが，復元に使ったデータ数が約 100 万名と大規模なため整数値に四捨五入して得られた正答数得点の精度に問題はないと判断した。

4.4　年度間比較の実施

4.4.1　テストの信頼性と対応づけ可能性の検討

ここでは具体的に小学校国語の場合を例にとって，経年変化分析調査を介した本体調査での年度間比較の実際を述べる。まず，そもそもテスト自体が児童生徒の学力を精度よく測定できているのかの吟味（信頼性分析：reliability analysis）と，互いに異なる設計の基で作成された仕様の異なるテストの得点を比較可能なように対応づけることができるか否かの確認（対応づけ可能性分析：linkability analysis）の必要がある。後者については Sato & Shibayama（投稿中）によって提案された，対応づけられた得点の信頼性という観点から判断するための指標を利用する。

まず，表 4.4 に示すように本体調査の信頼性係数の推定値は 0.855，項目数がその半分の 14 項目から構成される経年変化分析調査の分冊であっても 0.758 及び 0.788 と，いずれの調査においても十分な測定精度が担保されていることがわかる。テスト開発の観点からいえば，経験的一般的に精度の高さは小学校・国語が一番確保しにくく，その次が小学校算数，中学校国語，一番精度が高くなるのが中学校数学である。実際，資料 4-１でもその傾向がみられ，本体調査の中学校数学では 0.938 と極めて高い信頼性が確保されていることがわかる。また，公的な大規模テストに必要な測定精度として 0.7 程度以上というのが一つの目安に使われることがあるが，そのような中でも小学校国語で 0.855 という値は十分に高い精度であったと指摘できる。

また，本体調査全体と経年変化分析調査との相関係数も 0.75 であり，設計仕様・利用目的が異なるものの，その本質においては両者ともほぼ内容的に同じ小学校国語の学力を測定しているテストであることが分かる。また，全体と A 問題との相関が 0.943，B 問題との相関が 0.877 と高くなっているが，これは本体調査全体が本体調査 A 問題と B 問題の得点を合計しているため当然の結果である。しかし，A 問題と B 問題の相関が 0.668 であるのは，いずれの問題も同じ国語の学力を測定しつつも，少し異なる側面からその学力を測定しているともいえ，A 問題と B 問題の設計目的が数値としてあらわれたものと解釈して良いであろう。小学校算数，中学校国語，中学校数学になると A 問題，B 問題間の相関は 0.7 から 0.8 と小学校国語の場合に比べて高くなり（資料 4-1），設計目的通り A 問題，B 問題では互いに少し異なる学力を測定しつつも，両問題で問うている学力の重なりが大きくなっている，あるいは相互に影響し合った学力の構造を持っていると解釈できる。この議論は，平成 28 年度調査においても同様である（資料 4-２,3）。

表 4.4　平成 25 年度本体調査・経年変化分析調査の項目数・信頼性・相関係数（小学校・国語）

小学校・国語	項目数	信頼性係数（α）	本体調査全体	本体調査A問題	本体調査B問題
本体調査・全体	28	0.855	1.000	0.943	0.877
本体調査・A問題	18	0.783	0.943	1.000	0.668
本体調査・B問題	10	0.735	0.877	0.668	1.000
経年調査・分冊１	14	0.758	0.750	0.703	0.665
経年調査・分冊２	14	0.788	0.754	0.695	0.680

次に Sato 他の指標を使った対応づけ可能性の検討であるが，その考え方の基本は，仮に経年変化分析調査側のテスト得点を本体調査側のテスト得点に直接的に線形等化すると仮定し，対応づけられた経年変化分析調査側の得点（対応づけ得点）の信頼性がどの程度の範囲に入るかを両者の信頼性係数ならびに相関係数から推定するものである。シミュレーション等でおよそその値は 0.7 程度あれば良いことが分かっている。これを実際に示したのが表 4.5 である。

表 4.5　平成 25 年度調査において経年変化分析調査を本体調査に対応づけた場合の信頼性の範囲

小学校・国語	共　通　尺　度	
	本体調査・全体	
	下界	上界
経年調査・分冊 1	0.667	0.707
経年調査・分冊 2	0.670	0.712

小学校国語ではその下限値が 0.7 を切っているもののその値は約 0.67 であること，また資料 4-4 から資料 4-6 までに示す値も下限値であってもほぼ 0.7 程度は確保できていること，それに対して上限値は 0.7 を大きく超え，中学校数学では 0.8 までに及んでいること等から，対応づけは十分に可能であると考えられる。

次に実データ分布と復元得点分布の比較による復元精度の確認であるが，これは図 4.4 に示すとおりである。黒い実線が平成 28 年度中学校数学の累積曲線，青い累積曲線が復元したものであるが，ほとんど重なっており，十分な復元精度であることが分かる。

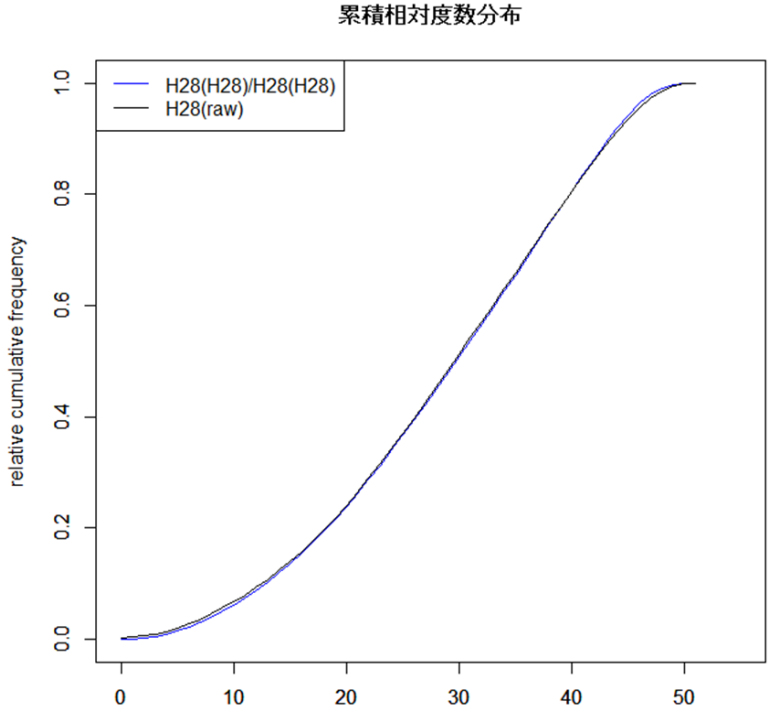

図 4.4　実分布と復元分布の累積相対曲線の比較（中学校・数学）

以上の検討より，本提案方法における対応づけ手続きは全体として妥当なものであると判断した。

4.4.2　対応づけの実際

表 4.2 に示すとおり，平成 25 年度の小学校国語は A 問題，B 問題を合わせると全部で 28 項目からなるため，正答数得点の範囲は 0 点から 28 点となる。受検した児童の総数 1,130,296 名であった。その得点分布の様子は表 4.5 に示すとおりである。

表 4.5　平成 25 年度本体調査小学校国語の正答数得点に関する度数分布表

小学校　国語				
正答数	度数	累積度数	相対度数(%)	累積相対度数(%)
0	1699	1699	0.150	0.150
1	2502	4201	0.221	0.372
2	4570	8771	0.404	0.776
3	7726	16497	0.684	1.460
4	11456	27953	1.014	2.473
5	16066	44019	1.421	3.894
6	21066	65085	1.864	5.758
7	26004	91089	2.301	8.059
8	31151	122240	2.756	10.815
9	36603	158843	3.238	14.053
10	41844	200687	3.702	17.755
11	47285	247972	4.183	21.939
12	52715	300687	4.664	26.603
13	57085	357772	5.050	31.653
14	61471	419243	5.438	37.091
15	65334	484577	5.780	42.872
16	67904	552481	6.008	48.879
17	69920	622401	6.186	55.065
18	70902	693303	6.273	61.338
19	71482	764785	6.324	67.662
20	69907	834692	6.185	73.847
21	66156	900848	5.853	79.700
22	60700	961548	5.370	85.070
23	53874	1015422	4.766	89.837
24	44371	1059793	3.926	93.762
25	33539	1093332	2.967	96.730
26	22374	1115706	1.979	98.709
27	11216	1126922	0.992	99.701
28	3374	1130296	0.299	100.000
合計	1130296		100.000	

一方，表 4.2 に示すとおり，平成 28 年度では出題された問題は 25 項目に減っているため正答数得点の範囲は 0 点から 25 点となる。平成 25 年度との共通の問題はない。また，平成 28 年度児童集団は平成 25 年度児童集団ではない。しかし，経年変化分析調査のデータを経由して，平成 25 年度本体調査の 28 個の項目と平成 28 年度の本体調査の 25 項目の計 53 個の項目は，IRT 等化によって困難度等が調整され同じ尺度（共通尺度）上で表現されている。

そこで，まず，平成 25 年度児童集団全員 1,130,296 名の尺度値（尺度得点，学力）を平成 25 年度の項目への児童ごとの反応パターン（正誤）を使って IRT 推定によって求める。求められた尺度値ごとに，第 3 章で説明した Lord & Wingersky（1984）の Recursion Formula（再帰式）を用いて，平成 28 年度のすべての項目に対するすべて反応パターンについての確率を計算し，これを項目ごとにすべて足しあわせる。この段階で，ある尺度に関する平成 28 年度得点に関する確率分布が得られたことになる。この操作を 1,130,296 名分の尺度値すべてについて行い，その合計をとる。これが表 4.6 に示すところの H25 年度集団が平成 28 年度全体調査を受検したと仮定した場合の復元得点分布になる。復元度数は IRT モデルに基づく理論的な予測値のため小数点第 3 位まで表示した。合計人数を求めるとその値は 1130295.990 となり，丸め誤差の範囲で実人数の 1,130,296 と一致している。これは，いわば，IRT モデル上で平成 25 年度児童生徒集団に，平成 28 年度全体調査を受検してもらった状況を仮想的に作り出したことになる。算数・数学や中学校においても同様の計算を行い，その結果は資料 3 に示すとおりである。

表 4.6　平成 25 年度集団が平成 28 年度本体調査を受検したと場合の復元度数分布（小学校国語）

正答数	復元度数
0	54.740
1	406.620
2	1360.700
3	3039.640
4	5458.710
5	8627.880
6	12503.410
7	16975.350
8	21935.340
9	27333.500
10	33182.710
11	39527.830
12	46406.600
13	53817.010
14	61691.990
15	69875.540

正答数	復元度数
16	78092.790
17	85902.210
18	92616.620
19	97191.750
20	98115.080
21	93388.160
22	80861.450
23	59565.530
24	32674.150
25	9690.680

合計	1130295.990

※予測値のため小数点以下第3位まで表示

次に上で得られた平成 28 年度復元得点分布をターゲットに平成 25 年度正答数得点を平成 28 年度得点に関する手続きを示したのが図 4.5 である。左右両方の図とも累積相対度数のグラフ（累積カーブとよぶ）が描かれている。縦軸はいずれもパーセントランクを表す。いま，平成 25 年度正答数得点 17 に着目し，そこから垂直線を伸ばし累積カーブと交わる点を見つける。その交点から左に水平

に移動しパーセンタイルランクを読み取ると 51.972 となる。今度は逆に右方向水平に直線を伸ばしていくと平成 28 年度復元得点分布の累積カーブと交わる点がある。この点からまっすぐ下に垂線を下ろすと垂線の足が 17.729 となることがわかる。この値が平成 25 年度得点を平成 28 年度得点に換算した値である。もちろん以上のことは目の子ではなく，第 1 章で開発したパーセンタイル等化のアルゴリズムを用いて数値的に求めることになる。その結果が表 4.5 となる。なお，換算された値は平成 28 年度の得点の範囲の 0 から 25 までに収まっているわけではなく，平成 25 年度正答数得点の 0 と 28 は，平成 25 年度換算では，それぞれ 1.786，25.326 に対応づけられていることに注意が必要である。このような現象はとくに異常なわけではなく，単にパーセンタイルランクと得点の範囲の組み合わせで起こっていることである。以上をまとめた結果が表 4.7 になり，第 1 列と第 2 列が求める対応表である。

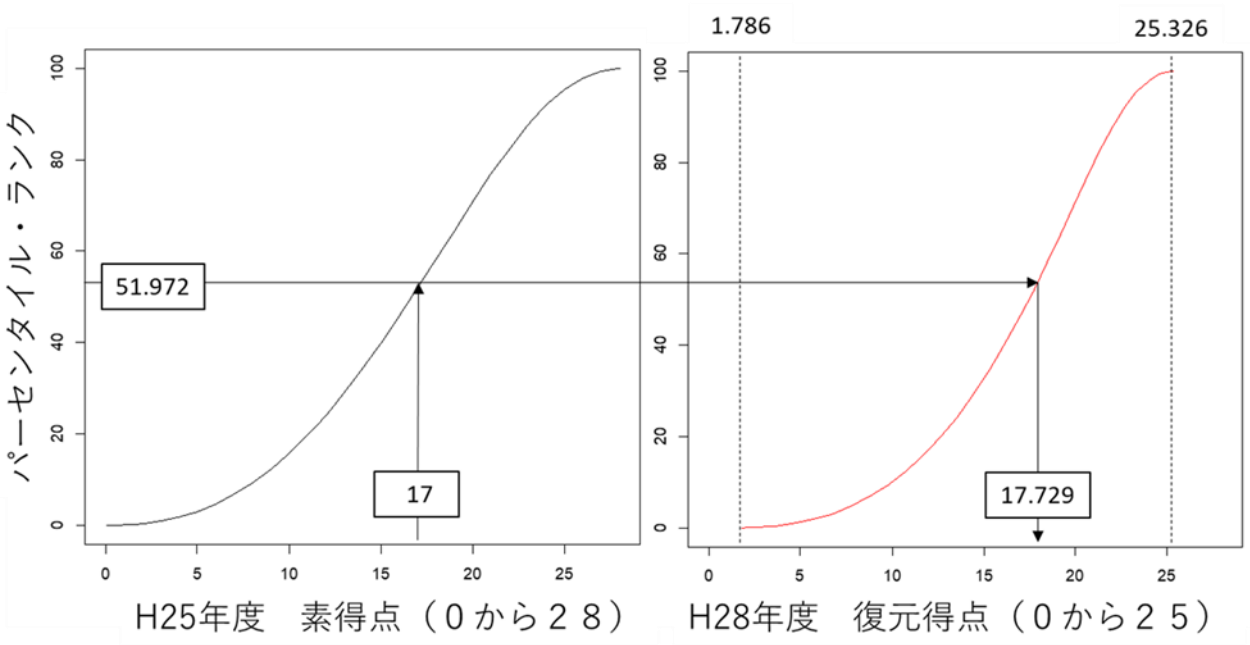

図 4.5　平成 25 年度正答数得点と平成 28 年度復元得点とのパーセンタイル等化の原理

表 4.7　対応表（換算表）とパーセンタイルランクとの関係（小学校国語）

小学校　国語				
H25年度 素得点	H28年度 換算点	累積度数	累積相対 度数(%)	パーセンタイル 順位
0	1.786	1699	0.150	0.075
1	2.872	4201	0.372	0.261
2	3.798	8771	0.776	0.574
3	4.769	16497	1.460	1.118
4	5.762	27953	2.473	1.966
5	6.767	44019	3.894	3.184
6	7.779	65085	5.758	4.826
7	8.783	91089	8.059	6.909
8	9.770	122240	10.815	9.437
9	10.745	158843	14.053	12.434
10	11.702	200687	17.755	15.904
11	12.640	247972	21.939	19.847
12	13.560	300687	26.603	24.271
13	14.450	357772	31.653	29.128
14	15.304	419243	37.091	34.372
15	16.137	484577	42.872	39.982
16	16.945	552481	48.879	45.876
17	17.729	622401	55.065	51.972
18	18.490	693303	61.338	58.202
19	19.223	764785	67.662	64.500
20	19.946	834692	73.847	70.755
21	20.646	900848	79.700	76.774
22	21.325	961548	85.070	82.385
23	22.007	1015422	89.837	87.454
24	22.655	1059793	93.762	91.800
25	23.309	1093332	96.730	95.246
26	24.008	1115706	98.709	97.719
27	24.573	1126922	99.701	99.205
28	25.326	1130296	100.000	99.851

以上の手続きによって求められた対応表を使って，平成 25 年度正答数得点を平成 28 年度復元得点分布に変換すれば，平成 25 年度児童生徒集団が平成 28 年度本体調査を受検したと仮定した場合の得点分布を求めることができる。この得点分布と平成 28 年度に得られた実際の正答数得点分布とを比較すれば，平成 28 年度得点上で平成 25 年度からの得点分布（学力分布）の変化の様子をとらえることが可能となる。

この様子を全国のすべての児童集団で見たのが図 4.6 である。H25 年度集団が H28 年度調査を受けたと仮定した場合の復元得点分布と H28 年度実データ分布の比較が累積相対曲線で描かれている。全国の年度間比較でみれば小学校・国語は若干高学力層が増えた分，低学力層も若干増えている様子もうかがえるが，全国としてはさほど大きな変化はみられない。

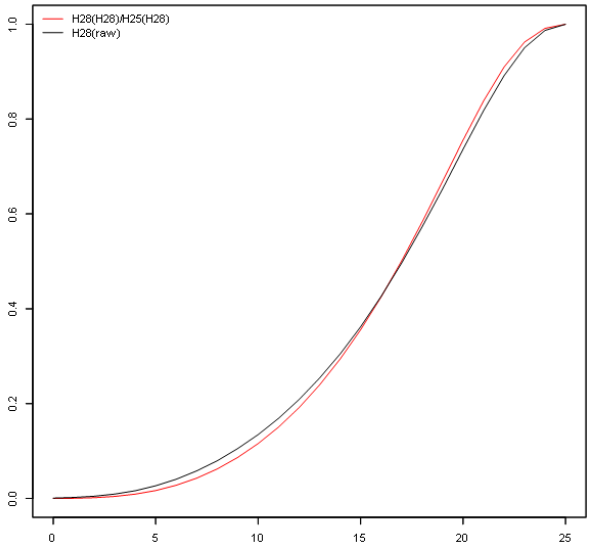

図 4.6　本体調査の得点分布の年度間比較（小学校国語：全数）

これを県別にみると，例えば図 4.7 に示すように，匿名化第 4 都道府県の場合は全国とほぼ同じ様子が読み取れる。

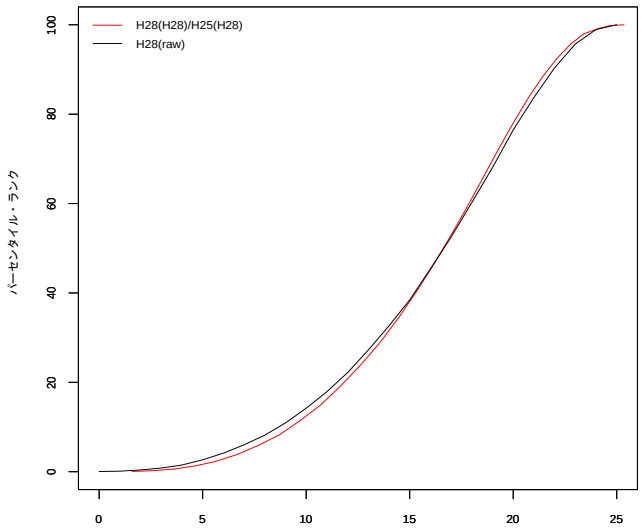

図 4.7　本体調査の得点分布の年度間比較（小学校国語：匿名化第 4 都道府県）

しかしその一方で匿名化第 25 都道府県では，赤色で示した平成 25 年度分布が，黒で示した平成 28 年度分布に異動していることが分かる。このことは平成 28 年度集団の方が平成 25 年度集団に比べて全体に学力が伸びていることを示すものである。おそらくこの都道府県ではこの間，なんらかの効果的な取り組みの大きな努力があったのであろう。このように，他の都道府県との年度内相対比較ではなく，自らの都道府県内の過年度分布との比較による振り返りが可能となることがこの方法のアドバンテージである。

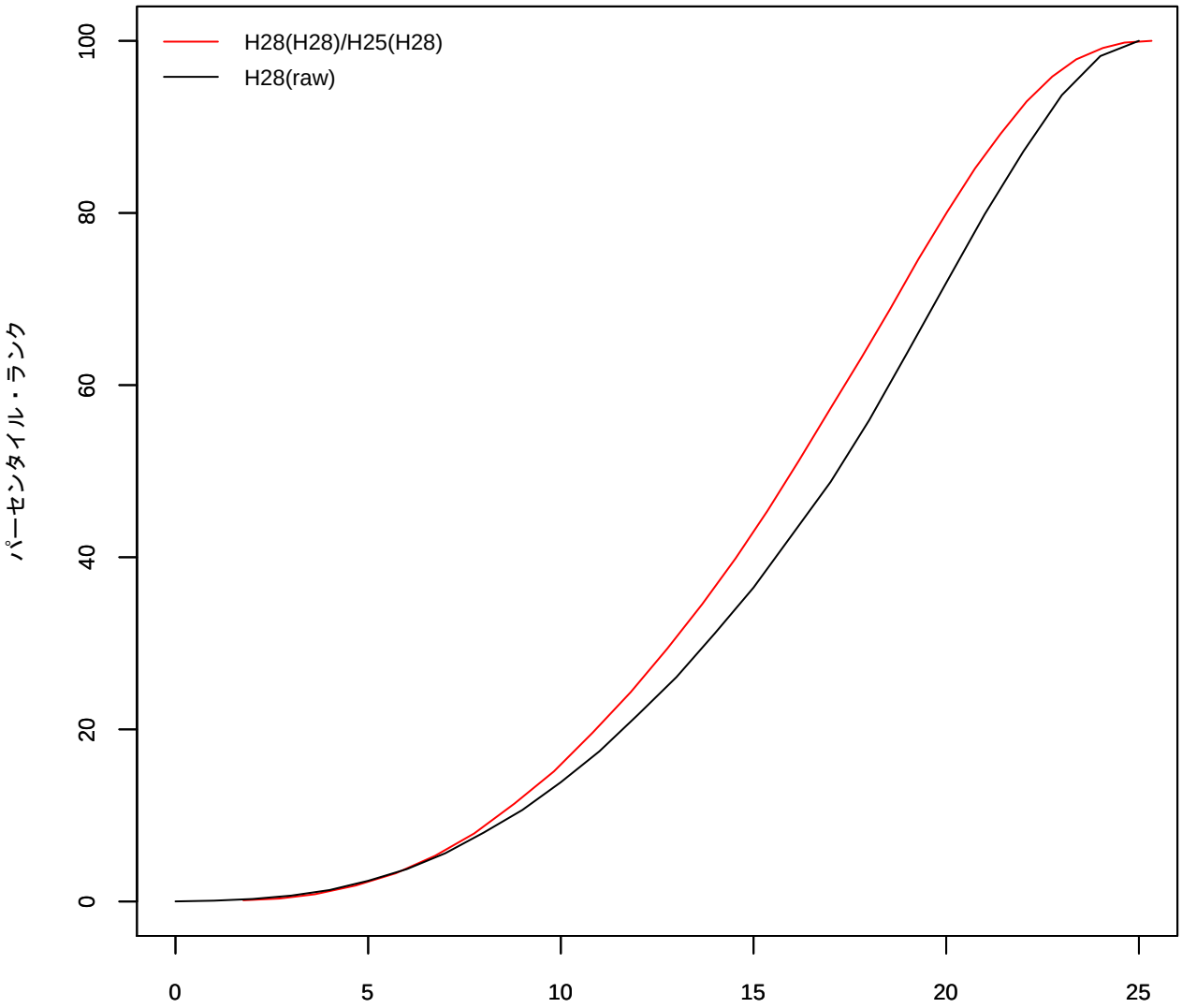

図 4.8　本体調査の得点分布の年度間比較（小学校国語：匿名化第 25 都道府県）

5. 推算値をもちいた下位領域ごとの得点比較の試み

5.1 下位領域ごとの年度間比較の難しさ

全国学力学習状況調査のテスト項目はそれぞれ一つもしくは複数の下位領域に対応している。その下位領域ごとにテスト得点の経年変化をみる場合の注意点は推算値に関する章ですでに述べたとおりである。端的にいえば，下位領域に含まれる項目数が少ないため年度間比較に必要な精度がテスト全体で比較するよりも著しく低下することがその原因である。

ここではそれを具体的に示すために，素点，点推定値，推算値の順にヒストグラムを例示し，それぞれの推定値が与える分布の特徴を考察する。なお，本章で扱うヒストグラムと推定密度関数（後述）の縦軸はすべて相対度数（密度）をとっていることに注意されたい。

例えば，小学校算数のテストには数量関係について問われるテスト項目が存在する。数量関係のみを問うている項目数は，平成 25 年度（以下，H25）で 32 項目中 7 項目が，また，平成 28 年度（以下，H28）では 29 項目中 5 項目が存在する。まずは素点の分布を用いて年度間の学力分布を比較する。分布の算出に利用した項目反応データは，どちらの年度とも経年変化調査に参加した受検者のうち，本体調査も受検している集団の項目反応データであり，項目母数は H28 年度の経年変化調査の項目母数を基準に等化してある。素点の分布では年度ごとに下位領域に対応する項目数が異なるため，これでは，たとえ復元得点分布を用いたとしても年度間の学力比較に支障をきたす。

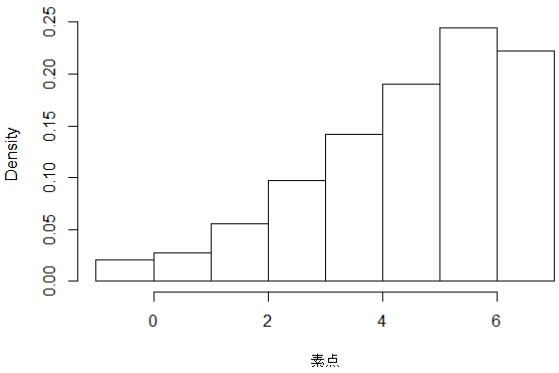

Figure 1　素点度数分布（H25）

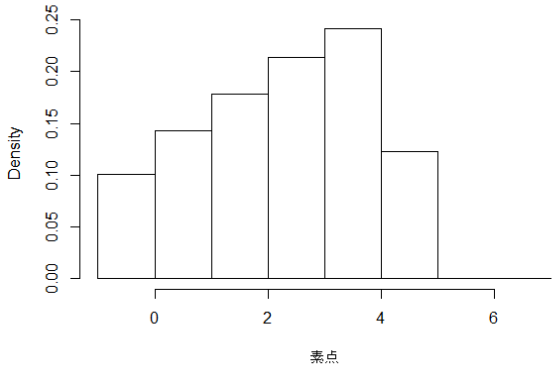

Figure 2　素点度数分布

つぎに，各テストから当該項目の項目反応と IRT 等化済みの項目母数を抜き出し，EAP，MAP，MLE 推定値を推定し，その分布を比較する。これら推定値の基本統計量は表 1 の通りである。なおベイズ推定値の事前分布には標準正規分布を用いた。

推定値	最小値	最大値	中央値	平均	標準偏差
EAP（H25）	-2.165	1.087	0.015	0.060	0.805
EAP（H28）	-1.489	1.226	0.035	0.002	0.790
MAP（H25）	-2.088	0.980	-0.002	0.012	0.758
MAP（H28）	-1.402	1.126	0.006	-0.020	0.734
MLE（H25）	-3.013	0.930	-0.114	-0.355	0.953
MLE（H28）	-1.548	1.071	0.009	-0.053	0.847

表 1　IRT 能力母数推定値の基本統計量

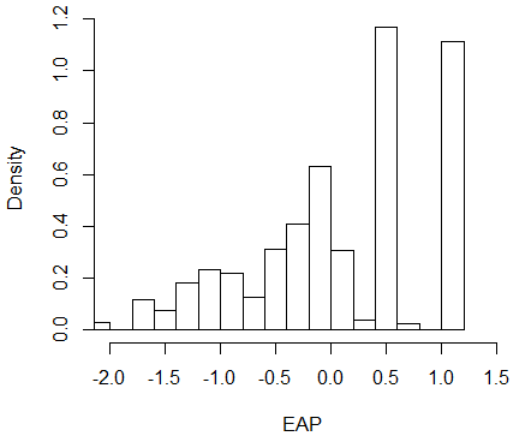

Figure 3　EAP 度数分布（H25）

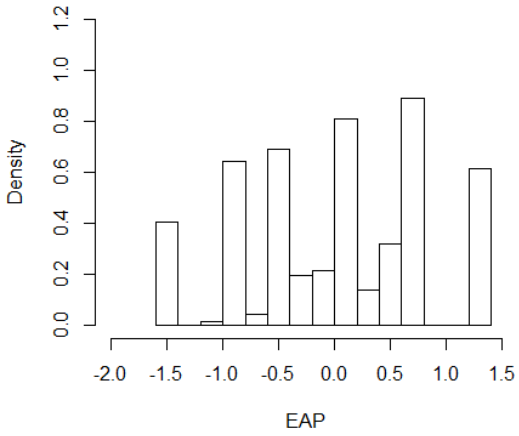

Figure 4　EAP 度数分布（H28）

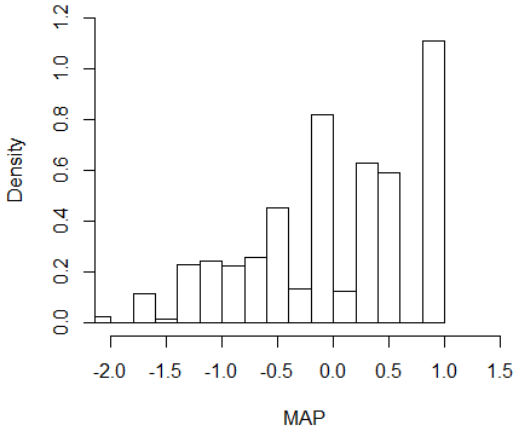

Figure5　MAP 度数分布（H25）

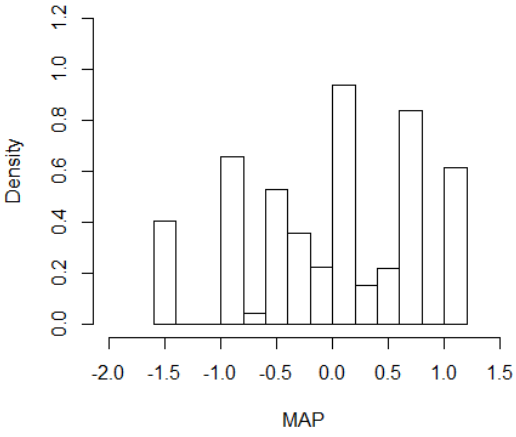

Figure 6　MAP 度数分布（H28）

ヒストグラムに描画すると，EAP，MAP，MLE 推定値はどれも起伏の大きい分布を示しており，統計量をから判断すると平均や標準偏差に違いが生じているのにも関わらず，分布の方ではその違いを確認しにくいものになっていることが分かる。これは IRT の能力母数が項目反応パターンに依存す

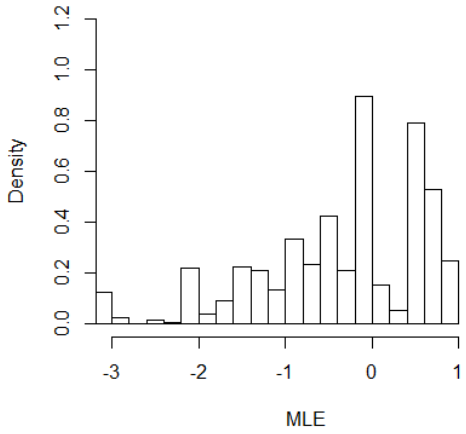

Figure 7 MLE 度数分布（H25）

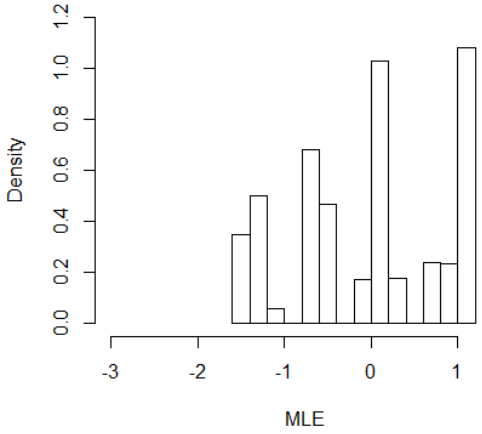

Figure 8　MLE 度数分布（H28）

るため，項目数が少ないと反応パターンが限定され，連続量である IRT の θ が離散量のように飛び飛びの値を取ることに起因する。このままでは，IRT 等化により尺度の単位と原点を揃えることができても，十分な確度をもって学力分布を比較することはできない。

5.2 推算値利用の可能性

推算値によるヒストグラムを描画する。推算値は事後分布のからのランダムサンプリングした値である。すなわち，サンプル数が十分なものであれば近似的に母集団の事後分布を表現することに他ならないだろう。推算値 10 組のデータを用いて描画したヒストグラムが以下の図である。

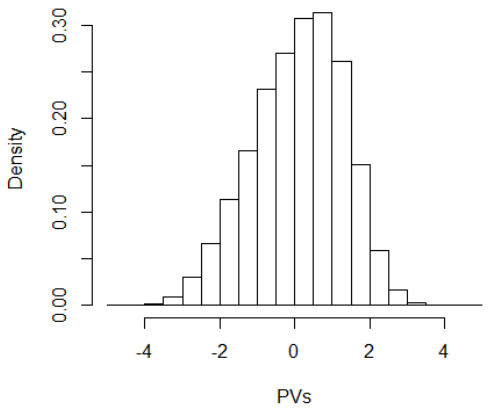

Figure 9　推算値ヒストグラム（H25）

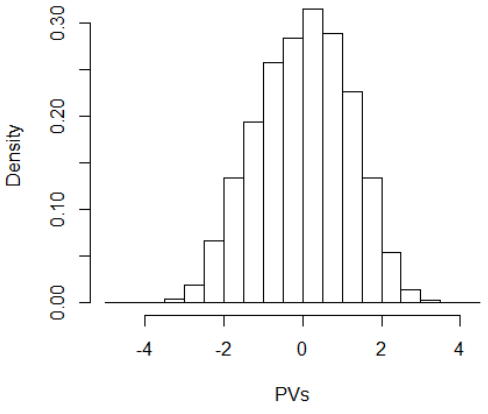

Figure10　推算値ヒストグラム（H28）

個人推定値を用いたヒストグラムよりもなめらかな分布になっていることが分かる。これは個人の能力のランダムネスを考慮した結果であり，推算値を利用するメリットの一つである。特に，ベイズ推定値に関しては事前分布に標準正規分布を用いたため平均から大きく離れた値は推定されにくい。推算値では事前分布の母数の取り方にもよるが，平均から離れた値でも推定されうる。

比較のために重ね合わせた図は以下の通りとなる。

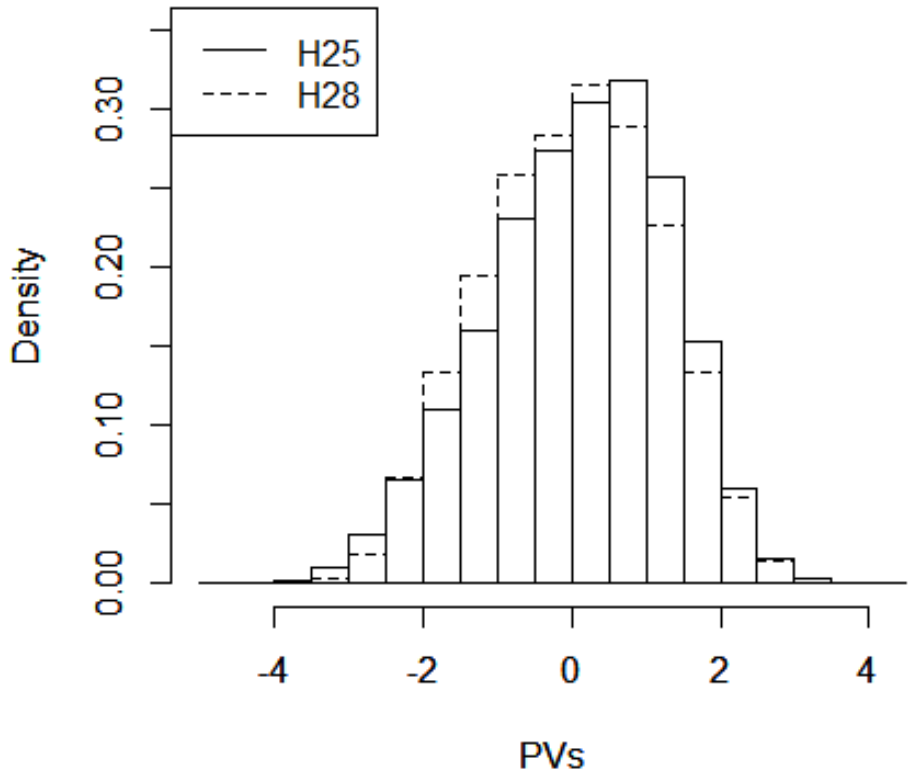

Figure 11　推算値による得点分布の比較

これまでの点推定値によるヒストグラムでは分布の形状を比較することは困難であったが，この図からは H28 は H25 に比べ，分布全体がわずかに左にゆがんでいる様子が分かる。統計量を見ると標準偏差に関してはほか推定値とは大きく異なり，H25 で SD=1.211，H28 で SD=1.172 であった。これは推算値の分散が大きいのではなく，点推定値では項目反応パターンの減少により分散が低く推定されてしまったことが原因であると考えられる。推算値を用いれば，もちろん事後分布には依存するが，項目反応パターンには依存しないため，推定に用いられる項目が少数であっても比較的正確な分散を推定できる（Wu, 2004）。なお，これら推定値の特徴の比較については Wu （2004） のほか，推算値とそのほか点推定値の統計量をシミュレーション分析により比較した von Davier et al. （2009） を参照されたい。

今回，推算値を用いて学力分布を描画するにあたり推算値 10 組の値をすべてもちいた。すなわち単純に 10 倍の θ の値を用いて相対度数を描画したことになるため，推算値のこの使用方法は一見問題があるようにも考えられる。しかし，推算値の本来の使用方法は組ごとに算出した平均や分散などの統計量の平均をとるというものである。今回の推算値によるヒストグラムの描画も，相対度数が平均や分散といった統計量を視覚的に表現している点を考慮すれば，無理のない使用方法であるといえるだろう。厳密な議論をすれば，推算値の組ごとに求めた分散の平均は，推算値 10 組すべての値の分散とは一致しない。しかし十分なサンプルサイズがあれば 10 組程度の推算値であってもその分散の値は，大きく変動しない。例えば先ほどの Figure 9 の H25 の推算値 10 組を，組ごとにヒストグラムを描画した場合，以下のようになる。どの組も分布の形状に大差はなく，したがって Figure9 のようなヒストグラムも，集団の学力分布をあらわす指標として活用可能であると言えよう。

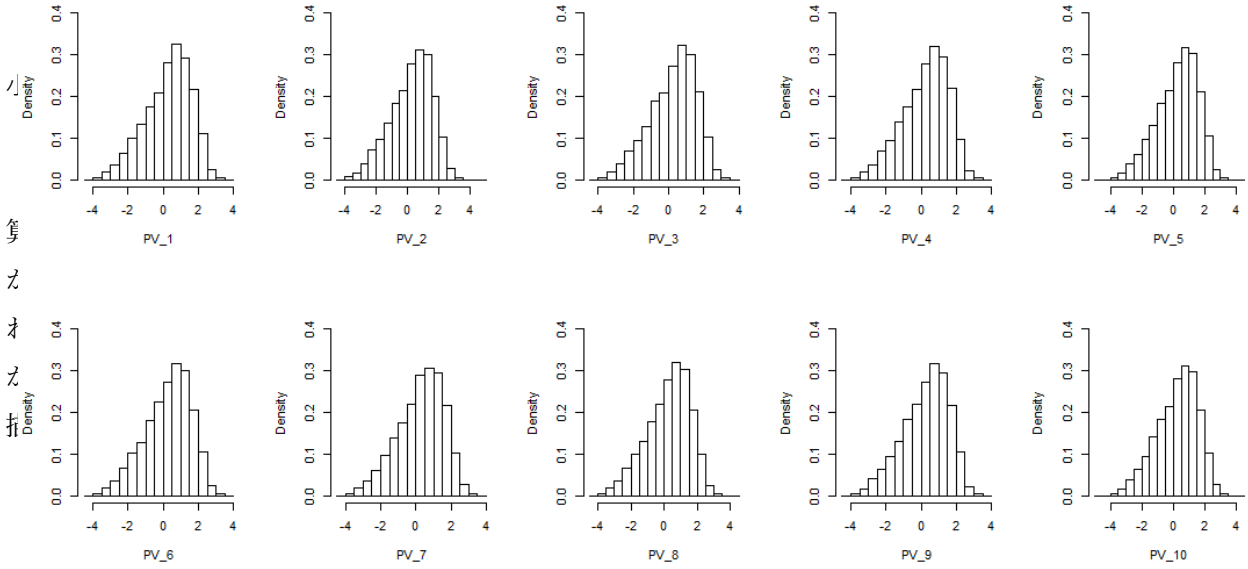

Figure 12　組ごとに求めた推算値のヒストグラム

5.3 推算値を利用した下位領域ごとの年度間比較の適用例

前節までは，点推定値の離散具合を確認するためにヒストグラムを使用していたので，それに併せて推算値のグラフにもヒストグラムを用いていた。しかし IRT 能力母数は本来連続量であり曲線で表現する方が望ましいため，ここではヒストグラムではなく密度関数で表現する。ヒストグラムと密度関数を重ね合わせたものが Figure13 である。このようにヒストグラムは指定した階級幅ごとの度数しか比較できないが，推定密度関数であれば，たとえ離散量であってもその間を数値的に補完して，なめらかな曲線として描くことができる。

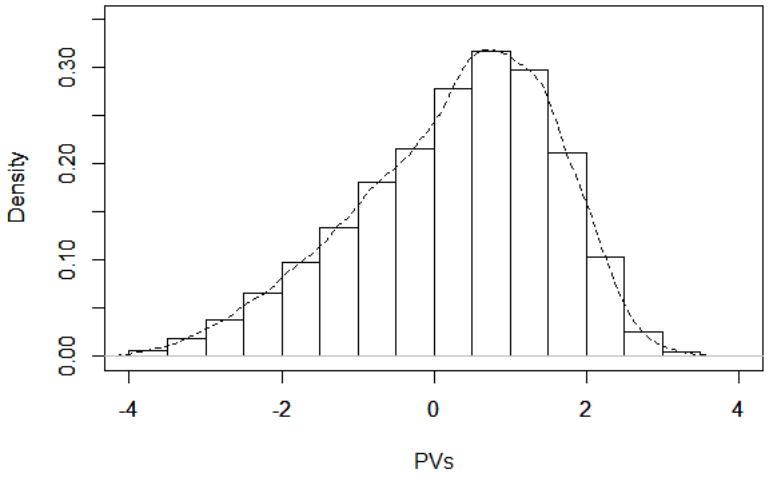

Figure 13　ヒストグラムと推定密度関数の重ね合わせ

それでは実際に小学校算数の IRT テスト得点を下位領域ごとに比較する。前節では少数項目という条件を扱うために，あえて数量関係だけに対応する項目を抽出して比較を行った。しかし複数領域に対応する項目を除いてしまうとほとんどの項目が比較対象とできないため，本節では複数下位領域に対応する項目も含めて，下位領域ごとの項目分類を行った。分類にあたっては国立教育政策研究所が公開している解説資料を参照し，学習指導要領により定められている算数の 4 下位領域，すなわち「数と計算」「量と測定」「図形」「数量関係」に対応する項目を特定した。各領域の項目数は以下の通りであった。

表２　領域別項目数

年度	全体	数と計算	量と測定	図形	数量関係
H25	32	11	11	6	11
H28	29	16	7	5	9

これらの項目を利用して推定した推算値 10 組の平均と標準偏差の平均をまとめたものが表 3 である。

表３　下位領域ごとの推算値統計表の平均

年度	全体	数と計算	量と測定	図形	数量関係
H25（平均）	0.220	0.175	0.210	0.189	0.192
（SD）	1.535	1.452	1.433	1.521	1.453
H28（平均）	0.115	0.083	0.104	0.103	0.142
（SD）	1.472	1.402	1.407	1.407	1.417

続いて，学力分布のグラフを出力する。先述したように推定密度関数で表現し，一つの領域の，二つの年度の分布を重ね合わせた。

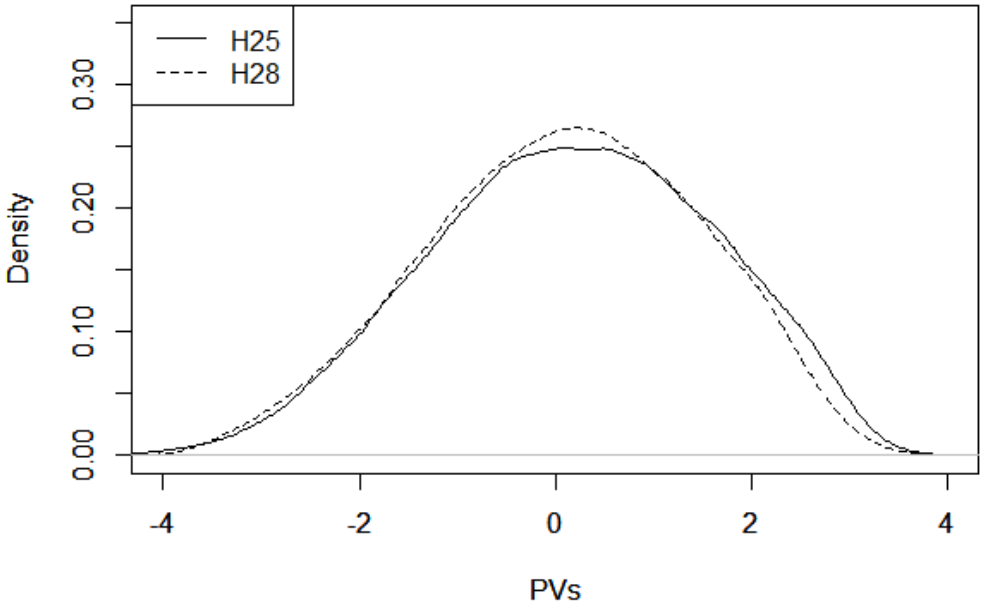

Figure 14　数と計算

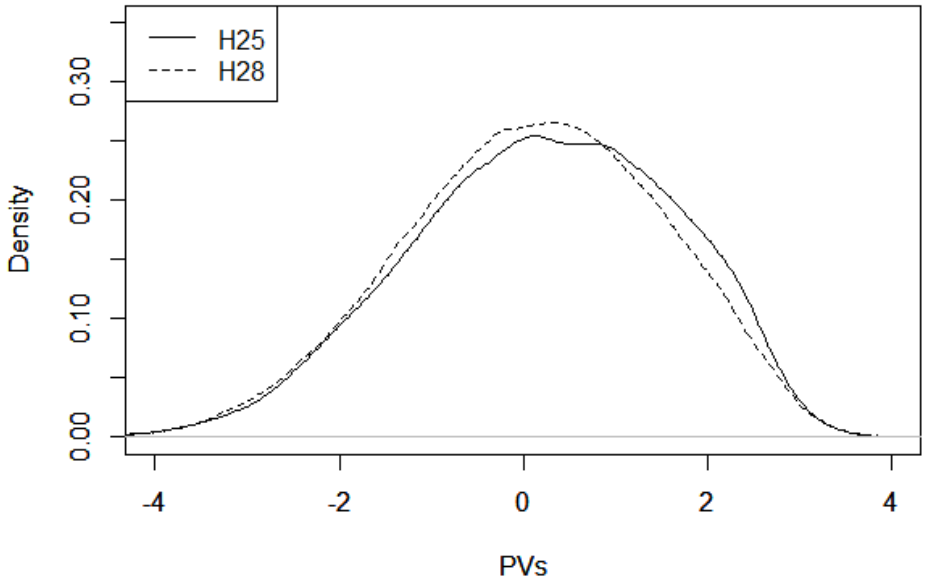

Figure 15　量と測定

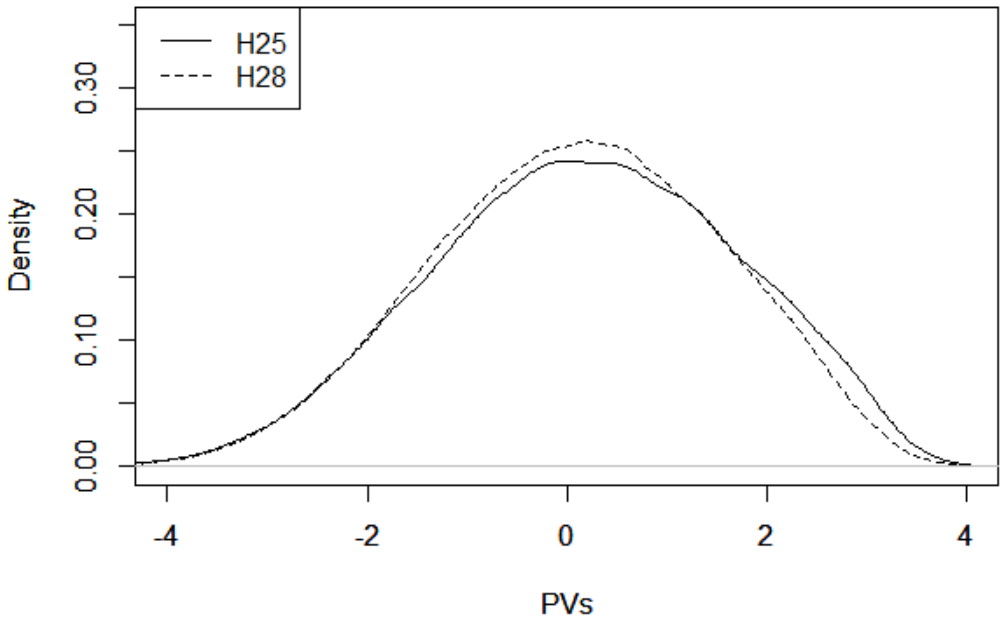

Figure 16　図形

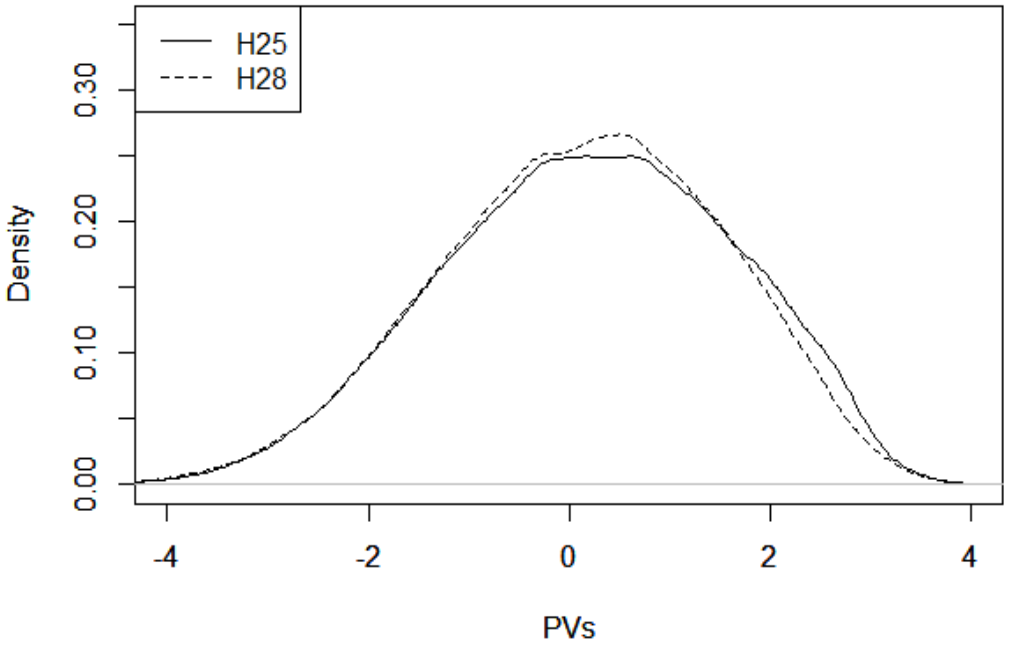

Figure 17　数量

Figure14～17 は，いずれも最大で 0.01~0.02 程度の相対度数のズレが確認できる。全項目の中で最も項目数が少ない「図形」の下位領域であっても，他下位領域と遜色ないレベルで集団の学力分布を再現し，比較可能である。推算値の値から推定密度関数を計算すれば，表３のような基本統計量だけでは分からない細かな差異も視覚的に確認できる。

柴山他（2013）や von Davier et al.（2009）が指摘するように，推算値を利用すれば能力分布のパーセンタイル値を比較的正確に推定できる。本節ではその推算値を，小数項目に限られる下位領域ごとに推定することで，より正確な領域ごとの年度間比較の可能性を示した。

文　献

Alina A. von Daivier, & Haiwen Chen （2013） The kernel Levine Equipercentile Observed-Score Equating Function, Resaerch Report, ETS RR-13-38, ETS.

von Davier, M., Gonzalez, E., & Mislevy, R. J. （2009）. What are plausible values and why are they useful? *IERI Monograph Series*, 2, 9-36

Dowle, M. & Srinivasan,A. （2017）. data.table: Extension of `data.frame`. R package version 1.10.4-3.

Han, T ., Kolen, M., & Pohlman, J. （1997）. A comparison among IRT true- and observed- score equatings and traditional equipercentile equating. Applied Measurement in Education,10,105-121.

Holland, P. W., & Dorans, J. N. （2006）. Linking and Equating. In R. L. Brennan （Ed.）, E*ducational Measurement* （4th ed.）, pp. 155-186）. Westport, CT: American Council on Education and preager.

Kolen, M. J., & Brennan, R. L. （2014）. Test equating, linking, and scaling: Methods and practices （3rd ed.）. New York: Springer-Verl ag.

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

熊谷龍一 （2009）. 初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発　日本テスト学会誌 5, 107-118.

熊谷龍一・野口裕之 （2012）. 推定母集団分布を使用した共通受検者法による等化係数の推定　日本テスト学会誌　8，9-17.

Little, R.J.A., and Rubin, D.B. （1983）. On Jointly Estimating Parameters and Missing Data. American Statistician, 37: 218–220.

Little, R.J.A., and Rubin, D.B. （2002）. Statistical Analysis with Missing Data （2nd ed..）,New York:Wiley

Lord, F. M., & Wingersky, M. S. （1984）. Comparison of IRT true-score and equipercentaile observed-
score “equatings”. Applied Psychological Mesurement, 8, 452-461.

Marco, G. L （1977）. Item characteristic curve solution to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.

Mislevy, R. （1991）. Randomization-based inference about latent variables from complex samples. Psychometrika, 56, 177-196.

Rubin, D. （1987）. Multiple imputation for nonresponse in surveys. New-York: Wiley.

Sato,Y. & Shibayama,T. （to be submitted） Reliability of linked scores under single group design: application to linkability analysis.

芝祐順 （1991）. 項目反応理論−基礎と応用− 東京大学出版会

柴山直，佐藤喜一，熊谷龍一，佐藤誠子（2012）全国規模の学力調査における重複テスト分冊法適用の試み.文部科学省平成 22 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
柴山直,熊谷龍一, 佐藤喜一,足立幸子,中野友香子（2013）.全国規模の学力調査におけるマトリックス・サンプリングに基づく集団統計量の推定について. 文部科学省平成 24 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
柴山直,熊谷龍一,後藤武俊,佐藤喜一,中野友香子（2014）.東日本大震災の学力への影響～IRT 推算値による経年比較分析～. 文部科学省平成 25 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
Stocking, M. L., & Lord, F. M. （1983）. Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-247.
Wu, M. （2005）. The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31, 114-128.