【原　著】

令和３年度大学入学共通テスト「倫理、政治・経済」を介した「倫理」と「政治・経済」の IRT True Score Equating

―2 パラメータ・ロジスティック・モデルを用いて―

橋本貴充*
荘島宏二郎*
宮澤芳光*
石岡恒憲*
前川眞一**

要　約

2021（令和 3）年度大学入学共通テストの公民の第 1 日程では、「倫理」と「政治・経済」の科目間で平均点差が 20 点以上となり、得点調整判定委員会で試験問題の難易差に基づくものと認められ、公民の科目間で得点調整が行われた。平均点に差をもたらす要因は、試験問題の難易差に限らず、例えば受験者の科目間の学力差も平均点差の原因となり得るため、受験者の科目間の学力差の影響を考慮して試験問題の難易差を議論する様々な方法が提案されている。「倫理」と「政治・経済」のデータは、「倫理、政治・経済」のデータと併せると、共通項目デザインのデータとなるため、共通項目デザインを前提とした方法を適用することが可能である。本稿では、「倫理」「政治・経済」「倫理、政治・経済」のデータに項目反応理論を適用し、「倫理」と「政治・経済」の試験問題の難易差を検討した。その結果、以下のことがわかった。まず、「公民の学力」を「大学入学共通テストの公民のテストで測られる能力」としたとき、2021（令和 3）年度の第 1 日程では、「倫理」の試験に比べ「政治・経済」の試験は難しい項目が多く、「倫理」の受験者に比べ「政治・経済」の受験者の「公民の学力」は平均的に低かった。その結果、「倫理」と「政治・経済」の平均点差が拡大してしまった。次に、項目反応理論を利用して同じ程度の「公民の学力」で得られると考えられる得点を比較したところ、「政治・経済」よりも「倫理」の方が高い得点であることがわかった。第三に、項目反応理論の true score equating を利用して、「公民の学力」の影響を考慮して「倫理」と「政治・経済」の等化を行うことを試み、その結果、得点に 20 点近くの加算があり得ることがわかった。以上の分析は、共通項目デザインのデータに適用したが、地理歴史、理科②などの共通受験者デザインのデータにも適用することができる。

キーワード：大学入学共通テスト、公民、共通項目デザイン、項目反応理論、等化

* 大学入試センター 研究開発部 試験技術研究部門
** 大学入試センター 研究開発部 客員教授

1 はじめに

2021（令和 3）年度の大学入学共通テスト（以下,「共通テスト」とする。）の公民の第 1 日程では、「倫理」と「政治・経済」の科目間で平均点差が 20 点以上となり、得点調整判定委員会で試験問題の難易差に基づくものと認められ、得点調整が行われた。平均点に差をもたらす要因は試験問題の難易差に限らず、例えば受験者の科目間の学力差も平均点差の原因となり得る。受験者の科目間の学力差の影響を考慮して試験問題の難易差を議論する方法には、(1) 前川（2020, 2022）のように、加算モデルを用いて、平均点を試験問題の難易度と科目選択者群の学力に分解する方法、(2) 石岡（2011, 2022）のように、選択しなかった科目の得点を欠測とし、ランダムフォレストを用いて補完する方法、(3) 同じく、選択しなかった科目の得点を欠測とするが、荘島（2022）のように完全情報最尤推定法を用いて周辺化する方法、(4) 吉村・荘島（2004）や荘島他（2022）のように、傾向スコアにより調整した共変量を用いて学力差を考慮する方法、(5) 大津（2011）のように、科目得点の背後に 1 次元の共通因子を仮定する方法などがある。これらの方法では、観測されたデータとして科目のテスト得点を使用し、項目への正誤の情報は使用しない。

ところが、共通テストおよび大学入試センター試験（以下,「センター試験」とする。）の「倫理」と「政治・経済」には、「倫理」および「政治・経済」と項目を共有する「倫理、政治・経済」という科目が 2012（平成 24）年度以降出題され、毎年 4 万人以上が「倫理、政治・経済」に解答している。このことは、「倫理」と「政治・経済」の試験が共通項目デザイン（common-item design, CID）（荘島, 2009）で実施され、「倫理」と「政治・経済」の科目間で等化が可能になっていることを意味する。これを利用し、橋本（2021）は 2021（令和 3）年度第 1 日程の「倫理」と「政治・経済」の平均点差について、「倫理は公民の学力の若干低い層が若干難しい問題に解答し、政治・経済は公民の学力のある程度低い層がある程度難しい問題に解答したために、平均点が大きく開いてしまったと考えられる。」と報告した。ただし、この報告は、正答率の比較のみを根拠とした分析によるものである。

共通項目デザインで実施されたテストで、受験者の学力差の影響を考慮してテストの難易差を議論するためには、項目反応理論（item response theory, IRT）がしばしば適用される。IRT は、項目ごとに困難度を推定できるだけでなく、受験者の能力（本稿では「公民の学力」に相当）という潜在変数を仮定した議論が可能であるという利点をもつ。さらに「受験者の能力」という困難度の指標と同じ尺度上の別の潜在変数を仮定することによって、困難度の指標を受験者の能力分布と独立に定義できるという利点がある。これは、受験者のグループごとの能力分布の差を考慮した分析ができることを意味する。冒頭で述べたように、平均点に差をもたらす要因は、試験問題の難易差に加え、受験者の学力差があり得るため、後者の影響を除いて試験問題の難易を議論するためには、IRT を用いた、能力に依存しない困難度指標が必要となる。

IRT の適用により、「倫理」受験者と「政治・経済」受験者の能力分布の差を推測できるだけでなく、能力分布に差があっても、同じ得点が同じ能力を示すように、テストを等化することもできる。テストの等化には様々な方法が提案されているが、observed score equating と true score equating に大別することができる（Kolen & Brennan, 2014）。observed score equating とは、等百分位法のように累積分布関数をそろえる等化方法である。true score equating とは、何らかの理論上の変数を誤差なく測定した「真の得点」をそろえる等化方法である。特に、IRT による true score equating（Lord, 1980; Lord & Wingersky, 1984）では、「能力パラメータ」の真の得点をそろえる。Kolen & Brennan（2014）によれば、IRT observed score equating に比べて IRT true score equating には、利点として、(a) 計算が容易である、(b) 変換が能力パラメータの分布に依存しない、の 2 点が挙げられ、欠点として、値をそろえる真の得点を実際に得ることはできないことが挙げられている。確かに能力パラメータの真の得点を得ることは不可能であるが、IRT true score equating の 2 点のメリットは、そのデメリットを上回るものであると考えられる。また、Kolen & Brennan（2014）

は、IRT の 3 パラメータ・ロジスティック・モデルに基づく true score equating の欠点として、当て推量パラメータの推定値の和を下回る得点や、満点では、能力パラメータの真の得点を推定できず、等化ができないことを挙げている。しかし、下方の得点については、2 パラメータ・ロジスティック・モデルを用いることでこの問題を回避できる。また、満点は、大学入試センターの得点調整が「変換前の満点は変換後の満点に」という変換を行っているため、変換前の満点に対応する能力パラメータ推定値が得られなくても、この方針を適用し、満点に変換すればよいと考えられる。さらに、たとえ 2 パラメータ・ロジスティック・モデルを用いても、満点と同様に 0 点に対応する能力パラメータ推定値も得られない。しかし、満点の場合と同様に、「変換前の 0 点は変換後も 0 点になるように」という大学入試センターの得点調整の変換方針を適用し、変換前の 0 点は変換後も 0 点とすればよいと考えられる。

本稿では、2021（令和 3）年度第 1 日程の「倫理」と「政治・経済」について、IRT の 2 パラメータ・ロジスティック・モデルを適用し、受験者の「公民の学力」の差を分離した「倫理」と「政治・経済」の難易差を議論する。また、IRT true score equating を行い、受験者の「公民の学力」の差を分離したときの「倫理」と「政治・経済」の等化を試みる。

2　方法

2.1　データ

2021（令和 3）年度共通テストの「倫理」「政治・経済」「倫理、政治・経済」の第 1 日程試験のデータに対し、正答を 1、誤答を 0、解答科目中にない項目への正誤を欠損値として使用した。2021（令和 3）年度共通テスト第 1 日程の「倫理」の項目数は 33、「政治・経済」の項目数は 31 であった。「倫理、政治・経済」の試験問題は、「倫理」に存在する試験問題の一部と、「政治・経済」に存在する試験問題の一部で構成され、2021（令和 3）年度共通テスト第 1 日程では「倫理」の 16 項目と「政治・経済」の 17 項目から構成されていた。そのため、「倫理」「政治・経済」「倫理、政治・経済」のデータを同時に扱えば、共通項目デザインを適用した試験が行われていることになる。

共通テストでは「倫理」と「倫理、政治・経済」、および「政治・経済」と「倫理、政治・経済」を同時に受験することはできない。しかし、「倫理」と「政治・経済」を同時に受験することは可能であるため、「倫理」と「政治・経済」の 2 科目を受験した受験者のデータも少数ながら存在する。したがって、分析対象のデータのイメージは図 1 のようになる。それぞれの条件の受験者数は図 1 を参照されたい。

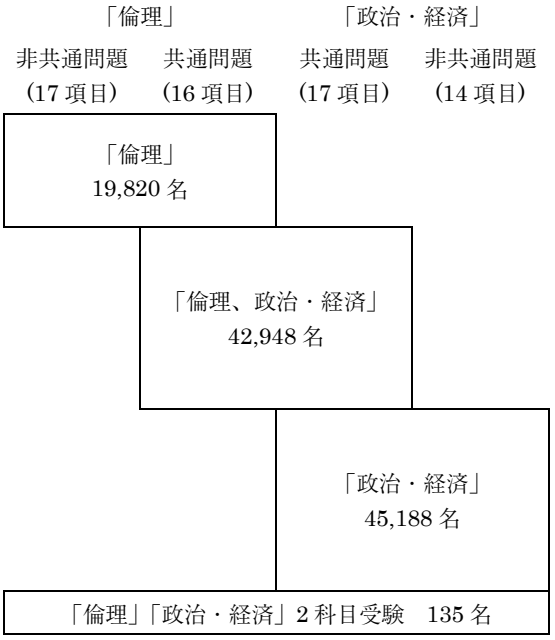

図 1　データのイメージ

2.2　モデル

本稿のデータに対して、IRT の代表的なモデルである 2 パラメータ・ロジスティック・モデルを適用した。2 パラメータ・ロジスティック・モデルは、識別力パラメータが a_j、困難度パラメータが b_j である項目 j に、能力パラメータの値が θ_i である受験者 i が正答する確率 $P(X_{ij}=1 \mid \theta_i, a_j, b_j)$ を次の式でモデル化する。

$$P(X_{ij}=1 \mid \theta_i, a_j, b_j) = \frac{1}{1+\exp\{-1.7a_j(\theta_i-b_j)\}} \quad (1)$$

識別力パラメータ a_j に定数 1.7 が掛けられているため、$a_j=1$、$b_j=0$ という項目では、θ_i の関数と

しての正答確率（項目特性曲線）が標準正規分布の累積分布関数に非常に近いものとなる。

本稿では、IRT の能力パラメータが「公民の学力」を表しているものとし、「公民の学力」を「共通テストの公民のテストで測られる能力」と定義する。その分布は、受験した科目に関わらず、「倫理」「政治・経済」「倫理、政治・経済」のうちいずれかを受験した全分析対象者の母集団平均が 0、母集団標準偏差が 1 であることを仮定した。

2.3 前提の確認

IRT の 2 パラメータ・ロジスティック・モデルでは、1 次元の能力パラメータを仮定している。本稿では、「倫理」「政治・経済」「倫理、政治・経済」をまとめたデータの 1 次元性を確かめるために、全 64 項目間の四分（テトラコリック）相関係数を計算し、固有値分解を行い、スクリープロットを確認した。これに加え、内的一貫性を確かめるために、四分相関係数、四分点相関係数（ファイ係数）、共分散を用いて、アルファ係数およびオメガ係数を計算した。

また、IRT の 2 パラメータ・ロジスティック・モデルでは、能力パラメータを所与とすれば、項目への正誤が条件付き独立となる、局所独立性を仮定している。本稿では、局所独立性を満たさない程度（つまり局所依存性の程度）の指標として、Yen（1984）の Q_3 統計量を計算した。Q_3 統計量とは、受験者の実際の正誤を 1 および 0 とし、そこから IRT で推定されるその能力の受験者の正答確率の推定値を差し引いた残差を計算し、残差どうしの相関係数として計算されるものである。

2.4 パラメータの推定

本稿のデータは、例えば「倫理」受験者の「政治・経済」の項目の正誤、「倫理、政治・経済」受験者の「倫理」の非共通問題の正誤など、欠損値を含んでいる。したがって、IRT のパラメータを推定する際には、欠損値を含んだままの矩形の項目反応パターン行列から一度に推定する方法である同時尺度調整法（川端, 2014）を適用した。同時尺度調整法には、等化係数を求める手続きが不要であること、既存の IRT 用プログラムのほとんどで実行可能であることなどの利点がある。

分析の実行には BILOG-MG（バージョン 3.0.2327.2）を使用し、図 1 の 4 群に多母集団を仮定して、2 パラメータ・ロジスティック・モデルを適用した。ただし、「倫理」受験者、「政治・経済」受験者、「倫理、政治・経済」受験者の間に「公民の学力」の差があると考えられるため、全体を通じて能力パラメータの分布の平均が 0、標準偏差が 1 となるように仮定した上で、それぞれの受験者集団の能力パラメータの平均を推定し、集団の「公民の学力」の高低を検討した。

2.5 「倫理」と「政治・経済」の比較と等化

2.2 節の式(1)は、項目パラメータの値が a_j と b_j である項目に、能力パラメータの値が θ_i である受験者が正答する確率を示している。これに、項目 j の配点 w_j を乗じ、全項目について合計した、

$$T(\theta_i) = \sum_j w_j P(X_{ij} = 1|\theta_i, a_j, b_j) \qquad (2)$$

は、能力パラメータの値が θ_i である受験者の、科目得点の推定値と解釈できる。この得点は、共通テストが返している配点で重みづけされた得点に対応している。$T(\theta_i)$ を θ_i の関数としたもののことを、テスト反応関数（test response function, TRF）という（図 2）。「倫理」と「政治・経済」について TRF を計算し、能力の同じ受験者が「倫理」と「政治・経済」を受験したときに取れると考えられる得点を比較した。

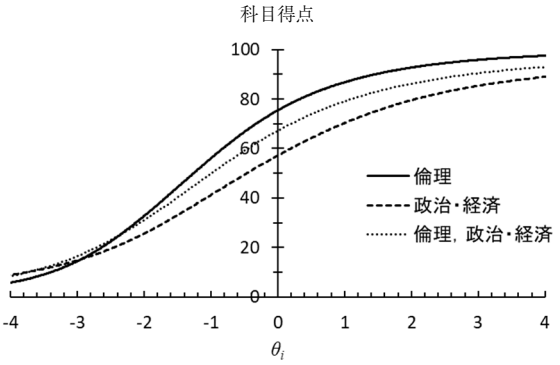

図 2 テスト反応関数（test response function, TRF）

科目間の難易を比較する際、同じ能力の受験者の得点ではなく、受験者集団の平均点がしばしば比較される。そこで、全分析対象者の能力パラメータの分布を標準正規分布で近似し、TRF に標準

正規分布の確率密度関数を乗じ、能力パラメータを積分消去し、科目得点の期待値を計算した。これは、それぞれの試験を全分析対象者が解答したときの予想平均点に相当する。

TRF を利用すると、「政治・経済」で Y_i 点を取った受験者の能力（θ_iという変数であるとする）の値が$\hat{\theta}_i$と推定されたとき、能力の値が$\hat{\theta}_i$である受験者の「倫理」の$T(\hat{\theta}_i)$を求めることで、「政治・経済」と「倫理」の等化が可能となる。本稿では、「政治・経済」と「倫理」の項目から、「公民の学力」という能力パラメータの真の得点の推定値$\hat{\theta}_i$を求め、以下の手順で IRT true score equating を行った。

1. Y_i=0 から 100 について、「政治・経済」の項目（項目通番 34 から 64）で、方程式

$$\sum_{j=34}^{64} w_j P(X_{ij}=1|\hat{\theta}_i, a_j, b_j) = Y_i$$

の解となる$\hat{\theta}_i$を、二分法アルゴリズムで求める。

2. 上で求めた$\hat{\theta}_i$と、「倫理」の項目（項目通番 1 から 33）のパラメータ推定値を、

$$\hat{Y}_i = \sum_{j=1}^{33} w_j P(X_{ij}=1|\hat{\theta}_i, a_j, b_j)$$

に代入し、$\hat{Y}_i$を求める。

3. 「政治・経済」の点 Y_i は「倫理」の$\hat{Y}_i$点に相当するとして等化を行う。ただし、$\hat{Y}_i$は四捨五入して整数に丸める。

3　結果

3.1　前提の確認

IRT の適用に先立ち、「倫理」「政治・経済」「倫理、政治・経済」をまとめたデータの 1 次元性を確かめるために、全 64 項目間の四分（テトラコリック）相関係数を計算し、固有値分解を行った（図 3）。その結果、第 1 次元の寄与率は 22.0％であるものの、第 2 次元以降の変化はなだらかであった。また、内的一貫性を計算した結果、四分相関係数、四分点相関係数（ファイ係数）、共分散から計算したアルファ係数はそれぞれ、0.935、0.882、0.872 であり、オメガ係数はそれぞれ、0.937、0.883、0.874 であった。したがって、本稿のデータは 1 次元の能力を測定していると仮定して差し支えないと考えられる。

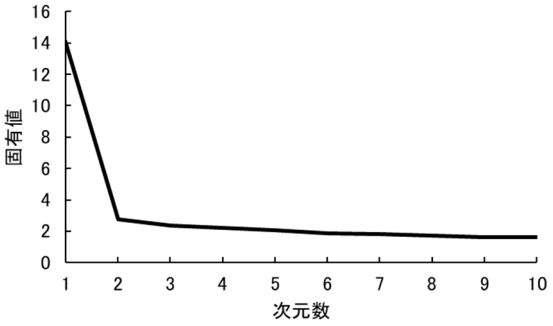

図 3　固有値のスクリープロット（10 次元まで）

局所依存性の指標である Q_3 統計量の値は、最大でも、「倫理」の項目 18 と「政治・経済」の項目 14 の間の 0.277 にとどまった。したがって、どの項目間でも局所独立性の仮定は満たされているとして差し支えないと考えられる。

3.2　項目パラメータ推定値

2 パラメータ・ロジスティック・モデルを適用した「倫理」および「政治・経済」の項目パラメータの推定値は、表 1 および表 2 のようになった。表 1 および表 2 において、「倫理、政治・経済」にも設問番号のある項目は、「倫理、政治・経済」との共通項目である。

表 1 より、「倫理」の 33 項目のうち、困難度パラメータの推定値が正の値になったものは 6 項目で、困難度パラメータ推定値の平均は－1.092、標準偏差は 1.041 であった。これに対し、表 2 より、「政治・経済」の 31 項目で困難度パラメータの推定値が正の値になったものは 12 項目で、困難度パラメータ推定値の平均は 1.549、標準偏差は 9.817 であった。設問番号 1 と 8 の困難度パラメータ推定値が極端に大きな正の値であるため、これら 2 項目を除いた 29 項目で平均および標準偏差を計算したところ、それぞれ－0.359、1.077 であった。困難度パラメータが正であることは、θ_iの値が 0、すなわち能力パラメータが平均的な値である受験者の正答する確率が 50％に満たない、難しい項目であることを意味する。したがって、「倫理」に比べ「政治・経済」は難しい項目が多かったといえる。

表1 「倫理」の項目パラメータ推定値

「倫理」問題番号	「倫理」設問番号	「倫理、政治・経済」設問番号	配点（「倫理」）	識別力	困難度
第1問	1	1	3	0.433	−0.550
	2	2	3	0.730	−2.210
	3	3	3	0.707	−0.584
	4	4	3	0.807	−2.720
	5		3	0.695	−0.919
	6		3	0.360	0.117
	7		3	0.523	0.356
	8		3	0.814	−2.065
第2問	9	5	3	0.626	−1.471
	10	6	3	0.773	−0.826
	11		3	0.916	−1.252
	12	7	3	0.993	−0.824
	13		3	0.469	−0.690
	14	8	3	0.256	1.070
	15		3	0.888	−0.590
	16		3	1.004	−1.815
第3問	17		3	0.650	−0.956
	18	9	3	0.207	1.136
	19	10	3	0.842	−2.107
	20	11	3	0.756	−0.908
	21		3	0.958	−1.227
	22		3	0.371	0.000
	23		3	0.541	−2.241
	24	12	3	0.542	−2.182
第4問	25		3	0.606	0.642
	26		2	0.299	−0.984
	27		3	0.689	−2.190
	28	13	4	0.460	−0.606
	29		3	1.288	−2.689
	30	14	3	0.822	−2.346
	31		3	0.920	−2.112
	32	15	3	0.721	−0.734
	33	16	4	0.410	−1.564

前述のとおり、「政治・経済」の設問番号1と8で、困難度パラメータの推定値が極端に高くなっていた。この2項目の正答率はそれぞれ30.8%および17.3%（設問8の正答率は「倫理、政治・経済」受験者も含む）であった。ただし、正誤と科目得点との相関は小さく、双列相関係数の値はそれぞれ0.071および−0.020であった。この2項目は難しい項目というより、「政治・経済」の他の項目と多少異なる方向性をカバーする項目であった可能性がある。

3.3 能力パラメータ推定値

科目ごとの、受験者の能力パラメータの母集団平均および母集団標準偏差の推定値は表3のようになった。能力パラメータの値は、高いほどそれぞれの項目に正答する確率が高くなる。したがって、この平均が高い集団は、「倫理」と「政治・経済」のそれぞれの科目が、「公民の学力」という同じ1次元の学力を測定していると仮定した場合に、学力の高い集団であるといえる。「倫理」受験者の能力パラメータの母集団平均の推定値は 0.038、「政治・経済」は−0.390であった。したがって、「倫理」「政治・経済」「倫理、政治・経済」の受験者の中で、「倫理」は「公民の学力」が平均より若干高い集団が受験し、「政治・経済」は「公民の学力」が平均より低い集団が受験していたといえる。また、「倫理、政治・経済」受験者の能力パラメータの母集団平均の推定値は0.394であり、「倫理、政治・経済」は「公民の学力」の高い集団が受験していたといえる。

表 2　「政治・経済」の項目パラメータ推定値

「政治・経済」問題番号	「政治・経済」設問番号	「倫理、政治・経済」設問番号	配点（「政治・経済」）	識別力	困難度
第 1 問	1		3	0.098	4.505
	2		4	0.530	0.079
	3		3	0.502	−1.163
	4		4	0.178	1.107
	5		3	0.459	0.472
	6		3	0.226	1.476
	7		4	0.129	0.837
第 2 問	8	17	3	0.017	53.938
	9	18	3	0.355	−0.289
	10	19	4	0.359	−0.049
	11		3	0.644	−0.854
	12	20	3	0.260	−1.679
	13	21	4	0.548	−0.656
	14		3	0.475	−0.104
	15	22	3	0.200	1.362
第 3 問	16	23	3	0.579	−1.144
	17	24	3	0.361	−0.680
	18	25	3	0.624	0.765
	19	26	4	0.287	1.068
	20		3	0.452	0.811
	21		3	0.657	−0.028
	22	27	3	0.188	−0.933
	23	28	4	0.486	−0.043
第 4 問	24	29	3	0.645	−1.849
	25		3	0.427	0.593
	26		3	0.674	−1.703
	27	30	4	0.534	−1.287
	28		4	0.922	−1.593
	29	31	3	0.304	−1.007
	30	32	2	0.582	−2.008
	31	33	2	0.652	−1.919

表 3　科目ごとの受験者の能力パラメータ分布の推定値

科目	受験者数	平均	標準偏差
「倫理」	19,820	0.038	1.020
「政治・経済」	45,188	−0.390	0.882
「倫理、政治・経済」	42,948	0.394	0.948
「倫理」「政治・経済」2 科目選択	135	−0.440	0.891
全体	108,091	0.000	1.000

3.4　能力別得点推定値

TRF の計算に先立ち、IRT の能力パラメータの推定値は配点の重みなしで推定されたものであるのに対し、ここでは実際の共通テストと同様に配点の重みをつけた得点を推定しているため、2.5 節の式(2)を用いて IRT を介して得られた得点推定値と、実際の得点との一致度を調べた。得点の推定値と実際の得点の差のヒストグラムは、「倫理」「政治・経済」「倫理、政治・経済」の順に、図 4～6 のようになり、五数要約および平均値は表 4 のようになった。「倫理」と「政治・経済」の得点推定値は実際の得点よりやや高く、「倫理、政治・経済」の得点推定値は実際の得点よりやや低い傾向があるものの、差の平均値は 2 点以内に収まり、四分位範囲も 7 点以内に収まっている。また、実際の得点、得点推定値、配点を一律 1 点にしたときの合計正答数の推定値、能力パラメータ推定値

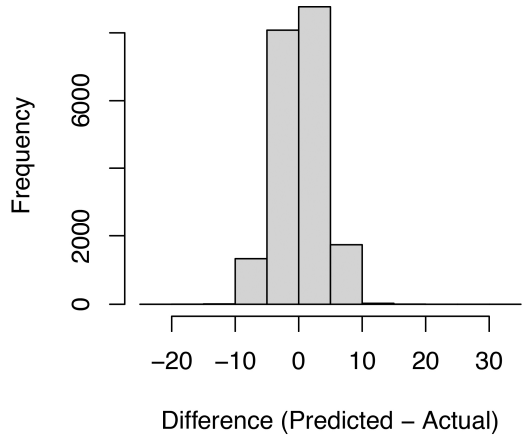

図 4 「倫理」の IRT による得点推定値と実際の得点の差のヒストグラム

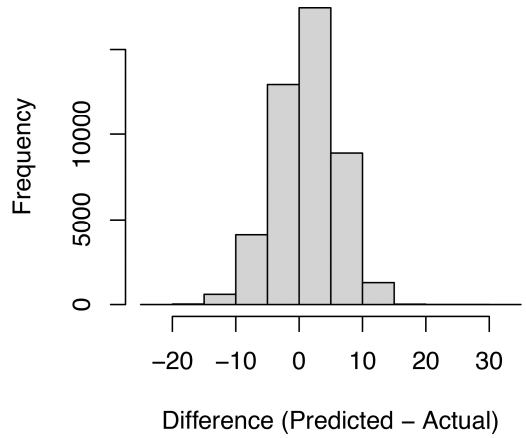

図 5 「政治・経済」の IRT による得点推定値と実際の得点の差のヒストグラム

の相関係数行列は表 5 のようになった。いずれの科目も相関係数は 0.96 を上回っており、実際の得点と推定値との間の相関は非常に強いといえる。

「倫理」と「政治・経済」について TRF を計算したところ、図 2 のようになった。$\theta_i = -3.07$ までは、同じ能力パラメータの受験者では「倫理」より「政治・経済」で高い科目得点が取れることが推定されていたが、$\theta_i = -3.07$ を超えると「倫理」の方が高い科目得点を取れるように推定された。平均的な受験者の能力といえる $\theta_i = 0$ で比較すると、「倫理」の科目得点推定値が 74.60 点であるのに対し、「政治・経済」の科目得点推定値は 56.03 点であり、18.57 点の差が開いていた。「倫理」と「政治・経済」の科目得点差が最も開いたのは $\theta_i = 0.15$ のときであり、「倫理」の科目得点推定値が 76.76 点、「政治・経済」の科目得点推定値が 58.13 点で、その差は 18.63 点であった。

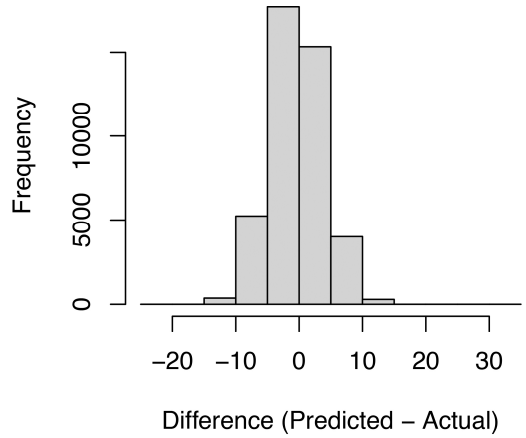

図 6 「倫理, 政治・経済」の IRT による得点推定値と実際の得点の差のヒストグラム

全分析対象者の能力パラメータの分布を標準正規分布で近似し、科目得点の期待値を計算した結果、「倫理」は 71.45 点、「政治・経済」は 55.00 点で、その差は 16.44 点であった。

「倫理、政治・経済」の TRF は「倫理」と「政治・経済」の間くらいであった。「倫理、政治・経済」の科目得点の観測された平均は 69.26 点であり、「政治・経済」の 49.87 点に比べ、「倫理」の 71.96 点にかなり近いが、能力パラメータの値が同じところで比較すると、「倫理、政治・経済」の難しさはさほど「倫理」に近いわけではないことがわかった。全分析対象者が解答したときの予想平均点は 64.15 点であり、「倫理」の期待値より 7.30 点低いものであった。

表 4 IRT による得点推定値と実際の得点の差の記述統計

科目	最小値	25%点	中央値	平均値	75%点	最大値
倫理	−19.12	−2.18	0.26	0.26	2.67	16.69
政治・経済	−21.84	−1.98	1.39	1.24	4.65	33.95
倫理、政治・経済	−14.15	−3.23	−0.45	−0.38	2.38	14.87

表 5　実際の得点，得点推定値，正答数推定値，能力パラメータ推定値の相関行列

	倫理			
	実際の得点	得点推定値	正答数推定値	能力パラメータ
実際の得点		.986	.986	.977
得点推定値			1.000	.977
正答数推定値				.977
能力パラメータ				
	政治・経済			
	実際の得点	得点推定値	正答数推定値	能力パラメータ
実際の得点		.966	.965	.966
得点推定値			1.000	.998
正答数推定値				.997
能力パラメータ				
	倫理，政治・経済			
	実際の得点	得点推定値	正答数推定値	能力パラメータ
実際の得点		.971	.971	.970
得点推定値			1.000	.986
正答数推定値				.985
能力パラメータ				

注：得点推定値と正答数推定値の相関は四捨五入すると 1.000 であったが厳密に 1 ではない（倫理は.9999995、政治・経済は.9999898、「倫理、政治・経済」は.9999687）。

3.5　能力パラメータを介した等化

表 6 は、IRT true score equating により作成した、「政治・経済」の得点を「倫理」の得点に等化する換算表である。「政治・経済」の得点の 0 点および 99 点以上に対応する θ_i の値は推定できなかった。しかし、「政治・経済」の得点の 98 点は 100 点に換算されているため、99 点以上も 100 点に換算してよいであろう。また、「政治・経済」の得点の 1 点は 0 点に換算されているため、0 点も 0 点に換算してよいであろう。この換算表に基づくと、「政治・経済」の得点が 55 点から 61 点までの場合に、等化によって 19 点加算されることになる。

4　考察

本稿では、2021（令和 3）年度第 1 日程の「倫理」と「政治・経済」のデータについて、「倫理、政治・経済」のデータと併せることで共通項目デザインのデータとし、IRT を適用することにより、受験者の「公民の学力」の差の影響を考慮してその難易差を検討した。その結果、次の 3 つのことが明らかになった。

まず、2021（令和 3）年度第 1 日程の「倫理」の試験に比べ「政治・経済」の試験は難しい項目が多く、「倫理、政治・経済」受験者に比べ「倫理」受験者の「公民の学力」は平均的に低く、「政治・経済」受験者はさらに低いという結果が得られた。橋本（2021）は「倫理は公民の学力の若干低い層が若干難しい問題に解答し、政治・経済は公民の学力のある程度低い層がある程度難しい問題に解答したために、平均点が大きく開いてしまった」と結論づけていた。本稿の分析によれば、表 3 より、「政治・経済」受験者の「公民の学力」は平均が −0.390 と推定され、表 2 より、「政治・経済」の 31 項目の困難度パラメータ推定値の平均は 1.549 であった。したがって、橋本（2021）の「政治・経済は公民の学力のある程度低い層がある程度難しい問題に解答した」という部分は本稿の分析で裏付けられた。一方、表 3 より、「倫理」受験者の「公民の学力」は平均が 0.038 と推定されたので、「公民の学力の若干低い層」ではない（ただし、「倫理、政治・経済」の 0.394 に比べれば低い）。また、表 1 より、「倫理」の 33 項目の困難度パラメータ推定値の平均は −1.092 であったため「若

表6 能力パラメータを介した「政治・経済」から「倫理」への得点換算表

得点 Y_i	$\hat{\theta}_i$	換算点$\hat{Y}_i$	得点 Y_i	$\hat{\theta}_i$	換算点$\hat{Y}_i$	得点 Y_i	$\hat{\theta}_i$	換算点$\hat{Y}_i$
0		(0)	34	−1.46	46	68	0.93	85
1	−12.45	0	35	−1.39	47	69	1.02	86
2	−8.85	0	36	−1.32	49	70	1.11	87
3	−7.25	1	37	−1.26	50	71	1.20	88
4	−6.28	1	38	−1.19	52	72	1.30	88
5	−5.62	2	39	−1.13	53	73	1.40	89
6	−5.12	3	40	−1.06	55	74	1.51	90
7	−4.73	4	41	−1.00	56	75	1.62	90
8	−4.41	5	42	−0.93	57	76	1.74	91
9	−4.15	6	43	−0.87	59	77	1.86	92
10	−3.92	7	44	−0.80	60	78	1.98	92
11	−3.71	9	45	−0.74	61	79	2.12	93
12	−3.54	10	46	−0.67	63	80	2.26	93
13	−3.37	12	47	−0.61	64	81	2.42	94
14	−3.23	13	48	−0.54	65	82	2.58	94
15	−3.09	15	49	−0.48	67	83	2.76	95
16	−2.97	17	50	−0.41	68	84	2.95	95
17	−2.85	18	51	−0.34	69	85	3.17	96
18	−2.74	20	52	−0.28	70	86	3.40	96
19	−2.64	22	53	−0.21	71	87	3.67	97
20	−2.54	23	54	−0.14	72	88	3.96	97
21	−2.44	25	55	−0.07	74	89	4.31	98
22	−2.35	27	56	0.00	75	90	4.71	98
23	−2.27	29	57	0.07	76	91	5.19	98
24	−2.18	30	58	0.14	77	92	5.77	99
25	−2.10	32	59	0.21	78	93	6.50	99
26	−2.03	34	60	0.29	79	94	7.46	99
27	−1.95	35	61	0.36	80	95	8.81	100
28	−1.87	37	62	0.44	80	96	10.91	100
29	−1.80	38	63	0.51	81	97	15.01	100
30	−1.73	40	64	0.59	82	98	31.51	100
31	−1.66	41	65	0.67	83	99		(100)
32	−1.59	43	66	0.76	84	100		(100)
33	−1.52	44	67	0.84	85			

干難しい問題」どころか「若干易しい問題」である。橋本（2021）の表現を本稿の結果で置き換えると、「倫理は公民の学力が 0.038 と若干高い層が困難度−1.092 の若干易しい問題に解答し、政治・経済は公民の学力が−0.390 とある程度低い層が困難度 1.549 のある程度難しい問題に解答したために、平均点が大きく開いてしまった」ということになる。難易差による平均点差を学力差がさらに広げてしまった事態は、橋本（2021）の見立てよりも強いものであったといえる。

次に、能力パラメータと、そこから推定される科目得点の関係より、同じ能力パラメータの受験者で比較すると、2 パラメータ・ロジスティック・モデルにおいて能力パラメータの非常に低い場合を除いて、「倫理」の方が「政治・経済」より高い科目得点を取れるという結果が得られた。このことから、「倫理」の方が「政治・経済」よりも易しいテストであったことが示唆された。

第三に、推定された能力パラメータの値を介して「政治・経済」と「倫理」の等化を行うと、「政治・経済」受験者の中には得点が最大で 19 点上昇する者がいることがわかった。受験者の科目間の

「公民の学力」の差を考慮した上で得点差を「完全に詰める」等化を行った場合、15 点を超える得点の上昇が発生する可能性があることは、本研究で得られた重要な知見といえるだろう。

最後に、本稿の分析の限界と今後の展開について述べる。第 1 の限界は、IRT の前提である。IRT の適用には、1 次元性と局所独立性が前提とされる。特に局所独立性に関しては、局所依存性のあるデータに IRT を適用すると、識別力パラメータが過大推定され、困難度パラメータが過小推定される問題が知られている（登藤, 2012）。本稿のデータは幸いにして 1 次元性も局所独立性も満たしているとみなせるものであった。しかし、項目を 1 つでも削除すると TRF が計算できなくなってしまうため、前提の妨げとなる項目を削除することはできない。前提が満たされない場合に、どのようにして TRF を求めるかが課題である。第 2 の限界は、0 点および満点付近の処理である。本稿のデータでは、「政治・経済」が 0 点、99 点、100 点の受験者で、対応する能力パラメータの値を推定できず、換算表では「政治・経済」が 1 点および 98 点のときの換算点を援用した。等化の目標科目（本稿では「倫理」）で、能力パラメータがある程度以下なら 0 点、ある程度以上なら 100 点に推定されているならば、本稿と同様に 0 点および 100 点に換算すればよく、仮に本方法をそのまま大学入学者選抜に用いても、その点で問題にはならない。しかし、目標科目で能力パラメータの推定値が高くとも TRF が 100 点にならない場合、等化の対象科目（本稿では「政治・経済」）の 99 点や 100 点に対し、得点を下げる等化を行うのかという問題が発生する。さらに、共通テストの多くの問題は多枝選択式であるため、IRT を適用するならば 3 パラメータのモデルの方が適していると考えられている。しかし、当て推量を考慮すると、特に低い得点で対応する能力パラメータの値が推定できず、true score equating による等化ができなくなってしまう。第 3 の限界は共通項目デザインである。本稿では、「倫理」と「政治・経済」に、それぞれの試験と項目を共有し、受験者の多い「倫理、政治・経済」のデータを加え、共通項目デザインとして分析を行ったが、このようにして試験問題の難易差を検討できるのは「倫理」と「政治・経済」の間だけである。しかし、多数の受験者が 2 科目を選択する「世界史 B」「日本史 B」「地理 B」の間、および理科②の「物理」「化学」「生物」の間では、共通受験者デザインを用いることで、本稿と同様の分析が可能となる。ただし、共通項目デザインと共通受験者デザインでどのように結果が異なるのか、今後検討が必要である。

参考文献

橋本貴充（2021）. 公民と数学の分析. 日本テスト学会第 19 回大会発表論文抄録集, 30-33.

石岡恒憲（2011）. Random Forest を用いた欠測データの補完に基づく大学入試センター試験科目間得点差. 応用統計学, *40*, 193-209.

石岡恒憲（2022）. ランダムフォレストを用いた欠測補完による得点調整. 日本テスト学会第 20 回大会発表論文抄録集, 36-37.

川端一光（2014）. 等化. 加藤健太郎・山田剛史・川端一光（著）R による項目反応理論（pp.243-281）. オーム社.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observedscore “equatings”. *Applied Psychological Measurement*, *8*, 453-461.

前川眞一（2020）. 加算モデルを用いた選択科目における集団の学力と試験の難易度の分離. 大学入試センター研究開発部リサーチノート, RN-20-06.

前川眞一（2022）. 加算モデルによる分析. 日本テスト学会第 20 回大会発表論文抄録集, 38-41.

大津起夫（2011）. 大学入試センター試験における科目別得点の非線形因子分析による比較. 大学入試センター研究紀要, *40*, 1-23.

荘島宏二郎（2009）. 項目反応理論. 植野真臣・永岡慶三（編）e テスティング（pp.23-48）. 培風館.

荘島宏二郎（2022）. 完全情報最尤法を用いた周辺平均推定. 日本テスト学会第 20 回大会発表論文抄録集, 34-35.

荘島宏二郎・橋本貴充・宮澤芳光・石岡恒憲・前川眞一（2022）. 傾向スコアを用いた令和 3 年度共通テスト公民科目間差の試算――国語・英語リーディング・英語リスニングを共変量に用いた場合――. 大学入試研究ジャーナル, *32*, 122-129.

登藤直弥（2012）. 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*, 81-91.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 124-145.

吉村宰・荘島宏二郎 (2004). 本追モニター調査の結果を利用した傾向スコア加重法による本追試験間の難易度比較. 大学入試センター研究紀要, *33*, 19-28.

IRT True Score Equating of "Ethics" Test and "Politics and Economics" Test through "Ethics, Politics and Economics" Test of the Common Test for University Admissions 2021: Using 2-parameter-logistic Model

HASHIMOTO Taka-Mitsu*
SHOJIMA Kojiro*
MIYAZAWA Yoshimitsu*
ISHIOKA Tsunenori*
MAYEKAWA Shin-ichi**

Abstract

In the 1st schedule tests of the Common Test for University Admissions (CTUA) in 2021, average scores of "Ethics" and "Politics and Economics" differed by more than 20 points. The score adjustment judging committee in the National Center for University Entrance Examinations (NCUEE) determined that this difference was caused by the difficulty difference of the tests, and implemented score adjustment for "Civics" subjects. Not only differences in test difficulty, but also differences in examinees' ability can lead to average score difference. Therefore, various methods have been proposed to examine differences in test difficulty with considering examinees' ability. Among them, methods for common-item design (CID) can be applied to "Ethics" and "Politics and Economics" tests, because "Ethics, Politics and Economics" test shares items with these two tests. This article applied the item response theory (IRT) to the "Ethics", "Politics and Economics", and "Ethics, Politics and Economics" tests, and examined differences in test difficulty with considering examinees' ability. We revealed the following. (1) In the 1st schedule tests of CTUA in 2021, the "Politics and Economics" test contained more difficult items than the "Ethics" test. In addition, average civics ability of "Politics and Economics" examinees was lower than that of "Ethics" examinees when the civics ability is measured by the CTUA Civics tests. Consequently, the average score difference between the "Politics and Economics" and "Ethics" tests increased. (2) Comparing examinees with the same civics ability, higher scores could be expected in the "Ethics" test than in the "Politics and Economics" test of CTUA in 2021. (3) We attempted score adjustment between "Ethics" and "Politics and Economics" tests through IRT true score equating. Consequently, it has been revealed that some "Politics and Economics" examinees' score could be added almost 20 points as the result of score adjustment. Although this analysis was applied to data designed as CID, it can be applied to data designed as common-examinee design (CED).

Key words: The Common Test for University Admissions, civics tests, common-item design, item response theory, equating

* Department of Test Technology, Research Division, The National Center for University Entrance Examinations
** Visiting Professor, Research Division, The National Center for University Entrance Examinations