

項目反応理論に基づく学力評価 —素点文化における問題点への警鐘—

中川 幸一 *

西村 光一 †

埼玉大学

日本大学

(Received May 31, 2021 Revised November 15, 2021 Accepted January 3, 2022)

概 要

In widespread achievement tests, their raw scores are usually used in order to evaluate examinees' academic ability. Item response theory (IRT) is a new theory, which is hopeful in that it has the possibility to replace traditional achievement tests based on its raw scores. Adoption of IRT enable us to evaluate the essential academic ability of examinees without being affected by each item's characteristics. Instead, some specific technical points caused by the theoretical framework of IRT, management of its item pools and its sample sizes, need to be considered. In this paper, paying attention to these points, we discuss the effective use of IRT in actual academic scenes.

1 はじめに

現代の学生の学力はテストによって評価されることが多い。テストの設計とテストデータの分析はテスト理論として数理的に理論化されている。本稿では、学力テストに関する理論として、古典的テスト理論（以下、CTT: Classical Test Theory）と項目反応理論¹⁾（以下、IRT: Item Response Theory）を紹介する。検査としての学力テストと、問題項目の性質に着目したテスト理論とを前提として、現実の教育場面での新たなテストの方法を議論する。

2 学力テスト

学力テストは受験者の学力を一連の問題項目の解答によって測定する。一般に学力テストは、さまざまな教育場面での学業の達成状況を評価することを目的として、教員から学生に対して実施されている。

学力テストは受験者の学力を測定するために実施される。一方、学力テストは、その評価上の取り扱いから、受験者にとってはより高い得点によってより望ましい評価を得ること、すなわち

*k-nakagawa@h6.dion.ne.jp

†nisimura@chs.nihon-u.ac.jp

¹⁾“Item Response Theory”の日本語訳としては、「項目反応理論」または「項目応答理論」といった訳語のいずれかが用いられる。同一の用語としての理解で差し支えない。

学力による“競技”として高得点を志向するという側面をも有する。しかしながら、適正な学力評価のためには、そのようなある種の競技性を排して、受験者の純粋な学力を過不足なく（過大でも過小でもないものとして）測定する必要があるといえる。

学力テストの適正な実施のためには、学力の測定に関する専門的な制度設計が求められる。以降では「学力」および「テスト」について概説するとともに、「学力テスト」に関する数理的な理論を紹介する。

2.1 学力の測定

学力は一般には「学業によって修得される能力」として扱われる。

学力は学力テストの得点として示される。学力とは受験者が内的に有する知的能力であり、直接的に測定することは難しい。そのような特徴から、学力は学力テストによって間接的に測定されることとなる。

学力の適正な評価のためには、適正なテストの構成を要する。われわれは日常のなかで、適正なテストによって算出された得点を、学力を適正に反映する指標として取り扱っているといえる。

2.2 テストの構成

テストは一般には「能力・特性に関する検査」として扱われる。

テストは一連の問題項目への解答（回答）を求めるものとして用いられる。教育場面での学力評価に際しては、複数の試験問題に対する解答の正誤によって学力の高低を評価する。ただし、前項で述べたように、学力テストの得点は学力の間接的な指標であり、学力の直接的な測定結果ではないことには留意しなければならない。

テストの適正な実施のためには、テストに関する系統的な知見を要する。テストを通じた受験者による問題項目への解答と、その解答を通じて検討しうる受験者と問題項目の能力や性質について、詳細に理解して評価することが重要であるといえる。

3 テスト理論

テスト理論²⁾とは、テストに関する科学的・統計的な測定と分析を扱う広範な理論である。テスト理論によって、われわれは受験者の能力・特性を測定するとともに、問題項目の性質についても検討することができる。

テスト理論の要点はテストの標準化にある。標準化とは、テストに用いられる問題項目の尺度化とテストによって示される得点の意味合いの等化とを意味する。テスト理論を用いることによって、さまざまな問題項目の性質を数量的に評価するとともに、その得点を相互に比較可能なものとして位置づけることができる。

以降、本稿ではテスト理論として CTT と IRT を概説する。

3.1 古典的テスト理論 (CTT: Classical Test Theory)

CTT は平均と分散を用いる旧来からのテスト理論である。CTT によって、受験者の正答率のほか、問題項目の妥当性と信頼性をも検討することが可能となる。CTT では分析対象のデータ

²⁾テスト理論では、学力評価に限らず、さまざまなテストが扱われる。そのなかでは、受験者・問題項目・解答といった用語について、受験者・質問項目・回答などの対応する表現もまた用いられる。

として正答数およびその和得点を扱う。

CTT の限界としては、問題項目の性質を受験者の学力と分離して扱うことができない点が挙げられる。CTT による分析では、各問題項目の解答傾向について、項目全体の解答傾向との比較によって検討することができる一方で、各受験者の解答傾向とは区別して扱うことが難しい。したがって、CTT に基づく評価では、別のテストを受験した場合（別の問題項目に解答した場合）の得点との均質な比較もまた容易ではないといえる。

以降では、CTT によって扱われる分析データとして「正答率と項目テスト相関」および「得点と測定精度」について説明する。

3.1.1 正答率と項目テスト相関

受験者数を N 人、テストの項目数を n 項目とする。このとき、受験者 i が項目 j に解答（回答）したかどうかを表す欠測指示子 z_{ij} を以下のように定義する。

$$z_{ij} = \begin{cases} 1 & \text{回答} \\ 0 & \text{欠測} \end{cases} \quad (i = 1, 2, \dots, N; j = 1, 2, \dots, n)$$

また、受験者 i の項目 j に対する項目反応 u_{ij} を以下のように定義する。

$$u_{ij} = \begin{cases} 1 & \text{正答} \\ 0 & \text{誤答} \end{cases} \quad (i = 1, 2, \dots, N; j = 1, 2, \dots, n)$$

ここで、欠測指示子と項目反応の関係は以下ようになる。

$$\begin{cases} \text{回答} \\ \text{欠測} \end{cases} = \begin{cases} \text{正答} \\ \text{誤答} \end{cases}$$

その項目に対して解答行動を起こした割合を解答率（回答率）といい、以下のように定義する。

$$\bar{z}_j = \frac{\sum_{i=1}^N z_{ij}}{N}$$

また、項目を提示された者のうち、正答した割合を正答率（通過率）といい、以下のように定義する。

$$\bar{u}_j = \frac{\sum_{i=1}^N z_{ij} u_{ij}}{\sum_{i=1}^N z_{ij}} = \frac{\sum_{i=1}^N z_{ij} u_{ij}}{N \bar{z}_j}$$

各項目が項目全体の傾向に沿っているかどうかを表す指標として、項目得点と和得点との相関（点双列相関 / ピアソンの積率相関係数）を項目テスト相関といい、以下のように定義する。

$$r(u_j, x) = \frac{\text{Cov}(u_j, x)}{\sigma_{u_j} \sigma_x}$$

3.1.2 得点と測定精度

得点（測定値） X は真値 T と誤差 E に分解することができると定義する．この分散は，真値と誤差は無関係であると仮定すると以下ようになる．

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

テストの結果（得点）の再現性に関する指標として，全体の得点の分散に占める真値の分散の割合を信頼性係数といい，以下のように定義する．

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

3.2 項目反応理論 (IRT: Item Response Theory)

IRT は項目と能力のパラメタ推定を用いる比較的新規のテスト理論である．IRT によって，一連の問題項目の性質と受験者の能力とを個別に推定することが可能となり，別のテストを受験した本人や別人との得点の比較が容易となる．

IRT による各問題項目の性質を示すパラメタとして，本稿では 2 パラメタによるモデル，すなわち 2 つの指標を用いるモデルによって説明する．

IRT の利用に際しては，十分な項目プールとサンプルサイズとを要する．なお，IRT に基づくテストの制度設計は，項目プールとサンプルサイズのほか，試験問題の非公開化および単問出題と選択解答による形式といった，いくつかの制約ともいべき要素を含んでいる．

以降では，IRT によって扱われる分析データとして「特性関数」「情報関数」およびそれらの関数によって表される曲線について説明する．

3.2.1 特性関数

テスト（項目）に対する受験者の反応から推定される属性を特性として扱う．この特性は直接観測することができない（顕在的でない）構成概念であるため，このことを強調して潜在特性とも呼ばれる．

潜在特性の尺度値（特性値） θ の受験者が項目に正答する確率を θ の関数として表したものをテスト（項目）特性関数といい，この曲線をテスト（項目）特性曲線という．

$$\begin{aligned} p_j(\theta) &= \int_{-\infty}^{a_j(\theta-b_j)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\ &= \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{a_j(\theta-b_j)}{\sqrt{2}}\right) \right) \\ &\approx \frac{1}{1 + \exp(-Da_j(\theta-b_j))} \quad (\text{ただし, } D = 1.702) \end{aligned}$$

上記の項目 j に関する項目特性関数は，上段を 2 パラメタ正規累積モデル，下段の近似モデルを 2 パラメタロジスティックモデル (2PLM: 2 Parameter Logistics Model) という．

ここで，パラメタ a ， b はそれぞれ

困難度 (b): 問題項目に対する「正答がその学力において難しいかどうかの程度」

識別力 (a): 問題項目に対する「正答がその学力を鋭敏に反映するかどうかの程度」

を表しており、

困難度 (b): $p(\theta) = 0.5$ のときの θ の値³⁾

識別力 (a): $\theta = b$ における特性曲線の傾き (傾きの最大値) に比例⁴⁾

によって定義する。

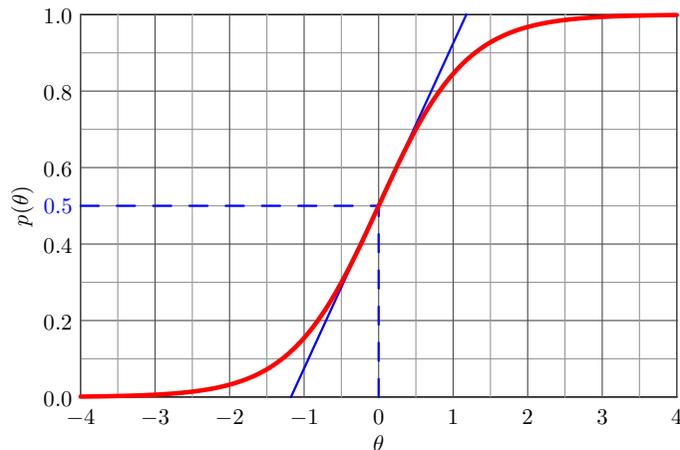


図 1: テスト特性曲線

図 1 は、困難度を 0.0、識別力を 1.0 としたときのテスト特性曲線である。

3.2.2 情報関数

テスト全体の測定精度として、一般的な統計学分野でフィッシャー情報量 (対数尤度関数の微分の二次のモーメント) と呼ばれている値を指標として表し、この情報量を曲線で表したものをテスト情報曲線という。

$$I(\theta) = D^2 \sum_{j=1}^n a_j^2 p_j(\theta)(1 - p_j(\theta))$$

誤差分散はテスト情報量の逆数 $\frac{1}{I(\theta)}$ 、標準誤差は誤差分散の正の平方根 $\frac{1}{\sqrt{I(\theta)}}$ で表される。ま

た、最大情報量は $\theta = b$ のときであり、情報関数の値は $\frac{D^2 a^2}{4} = 0.7225a^2$ となる。

³⁾ テスト特性関数の場合は $p(\theta) = 0.5$ 、項目特性関数の場合は $p_j(\theta) = 0.5$ となる。

⁴⁾ テスト特性関数の場合は $\theta = b$ 、項目特性関数の場合は $\theta = b_j$ となる。

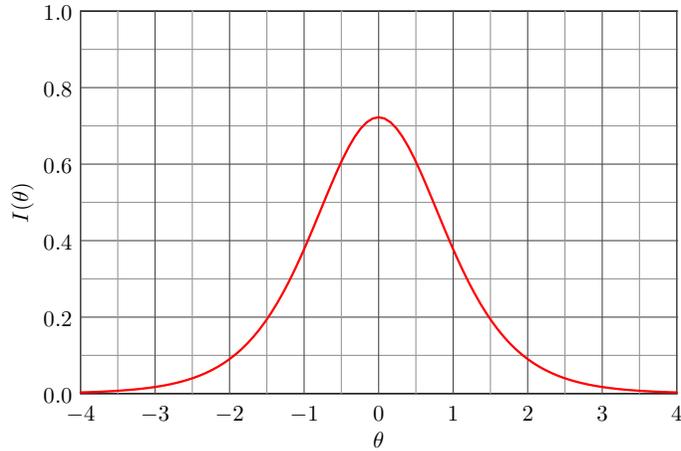


図 2: テスト情報曲線

図 2 は、困難度を 0.0、識別力を 1.0 としたときのテスト情報曲線である。面積は識別力に比例しており、また、 $\theta = b$ から離れるにしたがって、急速に情報量が小さくなるのがわかる。

3.2.3 特性曲線とパラメタの差

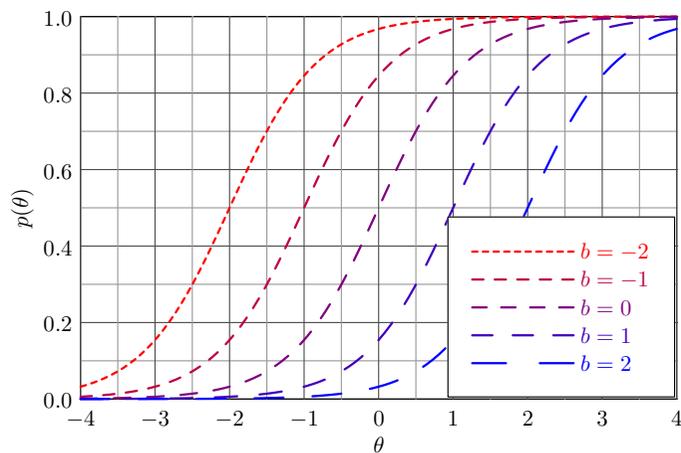


図 3: テスト特性曲線と項目困難度の差

図 3 は識別力を 1.0 で固定し、困難度をいくつか変化させたテスト特性曲線のグラフである。困難度が高いほどグラフが右にシフトしていることがわかる。これは、右にシフトするほど難易度が難しい項目であるということを表している。

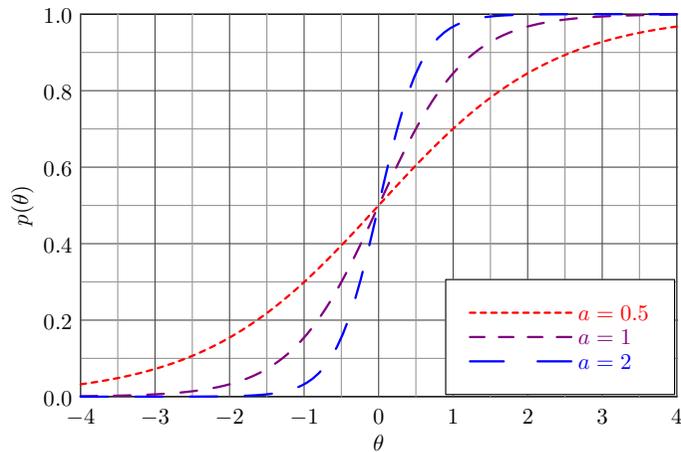


図 4: テスト特性曲線と項目識別力の差

図 4 は困難度を 0.0 で固定し，識別力をいくつか変化させたテスト特性曲線のグラフである．識別力が高いほど $p(\theta) = 0.5$ における傾きが大きいことがわかる．これは，受験者の特性値の違いが正答確率に鋭敏に反映されているということを表している．

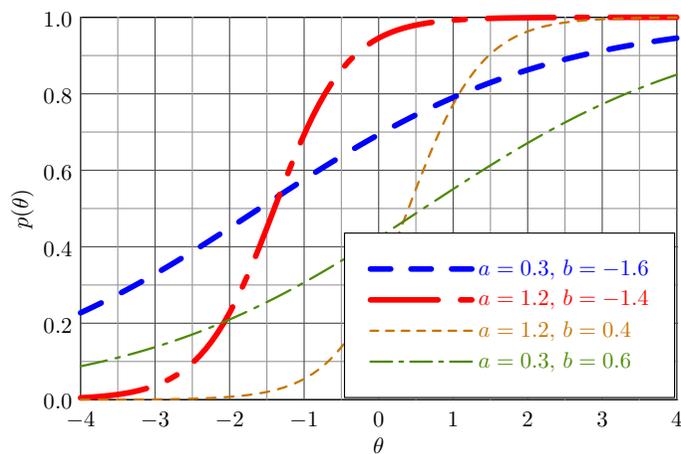


図 5: 2 パラメタモデルによる項目特性曲線の差

図 5 は以下のような特徴的な 4 つの項目特性曲線を描いたものである．

- 青（太線・ミシン目）：識別力が低く，易しい項目
- 赤（太線・一点鎖線）：識別力が高く，易しい項目
- 黄（細線・ミシン目）：識別力が高く，難しい項目
- 緑（細線・一点鎖線）：識別力が低く，難しい項目

青（太線・ミシン目）と赤（太線・一点鎖線）の困難度は同程度だが、 $\theta = -3.0$ 付近と $\theta = 0.0$ 付近で難しさ（正答率）が逆転していることがわかる。また、黄（細線・ミシン目）は、青（太線・ミシン目）よりもずっと困難度が高いが、 $\theta = 2.0$ 付近では青（太線・ミシン目）のほうが正答率が低いということもわかる。

3.2.4 情報曲線とパラメタの差

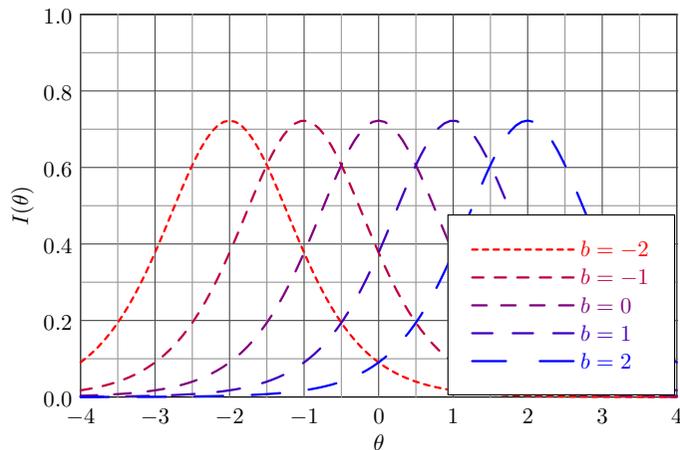


図 6: テスト情報曲線と項目困難度の差

図 6 は識別力を 1.0 で固定し、困難度をいくつか変化したテスト情報曲線のグラフである。困難度が高いほどグラフが右にシフトしていることがわかる。テスト情報曲線は項目情報曲線の重ね合わせとして表されるため、困難度にはばらつきがあるほうが、 θ が b から離れたところでも情報量が小さくなりにくくなる。

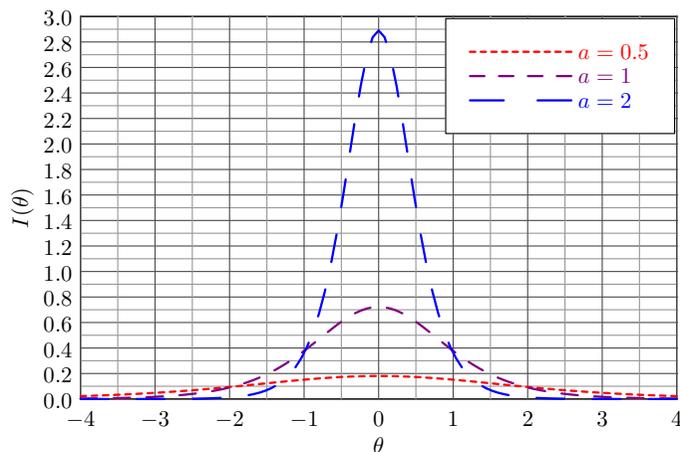


図 7: テスト情報曲線と項目識別力の差

図 7 は困難度を 0.0 で固定し、識別力をいくつか変化させたテスト情報曲線のグラフである。識別力が高いほど最大情報量が大きいことがわかる。識別力が高いほど最大情報量は大きいですが、 θ が b から離れると情報量が逆転する。そのため、識別力が高すぎるのもよくない。

4 CTT と IRT のシミュレーション比較

CTT と IRT の差異について、以降では疑似データを用いて比較する。CTT に基づく正答率の算出に加えて、IRT に基づく項目パラメタおよび受験者の θ を推定する。CTT に基づく素点得点と IRT に基づく尺度得点との比較を通じて、CTT と IRT の差異および IRT の特色について検討する。

4.1 計算機実験

現実的な状況を想定して、以下の条件で Mathematica 12.3 によるシミュレーションを実施した。

設問数： 10 問

選択肢数： 4

受験者数： 5000 人

ここで、受験者の能力が同じである場合に「局所独立性の仮定（一方の項目解答の正誤は他方の項目解答の正誤に影響を及ぼさない（各項目への正誤反応が独立である）という仮定）」を設けているため、選択肢の選択方法は完全にランダムとなっている⁵⁾。IRT の分析ツールとしては EasyEstimation Ver.2.1.5 を用いた。

表 1: 項目別正答率と項目パラメタ

| 項目番号 | 正答率 | a | b |
|------|---------|---------|----------|
| 1 | 24.34 % | 0.06943 | 9.62352 |
| 2 | 24.58 % | 0.11638 | 5.70632 |
| 3 | 25.38 % | 0.07623 | 8.34002 |
| 4 | 23.96 % | 0.06972 | 9.75921 |
| 5 | 24.98 % | 0.05221 | 12.34324 |
| 6 | 25.82 % | 0.04961 | 12.34324 |
| 7 | 24.94 % | 0.27950 | 2.42743 |
| 8 | 24.58 % | 0.06172 | 10.69514 |
| 9 | 24.84 % | 0.06748 | 9.66453 |
| 10 | 26.42 % | 0.06431 | 9.37997 |

⁵⁾疑似乱数発生アルゴリズムとして、拡張セルオートマトンアルゴリズムを用いた。なお、ほかに線形合同発生アルゴリズム、メルセンヌツイスタシフトレジスタ生成アルゴリズム、Intel MKL アルゴリズム、Wolfram 則 30 アルゴリズムのそれぞれによる検証を実施した結果、これらのアルゴリズムの種類による差異として本稿で特筆すべきものは認められなかった。

これより，表 1 を用いることで受験者の θ を推定することができる．テスト特性曲線より IRT による尺度得点を求めると図 8 のようになる．今回のシミュレーションでは，受験者の θ の範囲は $-11.249 \leq \theta \leq 21.388$ であったため，尺度得点の範囲は 7.6057 ~ 81.1054 点となった．

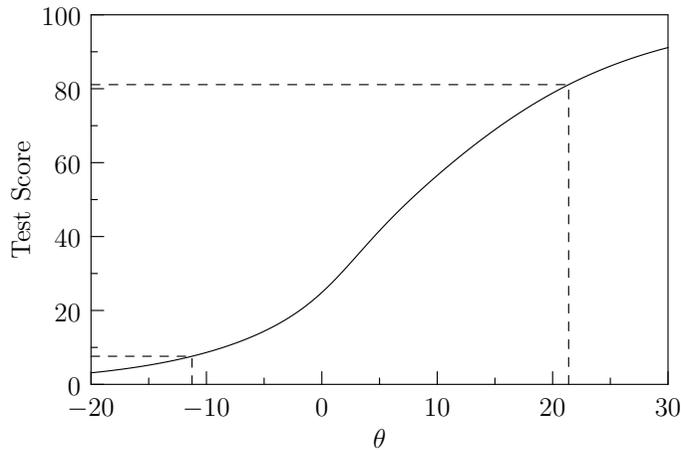


図 8: IRT による尺度得点

また，同一のシミュレーションデータにおいて，CTT に基づく 1 問 10 点として採点した結果をヒストグラムにしたものを赤・斜線，IRT に基づく尺度得点によって採点した結果をヒストグラムにしたものを青・水玉とし，重ね合わせたものが図 9 となる

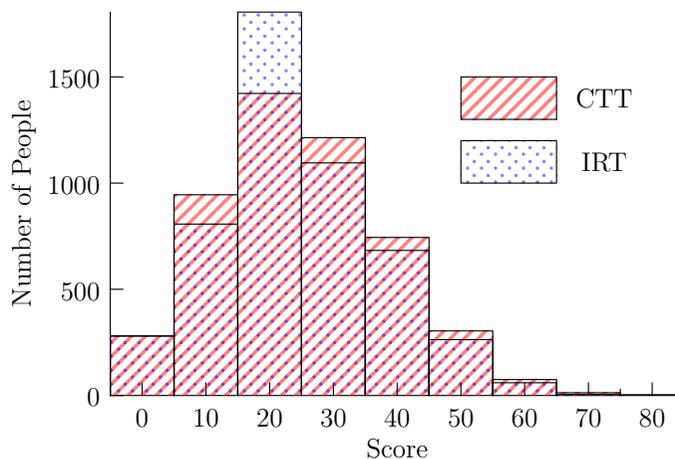


図 9: 得点のヒストグラム

各採点方法について，平均点と標準偏差を求めると表 2 のようになった．今回の規模のデータでは，平均と標準偏差ともに，双方の間に顕著な差異がみられなかったことがわかる．

表 2: 得点の平均と標準偏差

| | CTT | IRT |
|------|---------|---------|
| 平均 | 24.9840 | 23.8534 |
| 標準偏差 | 13.6995 | 13.0688 |

このように、一連の問題項目に対する解答を採点して得点化する限りにおいては、CTT と IRT の間に顕著な差異はみられない。このことは、CTT と比較して、IRT がなんらかの特殊な対象を分析しているということではなく、受験者の解答を“受験者の学力の指標”として、CTT と同様に測定・分析していることを意味する。

4.2 IRT の特色

では、CTT と IRT の間にみられる質的な差異とはなにか。前項で扱ってきた CTT と IRT のデータ比較に関連する主なものとして、次の 3 点が挙げられる。

第一に、CTT の素点得点と IRT による尺度得点は、その得点自体の意味合いが異なる。CTT では素点による得点が扱われるのに対して、IRT では項目パラメータを通じて算出される θ （および、その値を変換した尺度得点）が扱われる。CTT に基づく得点が「受験者の当該テスト（の各問題項目）に限った得点」を意味する一方で、IRT に基づく得点は「受験者の当該テスト（の各出題内容）に関する学力」を意味する。すなわち、IRT による θ （と尺度得点）は、CTT による素点得点よりも、より正確で安定的な評価をもたらすといえる。

第二に、問題項目の性質として得られる情報についても明確な差異がある。CTT によっても、識別力や困難度といった指標を得ることはできる。前述のうち、識別力には項目テスト相関、困難度には項目正答率がそれぞれ相当する。しかしながら、これらの指標はいずれも同じテスト内での相対的な位置づけを示すものであり、別のテストの問題項目との比較に適するものではない。一方、IRT による項目パラメータとしての識別力や困難度は、問題項目に関する普遍的な指標として、別のテストのさまざまな問題項目との比較にも適することとなる。

第三に、得点の等化 (equating) に関する差異もまた現実的に重要である。異なるテストによる得点を相互に比較可能とするためには、得点の等化という手続きが必要となる。等化に関しては、一般に線形変換法や等パーセントイル法といった手法が用いられる。ただし、CTT に基づく得点の等化は、どのような手法であってもその精度に問題を抱えている。一方、IRT に基づく得点の等化に際しては、それらの異なるテスト間に共通する項目（以下、共通項目）または共通する受験者（以下、共通受験者）を設けることによって、より精度の高い等化を実現することが可能となる。

5 教育場面での IRT の利用に関する考察

IRT はこれまでに一定規模の各種試験で広く利用されている。今後、IRT に基づくテストの実施は、どのような教育場面で想定されるのであろうか。

今後の IRT の利用に関しては、特に新型コロナウイルス感染症 (COVID-19) の影響下で、コンピュータに基づく試験方式 (CBT: Computer Based Testing) への適用が主に想定される。コン

コンピュータを用いる場合には、その都合上、これまで一般的に実施されてきた紙上の試験方式 (PBT: Paper Based Testing) と比較して、受験者がさまざまな時間に、さまざまな場所で、さまざまな問題項目に解答することもまた検討される。そのような試験の制度設計として、受験者の得点を (より正確かつ安定的なものとして) 相互に比較可能とするためには、IRT に基づくテスト設計が不可欠であるといえよう。同様に、多数の受験者が一回限りの試験を受験する旧来の形式に限らないという意味では、各種教育場面での定期的な学力試験のほか、いわゆる小テスト的に扱われる学習達成度の測定のような比較的小規模なテストに際しても同様の意味で有用であると考えられる。

以下、それらの教育場面での IRT 利用上の課題である項目プールとサンプルサイズの課題に関して、その展望を議論する。

5.1 項目プール

項目プール (項目バンク) は、問題項目の内容および性質を記録した、試験問題のデータベースに相当する。各問題項目の設問と解答選択肢、出題分野と作成年月日、および各項目パラメタ (困難度・識別力) が一覧として保持・利用される。

IRT に基づくテストでは、項目プールの管理を欠かすことができない。テストの実施に際しては、項目プールの中から、想定される受験者層に適したパラメタの問題項目を選定することとなる。そのため、IRT の利用に際しては、十分な項目プールを作成して有効に用いる必要がある。

項目プールについては、各種試験実施主体の間で共通の項目プールを利用するという方法も可能かもしれない。その際に共通項目や共通受験者が存在するなら、当該項目プール内の問題項目の利用に際して、項目パラメタの更新や得点の等化などに関しても実施上の利点があるといえる。

5.2 サンプルサイズ

IRT による項目パラメタの推定には、サンプルサイズにも気を配る必要がある。十分な精度の項目パラメタを算出するためには、各問題項目に対する多数の受験者の解答が必要となる。

サンプルサイズに関して、各種試験実施主体の間で共通項目や共通受験者を用いることができるならば、それらを相互に利用することで、合計で一定数のサンプルとして扱うこともできよう。それでもサンプルが不足する場合は、より少数のサンプルデータによって扱うことのできる IRT のモデルとして、項目パラメタが困難度のみに限られた 1 パラメタモデルの利用もまた検討できる。

6 おわりに

CTT と IRT という 2 種のテスト理論について、それぞれの理論間での定義・解釈の差異を表 3 にまとめておく。

本稿にて述べてきたとおり、IRT は CTT と比較して、テストの実施に関する非常に有用な理論といえる一方で、各種教育場面での現実的な利用上の課題をも抱えている。そのために、学力テストでの IRT の利用は、一定以上の規模の試験にほぼ限られている。

IRT の理論的・実践的な発展のひとつとしては、IRT を今後どのような現実の教育場面で利用できるかという論点が重要であるといえよう。現実的には、各種試験の CBT への移行に際して、

表 3: CTT と IRT の比較

| | 古典的テスト理論 | 項目反応理論 |
|------|---|---|
| 困難度 | 項目の正答率（通過率） | $p(\theta) = 0.5$ に対する θ の値 |
| 識別力 | 項目テスト相関 値が高いほど受験者の能力の高低をよ く区別する． | 特性曲線の傾きの最大値に比例 受験者の特性値の違いが正答確率にど の程度鋭敏に反映するか． |
| 得点 | 得点 = 真値 + 誤差 | テスト特性曲線 $p_j(\theta)$ |
| 測定精度 | 信頼性係数，標準誤差 テスト全体で一つの信頼性，一つの標 準誤差しか算出できない． | テスト情報関数，項目情報関数の標準 誤差 特性値のレベルごとに評価可能． |
| 依存性 | 標本に依存する． | 標本とは独立に推定される． |

それぞれの試験形式に合致した利用方法が検討されていくと考えられる．さらには，さまざまな教育機関や教育プログラムのなかで，共通項目や共通受験者の利用により，問題作成の負担を軽減しつつ，簡便かつ有効な形で IRT に基づくテストの実施を実現することもまた考えられるのではないだろうか．その数理的手法や各種指標に関する検討，およびその実務的な利用条件に関する検討が今後さらに進展することが望まれる．

謝 辞

概要の検討にご協力いただいた福田龍太郎氏と児玉公一郎氏，および本稿の修正に際して有用なご意見とご指摘をくださった査読者のみなさまに謹んで御礼を申し上げます．

参 考 文 献

- [1] 植野 真臣, 荘島 宏二郎: 学習評価の新潮流, シリーズ 行動計量の科学 4, 朝倉書店, 東京, 2010.
- [2] 豊田 秀樹: 項目反応理論 [入門編] (第2版), 統計ライブラリー, 朝倉書店, 東京, 2012.
- [3] 加藤健太郎, 山田剛史, 川端一光: R による項目反応理論, オーム社, 東京, 2014.
- [4] 熊谷龍一: 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, 2009.
- [5] Baker, Frank B., Kim, Seock-Ho.: *Item Response Theory: Parameter Estimation Techniques (2nd ed.)*, Marcel Dekker, 2004.
- [6] 村上 隆 (編著): 我が国の公的試験における得点等化の導入に向けた心理・教育測定的研究, 文部省科学研究費補助金研究成果報告書, 東京, 2003.
- [7] 独立行政法人大学入試センター: 大規模入学者選抜における CBT 活用の可能性について (報告), 独立行政法人大学入試センター, 東京, 2021. (https://www.dnc.ac.jp/research/cbt/cbt_houkoku.html)