第４章　IRT に基づく共通テストの実施

現在 CBT で行われている大規模試験の多くは，IRT（Item Response Theory，項目反応理論）というテスト理論を活用して実施されることが多い[30]。本章では，IRT とは何か，共通テストを IRT に基づいて実施する場合，どのような試験になるのか，従来の同一時刻一斉実施の試験と比べてどのような違いがあるのかなどについて述べた上で，IRT に基づいて試験を実施する場合の主な課題と必要な対応について述べる。なお，IRT 自体は必ずしも CBT で行う必要はなく，実際に PBT で行っているケースもある。

１．IRT に基づく試験

(1) IRT とは何か

○　IRT とは，各受験者の試験問題に対する正答・誤答に基づいて，試験問題の特性と受験者の能力を分けて推定する統計理論の一つである。IRT に基づいて試験を実施する場合，以下のようなことが実現できる。

◆　**異なる試験問題に解答した受験者同士の能力が比較できる（そのため試験の複数回実施が可能）**

受験者が異なる試験問題に解答した場合でも，不公平性を排除して得点を算出することが可能になるため，異なる試験問題に解答した者同士の得点も比較できる。このため，例えば，IRT に基づいて共通テストを実施すれば，現行では本試験，追試験をそれぞれ年１回のみ実施しているところ[31]，年に複数回の試験を実施できるようになる可能性がある。

◆　**統計的品質が管理された試験問題を出題できる[32]**

上記のように異なる試験問題に解答した受験者同士の能力を比較可能にするためには，難易度推定のための事前の予備調査[33]を実施して調査参加者にあらかじめ試験問題に解答しても

---

[30] CBT において IRT が活用されるケースが多い理由としては，一度に受け入れ可能な受験者数に限りがあり，同一時刻一斉実施ではなく IRT に基づき複数回に分割実施をせざるを得ないという消極的な理由から，CBT であれば IRT に基づく試験を効率的・効果的に実施することが可能であるという積極的な理由まで，様々なものが考えられる。

[31] 令和３年度共通テストは，新型コロナウイルス感染症の影響に伴う学業の遅れに対応できる選択肢を確保するため，本試験を２回実施した。

[32] ただし，IRT の出題方式は多様で，その中でも予備調査による事前の統計的品質管理が必須でない方式も存在する（分冊方式やアンカーテスト方式など）。そのような方式で実施する場合は，試験問題の品質が事前に明らかにされるわけではない。

[33] 予備調査の方法としては，受験者集団に近い集団を対象にした「プレテスト」の実施や，本番の試験において新作の試験問題（案）をダミーとして出題することなどがある。予備調査については，**【コラム⑪】**で詳述。

らい，その解答データから試験問題の品質を推定しておく必要がある。このプロセスの中で，品質が基準を満たさない試験問題（例えば，難しすぎる又は易しすぎる試験問題，能力の高い受験者と低い受験者で正答率があまり変わらない試験問題など）を除くことになる。すなわち，試験問題の統計的品質が一定程度確保された試験問題によって，試験を実施できるということになる。

○　IRT に基づく場合には，本番の試験前に予備調査を通して試験問題の統計的品質管理を行うことが一般的であるが，試験問題の品質を示す指標は「項目パラメタ」と呼ばれる。IRT モデルには，日本における多くの大規模試験で用いられている２パラメタ・ロジスティックモデルや１パラメタ・ロジスティックモデルをはじめ様々なものが存在する[34]が，項目パラメタのうち代表的なものが以下の二つである。

◆　**難易度（difficulty）パラメタ**

２パラメタ・ロジスティックモデルや１パラメタ・ロジスティックモデルを用いる場合，各試験問題の難易度パラメタは，受験者の正答確率が 50%になるところの能力値を意味する。

◆　**識別力（discrimination）パラメタ**

識別力とは，試験問題が受験者の能力の高低をどの程度敏感に捉えることができるかを示す指標である。通常，個々の試験問題を見ても，テスト全体の得点（estimated score）が高い受験者の正答率は高く，得点が低い受験者の正答率は低くなるが，受験者の能力値によって正答確率が大きく変化するのであればその試験問題の識別力は高く，あまり変化しないのであれば識別力は低いということになる。一般的に，試験問題には一定程度の高さの識別力が求められる。

○　IRT に基づく試験の実施方式には様々なものがあるが，代表的な実施方式としては以下のようなものがある。

◆　**リニア方式**

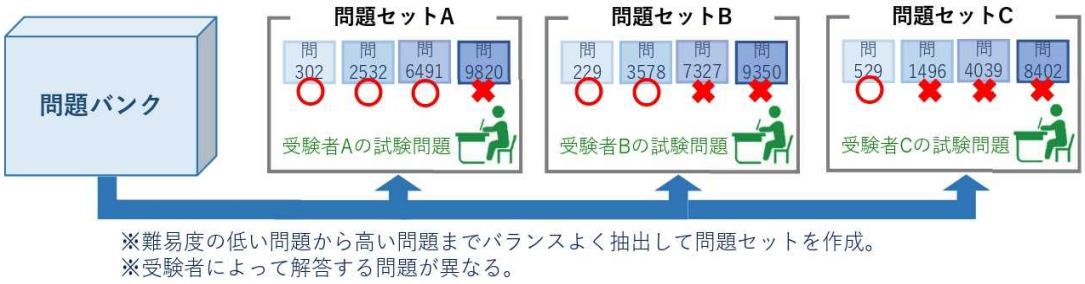
TOEFL iBT®テスト[35]，TOEIC® Program，医療系大学間共用試験 CBT 等において採用されている方式である。統計的品質のそろった試験問題セットをあらかじめ複数作成し，試験の実施日（時）や受験する国・地域の異なる受験者集団に対して異なる試験問題セットを使用して実施する。各試験問題セットの品質がそろっているため，試験問題セットの違いによる有利

---

[34] IRT モデルとしては，例えば，以下のようなものがある。
・１パラメタ・ロジスティックモデル：１種類の指標（難易度パラメタ）を用いるモデル。
・２パラメタ・ロジスティックモデル：２種類の指標（難易度パラメタ・識別力パラメタ）を用いるモデル。
・３パラメタ・ロジスティックモデル：３種類の指標（難易度パラメタ・識別力パラメタ・当て推量パラメタ（受験者が偶然に正答できる確率））を用いるモデル。
・部分採点モデル：試験問題の正誤だけでなく部分点も扱うモデル。

[35] TOEFL iBT®テストでは，リーディング・セクションとリスニング・セクションおいては IRT に基づいて得点の等化がなされているが，スピーキング・セクションとライティング・セクションにおいては，IRT ではなく等パーセンタイル等化法に基づく手法により等化がなされている。

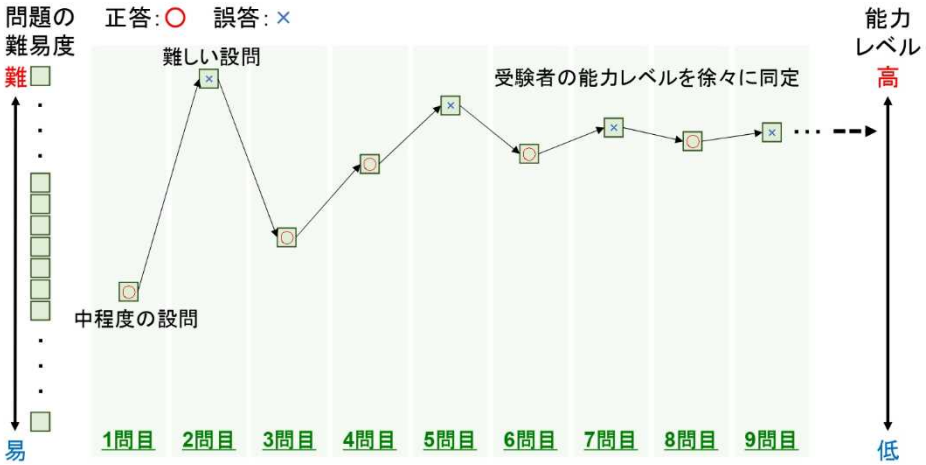
不利は基本的に生じない。



【図６】リニア方式

◆ **アダプティブ方式（コンピュータ適応型テスト・Computer-adaptive Testing；CAT）**

GMAT や TOEFL CBT®テスト（1998～2006 年に実施）[36]において採用されている方式である。具体的には，一人一人の受験者について，1 問ごとの解答の正誤に応じて能力を推定し，その受験者の能力測定の観点から最適な問題（例えば，正答確率が 50%ほどの，解けるか解けないかが不確実であるような問題）を出題する（このため，通常，正答すれば次により難しい試験問題が，誤答すれば次により易しい問題が出題される。）。受験者集団の能力差が大きい場合でも，個々の受験者ごとに難易度がカスタマイズされた形で試験問題を出題するので，より少ない出題数で能力を測定できるというメリットがある。



【図７】アダプティブ方式

○ IRT に基づいて試験を実施する際は，同一時刻に一斉実施する試験と異なり，本番の試験で出題する前に予備調査を行い，試験問題を本番の受験者に近い仮想受験者に解答してもらい，試験

[36] TOEFL CBT®テストの後継である TOEFL iBT®テストでアダプティブ方式を採用しなかった主な理由の一つに，試験後に人間が採点をする問題（新たに加わったスピーキング・セクションなど）が含まれるため，試験中に能力値を推定して適切な難易度パラメタの問題を出題することが困難になったことが挙げられる。

問題の品質をあらかじめ推定した上で，本番の試験を実施することになる。これを行うには，

・出題される問題が特定されないよう，多数の試験問題を用意する

・同じ試験問題を本番の試験で何度も利用するという設計の場合，試験問題を非公開とする[37]

といったことなどが必要になる。

○　また，IRT に基づく試験では，統計的品質管理を行った試験問題を大量に蓄積したデータベースを構築する場合が多い。このデータベースは「問題バンク」（item bank）と呼ばれる。試験実施時は，この問題バンクから，あらかじめ定められたルールに従って試験問題を選んで出題する。

【コラム⑨】２パラメタ・ロジスティックモデル

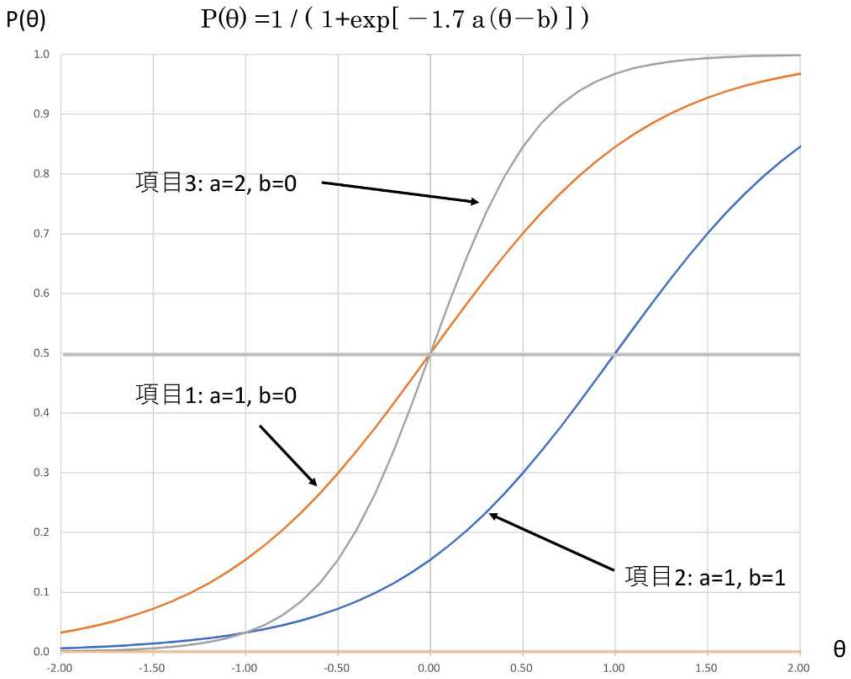
IRT では試験問題への受験者の反応の最小単位（例えば，共通テストであれば，各大問に含まれる小問の解答枠）を項目（item）と呼び，試験問題は項目の集合であると考える。そして，これらの項目は全て，受験者のある一つの特性（trait；ここではそれを能力と呼び，θ という記号で記す。）だけを測っていると考える[38]。

これらの項目の統計的品質は項目パラメタと呼ばれる数値で表現される。項目パラメタの数は用いられるモデルの種類により異なるが，ここでは，項目の識別力を表す a パラメタと項目の難易度を表す b パラメタの二つのパラメタを持つ２パラメタ・ロジスティックモデルと呼ばれるものに関しての説明を行う。

【図コラム⑨－１】に，共に受験者のある一つの特性を測っていると想定される三つの項目の項目特性曲線（item characteristic curve，項目反応曲線とも呼ばれる。）を示した。項目特性曲線は，受験者の能力が高くなるにつれ各項目への正答確率が増加していく様子を示したものである。

[37] 現行の共通テストの試験問題については，受験者に問題冊子の持ち帰りが認められているほか，試験翌日には多くの新聞に掲載され，後日，大学入試センターのウェブサイトにも掲載されるなど，広く公開されてきた。

[38] 各項目が受験者のただ一つの特性のみを測っているという仮定を一次元性の仮定と呼ぶ（**【コラム⑬】**）。



【図コラム⑨－１】項目特性曲線

【図コラム⑨－１】から明らかなように，能力が高くなればなるほど，どのような項目にも正答する確率が増加していき，最終的には（グラフの右部分で）全ての項目の正答確率は 1 に近づく。また，能力が低くなれば正答確率は減少していく。能力を表す θ は理論的には-∞から+∞の値をとるが，図には -2 から 2 までの範囲を示している。

【図コラム⑨－１】で，例えば，項目１と項目２の項目特性曲線を比べると同じ θ の値でも項目１の正答確率の方が高くなっていることが分かるが，このことは項目１の方が易しい項目であることを示している。これを項目パラメタの違いで見ると，二つの項目の a パラメタの値は共通であるが項目２の b パラメタの値が項目１の b パラメタの値より大きい。この b パラメタの値は図中に縦軸の 0.5 のところに引いた横線と項目特性曲線との交点の θ の値であり，正答確率が５割となる θ の値を意味しているが，その意味で，b パラメタを「難易度パラメタ」と呼ぶ。

また，項目１と項目３を比べると，b パラメタの値は 0 で共通であるが項目３の a パラメタの値が大きく，それが，項目特性曲線の θ が 0 付近での傾きの大きさに反映されている。これは，項目１では θ が 0 から 1 へ変化しても正答確率の変化は 0.35 程度であるが，項目３の場合は 0.45 も変化するというように，θ の変化に対する正答確率の変化の大きさ（感度）を表す指標であり，a パラメタは「識別力パラメタ」と呼ばれている。言い換えれば，識別力が高い項目ほど，小さな θ の増加であっても敏感に反応して正答確率が増加する。

なお，項目特性曲線は a パラメタと b パラメタの値と指数関数 exp[]を用いて

$$P(\theta) = 1 / (1 + \exp[-1.7\, a\, (\theta - b)])$$

というロジスティック関数の形で示すことができる。

項目の識別力が異なる場合，項目の難易度の順番が受験者の能力に依存して変化する場合がある。例えば，項目１と項目３の難易度パラメタ（b パラメタ)の値は同じであるが，項目特性曲

線を比べると能力の上位層にとっては項目３の方が易しく，能力の下位層にとっては項目１の方が易しい。

（執筆：前川眞一（(独)大学入試センター特任教授，東京工業大学名誉教授））

(2) 異なる試験問題に解答した受験者同士の能力の比較

○　異なる試験問題セットに解答した受験者同士の能力を比較する際に，試験問題セット間の統計的品質の差異に対して何の手立ても講じなければ，不公平が生じる可能性がある。

○　この不公平を排除するために，異なる試験問題セットを受験した各受験者のテスト結果を共通の尺度上の得点で表現し，複数の試験問題セットの受験者間で得点を相互に比較することを可能にする統計的処理を行う。この統計的処理のことを「等化（equating)」という。

○　IRT に基づく試験の得点の表示に当たっては，各設問に配点は設定せず，各設問の項目パラメタを用いた IRT の数式により受験者の能力値を推定し，それを基に得点を算出するという方法を採ることが多い[39]（**【コラム⑩】**)。

【コラム⑩】IRT に基づく試験の得点の算出方法

本コラムでは，各項目に配点は設定せず，受験者の各項目（例えば，共通テストであれば，各大問に含まれる小問の解答枠）への反応パターン（正答したら○，誤答したら×）を基にして，項目の難易度パラメタと識別力パラメタを使って受験者の能力値 θ を算出する場合の得点の算出方法を紹介する。能力値の定め方には様々な方法があるが，ここでは，最尤推定法（maximum likelihood estimation method）と呼ばれる能力値の推定方法を説明する。

例えば，３問の項目から成る試験の正誤反応は，全問正答と全問誤答を除けば[40]，【表コラム⑩－１】に示した S1 から S6 の 6 パターンとなる。例えば，項目 1 にのみ正答し他の項目には誤答した場合は 「S1　○××」で表される反応パターンを得る。そして，それぞれの反応パターンに対応し θ の値が定まる。

【表コラム⑩－１】３問の項目から成る試験の正誤反応のパターン

反応パターン	S1	S2	S3	S4	S5	S6
	○××	×○×	××○	○○×	×○○	○×○

[39] 難易度等のそろった等質な試験問題セットを使用する場合は，様々な条件を考慮した上で，現行の共通テストのように正答した設問の得点を足し上げた点数（素点）により示すことも可能である。

[40] 全問正答・全問誤答時の能力値は理論上無限（-∞や+∞）となり推定できないため，本コラムの説明には含めない。なお，実際の応用場面では，全問正答・全問誤答の受験者に与える能力値をあらかじめ定めておくケースもある。

項目特性曲線を用いれば，能力がθの受験者がこれらの反応パターンを得る確率を計算することができる。すなわち，各項目への正誤反応は独立であることを仮定すれば，例えば，S1 のような反応パターンをする確率は，能力値がθの受験者の項目 1 に正答する確率と，項目 2 に誤答する確率，そして項目 3 に誤答する確率を掛け合わせたものとして得られる。いま，各項目への項目特性曲線の値を P1(θ)，P2(θ)，P3(θ) とすれば，この確率は　P1(θ) × (1- P2(θ)) × (1- P3(θ))　となる[41]。

この確率は，それをθの関数と見た場合に尤度（likelihood）と呼ばれるが，それをθの値が -0.5，0，0.5 の場合に計算したものが【表コラム⑩－２】である。

【表コラム⑩－２】特性値の尤度と最尤推定値

	尤度			最尤推定値
反応パターン	θ=-0.50	θ=0.00	θ=0.50	
S1 ○××	0.23	0.21	0.08	-0.3
S2 ×○×	0.04	0.04	0.01	-0.3
S3 ××○	0.10	0.21	0.18	0.1
S4 ○○×	0.02	0.04	0.03	0.1
S5 ×○○	0.01	0.04	0.08	0.7
S6 ○×○	0.04	0.21	0.41	0.7

これらのθに対応する項目特性曲線の値（正答確率）とそれを 1 から引いた (1-P(θ))の値（不正解の確率）は，【図コラム⑨－１】から読み取り，【表コラム⑩－３】に示した。例えば，「S1 ○××」という反応パターンを得る確率は，θ=-0.50 の時には 0.30 × 0.93 × 0.85 = 0.23 であり，θ=0 の場合には 0.50 × 0.85 × 0.50 = 0.21 となる。

【表コラム⑩－３】θに対応する項目特性曲線の値（正答確率）と (1-P(θ))の値（不正解の確率）

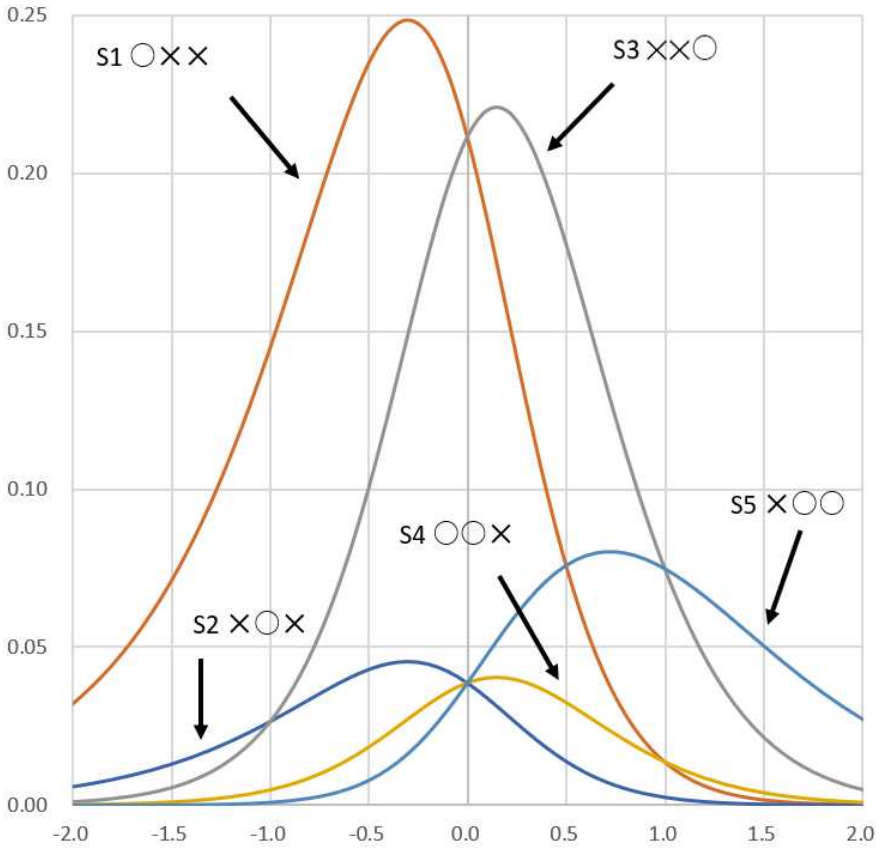
		θ=-0.50	θ=0.00	θ=0.50
P(θ)	問題 1	0.30	0.50	0.70
	問題 2	0.07	0.15	0.30
	問題 3	0.15	0.50	0.85
1-P(θ)	問題 1	0.70	0.50	0.30
	問題 2	0.93	0.85	0.70
	問題 3	0.85	0.50	0.15

これを -2< θ <2 の範囲で全て計算しグラフにしたものが【図コラム⑩－１】の「S1　○××」のグラフであり，この反応パターンに対応するθの尤度関数（likelihood function）と呼ばれ

[41] この，各項目への正誤反応が独立であるという仮定を局所独立性の仮定と呼ぶ（**【コラム⑬】**）。

る。この尤度関数を見ると，θ=-0.3 付近でその値が最大となるが，この，最大の正答確率を与える θ の値を持って，反応パターン「S1　○××」に対応する能力値の推定値とする方法を θ の最尤推定値（maximum likelihood estimate）と呼ぶ。

もう一つの例として「S5　×○○」を考えてみると，能力値が θ の受験者がこの反応パターンを得る確率は，項目 1 に誤答する確率 × 項目 2 に正答する確率 × 項目 3 に正答する確率であるから，例えば，θ=-0.50 の場合には 0.70 × 0.07 × 0.15＝0.01 であり，θ=0 の場合には 0.50 × 0.15 × 0.50＝0.04 である。そして，それを全ての能力値で計算したものが【図コラム⑩－ 1】の「S5　×○○」のグラフであり，その最大値（最尤推定値）はおよそ θ=0.7 付近である。



【図コラム⑩－ 1】尤度関数

【表コラム⑩－ 2】には，θ のおよその最尤推定値を示してあるが，この表から以下のことが分かる。まず，最尤推定値は単純な項目正答数とは一対一に対応していない。例えば，S1，S2，S3 はそれぞれ 1 問だけ正答しているパターンであるが，S3 に対応する θ の推定値の値が大きい。また，S4，S5，S6 に関しても，同じ 2 問だけ正答しているにも関わらず S4 に対応する θ の推定値が低い。また，1 問だけ正答している S3 と 2 問正答している S4 に対応する θ の推定値が同じ値となっている。これは，IRT の 2 パラメタ・ロジスティックモデルにおける θ の推定値が，単純な項目正答数の関数ではなく，識別力パラメタ（a パラメタ）で重みをつけた重み付き正答数の関数として計算できることに由来している。上記の例でいえば，項目 3 の a パラメ

タの値は 2 であるため，S1 から S6 の重み付き正答数は 1, 1, 2, 2, 3, 3 となるが，識別力の高い項目 3 に正答することは，項目 1 と 2 の二つに正答することと同等として取り扱われていることになる。

次に，S1 と S2 を比べてみると難しい項目である項目 2 に正答した S2 に対応する θ の値が大きめに推定されても良いような気がするかもしれないが，この二つの反応パターンに対応する重み付き正答数は同じであるため，θ の値は等しく推定されている。この点，S2 という反応パターンは，易しいはずの項目 1 にも正答できなかった受験者の反応と解釈すれば納得がいくかもしれない。なお，S1 は易しい項目に正答，難しい項目は誤答という反応パターンであり，逆に S2 は難しい項目に正答，易しい項目には誤答という反応パターンであるため，S2 は起こりにくい反応パターンであると考えられるが，そのことは【表コラム⑩－２】に示した S2 の反応確率の低さに反映されている。

また，最尤推定法で与えられるのは能力値 θ の点推定値であるが，必要とあれば，推定の誤差（error）を考慮した，一定の幅をもって推定される区間推定値を与えることもできる。たとえば，S3 という反応パターンに対応する θ の点推定値は 0.1 であるが区間推定としては（-0.1 ≦ θ ≦ 0.3）という様に表現することも可能である。その際の区間の幅は，試験を構成する項目数が多いほど，また各項目の識別力が平均的に高いほど狭くなることが知られている。

なお，たとえば最尤推定法を用いて反応パターンに対応する θ が得られた場合，それをそのままの形で成績として用いると，例えば，θ=0 という結果はあたかも「能力がない」と誤った解釈を受験者にされたり，また，θ が負であるとやはり解釈上誤解や違和感を受験者に与えたりするおそれがある。そこで，例えば，線形変換により，能力値×100+500 と変換すれば，得点は以下の【表コラム⑩－４】のように示される。

【表コラム⑩－４】能力値と得点

能力値 θ	-4	-3	-2	-1	0	1	2	3	4
得点	100	200	300	400	500	600	600	800	900

（執筆：前川眞一（(独)大学入試センター特任教授，東京工業大学名誉教授））

２．IRT に基づく共通テストの実施方法と主な課題

○　仮に共通テストを IRT に基づいて実施すれば，１．で述べたような IRT の特徴により共通テストを以下のような形で実施することが可能となる（第２章２．③で詳述）。

・試験日時の複数回設定が可能
・一人の受験者による複数回受験
・受験者の能力の経時的な変化の把握

○　一方，IRT に基づいて共通テストを実施することを目指す場合，以下に挙げるような論点につ

いて検討が必要である。

(1) 問題作成

○　第1章1．(3)にもあるとおり，現行の共通テストでは6教科 30 科目の試験問題を本試験用，追試験用の2セット作成している。科目ごとに設置されている問題作成部会（約 460 人の大学等の教員）が，年間 50 日程度，大学入試センターのセキュリティエリア内で問題作成の会議を行っている。また，問題点検第一部会（約 160 人の大学等の教員及び高等学校関係者）が，年間 15 日程度，大学入試センターのセキュリティエリア内で試験問題の点検を行っている。

○　IRT に基づいて試験を実施する場合には，基本的に，試験問題を事前の予備調査で，本番の受験者に近い仮想受験者に解答してもらうことが必要になるため，本番の試験で出題される問題が特定されないよう，1科目当たり通常数千～数万問の試験問題を用意する必要がある。作成する問題が大幅に増えるため，現行の共通テストの試験問題と同等の質の維持が求められれば，大量の問題作成を支える十分な数の問題作成者の確保が必要となり，確保できない場合，試験問題の質が落ちる又は問題作成が不可能に陥ることになる。また，問題作成者が多くなればなるほど，問題作成に関する細目(item specification)を緻密に作成し，試験問題に求められる条件を確実に事前に共有することで，誰が問題作成を担当しても，ある程度均質な問題を作成できる体制作りが重要になってくる。

○　その場合，大学等の教員を中心とした現行の問題作成体制とは異なる，大学等の教員以外の人材を参画させる新たな体制を構築することも考えられる。例えば，①アイテム・ライター（問題素案作成を作成し，それを試験問題として適切な形に整える者）を専属で大学入試センターで雇用すること，②問題作成業務を個別に委託すること，などが考えられる。しかしながら，こうした従来と大きく異なる方法を導入することについては，大学が共同して実施する共通テストの在り方として，国民的な議論が必要である。

○　また，予備調査の実施には，本番の受験者に近い仮想受験者を相当数集める必要があるが，広い学力層が受験する共通テストの場合，それに応じた幅広い能力分布をもつ仮想受験者を用意する必要がある。そのような集団を集めることの実現可能性や，予備調査実施時に生じ得る試験問題の漏洩の可能性等についても考慮する必要がある。

○　大量の問題を作成して予備調査を行うには，多額の経費も必要となる。CBT 問題作成ワーキ

ンググループ[42]（以下「問題作成 WG」という。）の報告[43]によると，「情報Ⅰ」を CBT で実施するための試験問題を 5 年間で 1 万問作成する場合，1 年当たり 1 億 125 万円，5 年で総額 5 億 625 万円を要すると試算されている（**【コラム⑪】**）。

○ 大学入試センターの令和 2 年度予算において，PBT で実施する共通テスト 1 科目の問題作成に要する経費は平均すると約 4,100 万円[44]である。したがって，単純計算すると，5 億円をかけて作成した問題を約 12 年間使用し続けることができれば，合計のコストでは現状とほぼ変わらない。しかしながら，IRT の場合には，試験を何回か実施する中で，試験問題の曝露（exposure；試験問題が受験者の目に触れること）や漏洩により項目パラメタが影響を受けてしまうため，試験問題の再利用には限界がある。試験の測定精度を維持するためには，項目パラメタが変動した試験問題を廃棄し，新しい試験問題を問題バンクに追加していくことが必要となることから，実際には，上記問題作成 WG の報告以上の経費が必要になってくると考えられる。

○ 特に，共通テストはハイステークスな使われ方をしている試験であることから，問題漏洩のリスクも高く，一方では，試験の精度を保つ必要性も高い。このため，新たに試験問題を作成して，実際に使用する試験問題を入れ替えていく必要性は，他の試験以上に高いと考えられる。加えて，共通テストは高等学校学習指導要領を踏まえて実施することとされているが，学習指導要領はおおむね 10 年に一度改訂されており，改訂時にはその改訂内容を踏まえた対応も必要となるため，教科・科目の再編状況によっては，多くの問題を入れ替えることが求められる可能性がある。

＜主な課題＞

- 作成すべき問題数が大幅に増える。試験の実施方法や受験者数によっては 1 科目数千～数万問の問題を作成することが求められる。
- 同じ試験問題を本番の試験で何度も利用するという設計の場合，試験問題の曝露や漏洩への対応で頻繁な問題入替え・追加が必要。

＜必要な対応＞

- 十分な数の問題作成者の確保，又は大学等の教員以外の人材を参画させる新たな問題作成の体制の構築が必要。
- 実施経費の増加に伴う財政負担について検討することが必要。

---

[42] CBT 活用検討部会の下に設置された有識者会議。共通テストにおいて教科「情報」の試験を CBT で実施する場合，特に問題バンクを構築して CBT-IRT で実施すると仮定した上で，試行的に問題バンクを構築し，主に問題作成の生産性を中心にそのフィージビリティ（実現可能性）を検証した。

[43] 付録 2「問題バンク構築に係る調査研究について～CBT-IRT での共通テスト「情報」の問題作成に係るフィージビリティの検証～（報告）」。

[44] 大学入試センターの令和 2 年度予算において，PBT で実施する共通テスト（6 教科 30 科目）の試験問題作成に要する経費として約 12.3 億円が計上されており，単純計算で 1 科目当たりの平均試験問題作成経費は 4,100 万円となる。

【コラム⑪】問題バンクの構築・メンテナンス

本文(1)では問題バンクについて触れたが，そもそも問題バンクはどのように構築するのか，また，一度構築した後に，どのようなメンテナンスが必要なのだろうか。

○必要問題数の算出

問題バンクには，一般に数千～数万問の試験問題を蓄積することが求められるとされるが，実際には，各試験において蓄積すべき問題数は，試験をめぐる状況に応じて異なり，ケース・バイ・ケースでの算出が必要である。

問題バンクに蓄積すべき問題数を変動させる要因としては，①一つ当たりの試験問題セットに含める問題数，②作成すべき試験問題セットの数，③出題領域の細分化の度合いなどが挙げられる。また，後述するように，試験問題（案）に対する予備調査を実施する場合，④問題の採択率（予備調査で使用した問題のうち基準を満たしたものの割合）や，⑤問題を何回繰り返して使用するか（どの程度なら繰り返し使用できるか）等についても考慮する必要がある。

○試験問題（案）の作成

問題作成体制を整えた上で大量の試験問題（案）を作成する。なお，次に述べるように，予備調査のデータを基に推定した難易度や識別力の値が統計的な基準を満たしたもののみを問題バンクに登録することになるため，実際には，問題バンクへの登録が必要な試験問題数を超える試験問題（案）を作成する必要がある。

なお，出題領域や下位分類を設定する場合には，試験問題（案）の作成時にメタ情報（出題領域や下位分類を示すラベル）を付与することで，試験問題の効率的な管理が可能になる。

○試験問題（案）の項目パラメタの推定

次に，作成した試験問題（案）の項目パラメタを推定するための予備調査が必要になる。このための方法としては，以下のように，本番の受験者に近い仮想受験者を対象にした「プレテスト[45]」の実施や，本番の試験において新作の試験問題（案）をダミーとして出題するなどがある。

◆ プレテストの実施

プレテストは，本番の受験者に近い仮想受験者に協力を求め，試験問題の項目パラメタをあらかじめ推定する方法である。多数の問題について一定数以上の仮想受験者の解答を依頼することになるため，複数種類の試験問題セットを用いて，効果的な解答データ収集を計画することが必要である。

このプロセスは，いわば，仮想受験者が手分けをして解答を行うイメージとなるが，多数の問題に対してプレテストを行うため，同一の試験問題セットに含まれる問題の組み合わせ方には留意が必要である。例えば，互いにヒントとなりそうな問題ペアがあった場合に

[45] フィールドテスト，試行テストとも呼ばれる。

は，別々の問題セットに入れることが望ましい。反対に，同時に出題すべき問題ペアについては，同一の試験問題セットに含めることが望ましい。なお，この方法を用いる場合，実際の試験が行われる前の段階で，相当数[46]の仮想受験者が問題の内容を知ることになる。もちろん，誓約書等により機密保持を約束させる，書類等の持ち出し等を認めないといった措置を講じることは考えられるが，「記憶による持ち出し」を完全に防ぐことはできないため，試験問題に関する情報が一定程度漏洩する可能性は否定できない。

◆　本番の試験で新作試験問題をダミー試験問題として出題

本番で解答する受験者の得点算出用の試験問題に，「ダミー試験問題」として新作試験問題を混ぜて解答させ，その解答データから難易度を推定するという方法もある。この方法の利点は，仮想受験者ではなく本物の受験者から，試験問題解答へのモチベーションが最も高い状態で，新作試験問題に対する解答データを得られる点である。一方で，従来の日本の試験の多くでは，与えられた試験時間をどのように使うかについても，測定する能力の一環として捉えられてきた部分があることから，仮に，ダミー問題を加える場合には，試験時間の在り方について慎重な検討を要する。

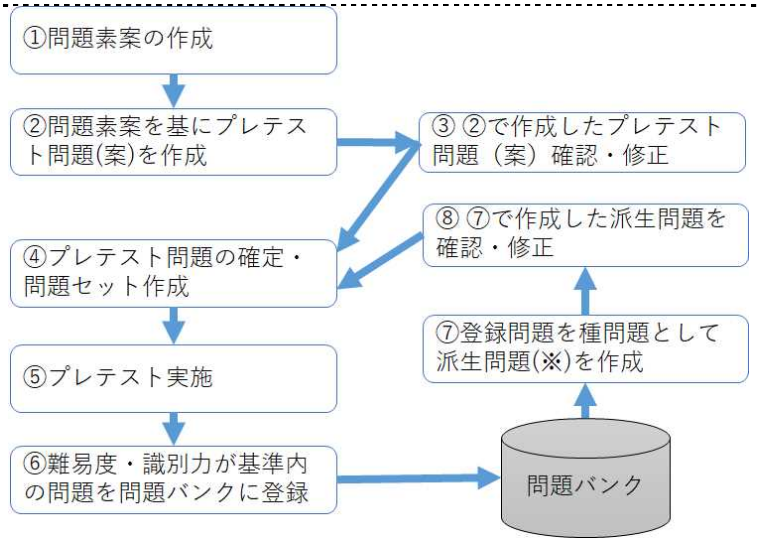
各試験問題（案）の解答データの収集・分析後，項目パラメタの値が条件を満たすもののみ，項目パラメタの値を付与した上で問題バンクに蓄積する。項目パラメタの値が条件に合わない（難易度パラメタが高すぎる又は低すぎる，識別力パラメタが低すぎる等）場合には，該当する試験問題（案）を廃案にする，再修正を施して改めて予備調査を行うなどする。

○問題バンクのメンテナンス

問題バンクは一度構築した後も継続的なメンテナンスが必要である。これは，試験を何回か実施する中で，試験問題の曝露や漏洩によって項目パラメタが変動するためである。試験の測定精度を維持するためには，項目パラメタが変動したと考えられる試験問題を廃棄し，新しい試験問題を問題バンクに追加していくことが必要となる。また，学習指導要領の改訂など，試験で測定しようとする能力や問う内容が変わった場合にも，新しい試験問題と入れ替える必要がある。

問題作成 WG においては，共通テスト教科「情報」の問題バンクを構築する場合の問題作成プロセスや問題作成体制を具体的に想定し，そのプロセスや体制を，小規模ではあるが実際に取り組んでみたところである。

[46] どの IRT モデルを用いるかに応じて異なるが，少なくとも 1 問当たり数百人分の解答が必要となる。



【図コラム⑪－１】問題作成 WG で取り組んだ問題バンク構築の流れ

この取組と同様の方法で５年間かけて１万問の試験問題を問題バンクに蓄積する場合，問題作成や予備調査（ここではプレテストを想定）の実施などのため，【表コラム⑪－１】にあるような人員・経費が必要となる。必要な経費については，１年当たり１億 125 万円，５年間総額で５億 625 万円と試算されている。

【表コラム⑪－１】１万問の試験問題を問題バンクに構築する場合に必要な人員・経費

	必要人数〔年〕	必要経費　〔円／年〕
問題素案作成（①）	50（各作成者が作成する素案数：20 問）[47]	0.5 万×20 問×50 人=500 万
第一部会（②）	20 人（会議日数：50 日）	50 万×50 日＝2,500 万
第二部会（③）	20 人（会議日数：50 日）	50 万×50 日＝2,500 万
問題管理委員会（④）	10 人（会議日数：25 日）	25 万×25 日＝625 万
予備調査（プレテスト）（⑤）	参加者　20,000 人[48]	0.2 万×20,000 名＝4,000 万

※表中の丸数字は，対応する【図コラム⑪－１】のプロセスを指す。

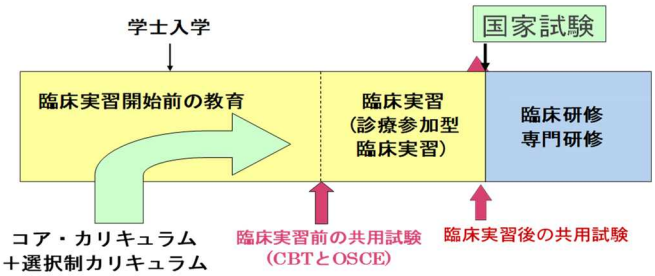
【コラム⑫】医療系大学間共用試験の問題作成について

医療系学部（医学部，歯学部）教育については，その到達目標を示した「モデル・コア・カリキュラム」が国において策定されており，各学部ではこれに基づいてカリキュラムが編成されて

[47] 各作成者が作成できる素案数は，科目の特性や出題形式により異なる。

[48] プレテスト参加者数は，
・1 問当たり延べ 200 人の参加者が必要
・参加者 1 人が一つの試験問題セットに含まれる 25 問を解答
という仮定に基づいて算出した。

いる。医療系大学間共用試験（以下「共用試験」という。）は，臨床実習に参加する学生の知識と技能・態度がモデル・コア・カリキュラムに定める水準に到達していることを確認するための大学間共通の評価システムで，CBT-IRT で実施されている。



【図コラム⑫－１】医師養成における共用試験の位置付け

大学間共通で臨床実習開始前の学生の到達度を確認する強いニーズがあった一方，学修進度や臨床実習開始時期が大学により異なり，到達度を確認する試験を全ての大学で同時期に行うことはできなかったため，異なる時期・場所でも公平に実施・評価できる試験が必要とされた。このため，共用試験のうち知識に関する部分の試験は CBT-IRT で実施することとなったのである。

ハードウェアの使用方法は LAN 方式[49]で，試験実施機関である医療系大学間共用試験実施評価機構（CATO）[50]のサーバに試験問題が配信され，各大学のパソコンやネットワーク（LAN）を使用して実施する。試験で使用するハードウェアは各大学が用意し，各大学は実施環境が適切かを動作確認キットで確認する。受験者数は医学系・歯学系合わせて年間約 11,000 人，受験料は令和２年度時点で 25,000 円である。

共用試験の出題は，CATO で構築する問題バンクから行われる。以下で共用試験の問題バンクの構築やメンテナンスについて紹介する。

○試験問題の概要

受験者は以下の６ブロック（各ブロック 60 分），計 320 設問の試験問題を解答するが，320 問のうち約 80 問がダミー試験問題（解答データを得ることを目的に本番の試験問題に混ぜる新作試験問題で，解答内容は試験の成績には反映されない）[51]である。成績表示は「合格」「不合格」の２種類である。

ブロック１～４：単一多肢選択形式 60 設問
ブロック５　　：多選択肢連問形式 40 設問（鑑別診断，病態等）
ブロック６　　：順次解答４連問形式 40 設問（臨床推論）

---

[49] LAN 方式については，第３章１．(2)で詳述。

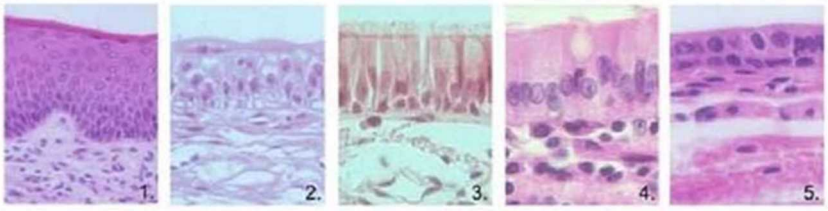
[50] 試験実施業務を支える組織体制の確立や，試験実施用システムの開発，試験問題の蓄積，学生の成績と試験問題のセキュリティ確保，試験監督の派遣等のために，試験実施機関として医療系大学間共用試験実施評価機構（CATO）が設置された。

[51] ダミー試験問題については，**【コラム⑪】**で詳述。

例題８　B-1-(2)-①-1)　上皮組織と腺の構造と機能を説明できる。

膀胱や尿管の内面をおおう上皮はどれか（ 写真はカラー表示 ）。

◎A．1
◎B．2
◎C．3
◎D．4
◎E．5



例題47　E-2-(2)-6）生化学的検査項目を列挙し目的, 適応と異常所見を説明し，結果を解釈できる。

食事の影響を受けやすい血液生化学検査はどれか。

◎A．尿酸
◎B．総蛋白
◎C．ナトリウム
◎D．クレアチニン
◎E．トリグリセリド

例題53　E-3-(5)-③-1）胸部診察で確認すべき項目を列挙し，視診，触診，打診と聴診ができる。

心臓の聴診上、Ⅰ 音と同時に開始し Ⅱ 音まで続く雑音を聴取した。この心雑音はどれか。

◎A．連続性雑音
◎B．収縮中期雑音
◎C．全収縮期雑音
◎D．収縮後期雑音
◎E．拡張早期雑音

【図コラム⑫－２】共用試験（医学系）の公開問題の例

※共用試験のその他のサンプル問題は**参考資料４**を参照

○試験問題（案）の作成

問題バンクを作成するために，平成 14～17 年の 4 年間にわたりトライアルを行った。問題作成依頼は CATO から問題入力システムなどを含む問題作成キットを提供して参加大学の教員に問題作成を依頼した。4 年間で，医学系 30,000 問以上（参加 80 大学），歯学系 25,000 問以上（参加 28 大学）の試験問題が提出された。

大量の問題を集めることができた理由としては，受験者層が臨床実習開始前の医療系学部生に絞られており問題を作成しやすかったこと，教員に共用試験の必要性が理解されていたこと，国家試験の作問経験者が多かったことなどが挙げられる。なお，トライアル期間中は問題作成者に対して謝金等は支払われていない。

○問題バンクの蓄積

トライアル期間中医学系約 28,000 人以上，歯学系約 10,000 人以上の学生が，共用試験システムを用いて試験を実施した。試験結果については CATO の事後評価解析委員会により評価が行われ，４年間で医学系が合計約 12,000 問，歯学系が合計約 8,000 問を問題バンクに蓄積した。

○問題バンクのメンテナンス

難易度に著しい変化が生じた問題，受験者全員がほぼ正解する問題，現在の状況と合わなくな

った問題などを問題バンクから削除する作業を毎年実施している。医学系については，プール問題の管理などを実施する専門部会が 1 回 3 人で年 80 回程度の実施で，問題バンクに蓄積されている問題の評価，新しく問題バンクに入れる問題の評価，事後評価等を行っている。また，不足となったモデル・コア・カリキュラムの領域の問題を新たに募集するなどして，新陳代謝を行っている。

○共用試験の CBT-IRT について

上述のように大変な労力をかけて行われている共用試験が普及してきた一つの理由を示す。平成 26 年から全国医学部長病院長会議が, 共用試験の統一的な合格基準を設定し, 合格者に Student Doctor の証明書を発行している。項目反応理論より求めた IRT 標準スコアの数年にわたる安定性，理論的な妥当性などから，この統一的な合格基準（推奨最低合格ライン）に，IRT 標準スコアが採用された。他の試験との相関が高いこと（各大学における報告等），合格基準周辺の標準誤差が十分に小さいこと（共用試験実施評価機構実施の講演会）が報告されている。また，内容妥当性については，問題作成がモデル・コア・カリキュラムに準拠していることがあり，さらに多くの教員が繰り返し問題のチェックを行っていることから適切な問題が提供されている。このように，十分な質保証を得るためのツールとして IRT 標準スコアが評価されてきている。

（執筆：仁田善雄（公益財団法人医療系大学間共用試験実施評価機構理事））

(2) 試験問題の非公開

○ IRT に基づく試験の得点の信頼性を維持するためには，各試験問題の項目パラメタの変動がないようにする必要がある。しかしながら，試験問題を原則非公開とし，同じ試験問題を本番の試験で何度も利用するという試験設計の場合, 試験問題の曝露や漏洩による試験問題の品質の変動が生じ得る。**【コラム⑪】**のとおり，受験者等が記憶に基づいて問題を持ち出すこと，いわゆる「記憶による持ち出し」は防ぎようがなく，意図的あるいは意図しない形での問題漏洩の可能性は否定できない。既存の IRT に基づく他の試験の事例を踏まえても，試験問題を非公開としていても，「記憶による持ち出し」は発生しており，漏洩を完全に防ぐことはできていないというのが実態である。

○ 試験問題の漏洩に対処するための実効的な方策としては,

・問題バンクに蓄積する問題数を多くする

・問題バンク内の試験問題の品質を常に確認し，曝露や漏洩により品質が変動した疑いのある問題を問題バンクから速やかに引き下げ，新規問題を追加する

などが考えられ，これらを徹底することが求められる。しかしながら，(1)で述べたように，問題の追加や入替は容易ではない上，それらを徹底しても，漏洩による影響を確実にゼロに抑えるこ

とは難しい。

○　また，現行の共通テストの試験問題が試験終了後に公開されていることを踏まえると，同じ試験問題を本番の試験で何度も利用するという試験設計とする場合，一部のサンプル問題は公開するとしても，共通テストの試験問題が原則非公開になることについて，受験者や保護者を含む社会全体の理解を得る必要がある。また，共通テストにおいては，「高等学校教育へのメッセージ性」を重視した問題作成を行っているが，試験問題が非公開になることで，試験問題を教育現場で活用することが困難になることについても留意が必要である。

○　なお，現行の共通テストの結果も，大学によって様々な形で使われている。共通テストの得点で合否が決まるようなハイステークスな使われ方がある一方で，共通テストの得点だけで合否が決まるのではなく，他の資料等も総合した上で判断するようなローステークスな位置付けである場合には，仮に，受験者が試験問題の一部を事前に知っていても，難易度や識別力等の項目パラメタや解答時間の経時的な変化が極めて小さくなる可能性がある。こうした場合であって，試験の実施や結果の利用に支障がないことが確認できれば，試験問題の漏洩による影響を考慮する必要性が低い場合も考えられる。また，そうした試験の場合には，全試験問題をあらかじめ公開することも選択肢としては考えられる。

＜主な課題＞

● 同じ試験問題を本番の試験で何度も利用するという設計とする場合，試験問題が原則非公開になる。

＜必要な対応＞

● 得点の信頼性を維持するためには，試験問題の漏洩を防ぐことが必要（ただし，漏洩の影響をゼロにすることは不可能であり，そのことに対する理解も必要。）。

● （一部のサンプル問題を除いては）試験問題を教育現場で活用できなくなることへの理解が必要。

(3) 出題形式（大問か小問か）

○　現行の共通テストは，従来のセンター試験の出題形式を引き継ぎ，原則として大問形式で出題されている。ここでの大問とは，「一つのストーリーに沿って複数の小問[52]を組み合わせたまとまり」のことを言う。大問形式による出題は，文脈を把握しながら思考する活動に取り組ませることができるなど，教育学的見地からも価値があると指摘されている。一方，大問形式のまま IRT に基づいて共通テストを実施する場合，次のようなことを検討する必要がある。

[52] 小問とは，受験者が解答を行う最小の単位である。

○　まず，大問形式によっても，局所独立性（local independence）の仮定（受験者の能力が同じ場合であれば，一方の試験問題の正誤は他方の試験問題の正誤に影響を及ぼさないという仮定）を満たすことができるかどうかである[53]。ある試験を IRT に基づいて実施する場合，当該試験の問題に対する受験者の解答データが，局所独立性の仮定を満たしていることが求められるため，ある問題が次の問題を正解するための前提になるなど，問題間での強い連動性がないことが必要になってくる（**【コラム⑬】**）。

○　もっとも，大問形式であるからといって，直ちに局所独立性の仮定が失われるわけではなく，受験者のデータを分析した結果も踏まえた総合的な判断が必要である。大問形式を用いたとしても，局所独立性の仮定が十分満たされている場合は，大問内に含まれる各小問について正誤の採点を行い，２パラメタ・ロジスティックモデル等の２値型（正答と誤答の二つのみが存在する）項目反応モデルを適用して，能力値を推定することが考えられる（**【コラム⑨】**）。仮に，局所独立性の仮定が満たされないと判断された場合は，大問内の正答数得点（その大問で何問に正答したか；大問内の部分点のような扱いになる）を用いて，部分採点モデルや段階反応モデル等の多値型（採点結果が二通りより多くなる）項目反応モデルを適用すること[54]なども考えられる。大問形式に対してどのような項目反応モデルを適用するかは，受験者の解答データの統計学的特徴，試験の目的や得点の利用方法等に応じて異なり得るだけでなく，分析者によっても判断が分かれ得るところである。

○　また，問題バンクを用いる場合には問題を大量に作成する必要があり，大問形式の問題は小問形式に比べて大量作成が困難であると予想されるため，どのように問題作成を行うかについて検討が必要である。

○　さらに，大問形式の試験問題は小問形式の試験問題に比べて印象に残りやすく，「記憶による

---

[53] 本文では主に局所独立性の仮定に関して説明したが，局所独立性の仮定と次に述べる一次元性の仮定は密接に関連している。大問形式のまま IRT を導入する場合，一次元性（unidimensionality）の仮定を満たすかについても，より丁寧な検討が必要となる。一次元性の仮定とは，一つの教科・科目の試験に含まれる問題の正誤データの背後に一つの能力次元が想定できるとする仮定のことである。一次元性の仮定は，本来 IRT だけに限定されたものではない（例えば，現行の共通テストにおいても各設問の配点を足し上げて得点を求めることが多く行われるが，一本の数直線上に布置される合計得点を使用するということは，暗黙のうちに設問間の一次元性を仮定していると考えられる。）。しかし，大問形式のまま IRT を導入した場合，試験全体の出来不出来を左右する能力次元に加え，各大問で扱われるテーマや測定内容・領域などに起因して大問内の項目の出来不出来を左右する個別の次元が見出される可能性があるため，一次元性の仮定を満たすのかについてより丁寧な検討が求められるのである。

[54] 例えば，ある大問が四つの小問（配点各１点）から構成されている場合，この大問の得点は，全問誤答の０点から全問正答の４点までの５段階の採点結果をもつことになる。このような大問を２値項目と捉えてしまうと，全問誤答と全問正答の二通りの採点結果しか想定しないことになり，適切でない。

持ち出し」が，より容易であると考えられる。しかし，(2)にあるとおり，同じ試験問題を本番の試験で何度も利用するという設計とする場合，試験問題の非公開を前提に実施する必要があるため，「記憶による持ち出し」が容易な形式での出題は，IRT にはなじまない可能性があることにも留意が必要である。

○　以上のように，共通テストを IRT に基づいて実施する場合，従来の出題の在り方にも影響を与える可能性がある。どのような出題形式とするのが適切かについては，様々な見地から十分に検討することが求められる。

＜主な課題＞

● 現行の共通テストのように大問形式で出題する場合，以下の課題がある。
　・問題の大量作成が困難。
　・「記憶による持ち出し」が小問形式に比べて容易。

＜必要な対応＞

● どのような出題形式とするのが適切かについて，様々な見地から十分に検討することが必要。

【コラム⑬】局所独立性の仮定，一次元性の仮定

本文(3)及び脚注 53 で言及した局所独立性の仮定及び一次元性の仮定について説明する。

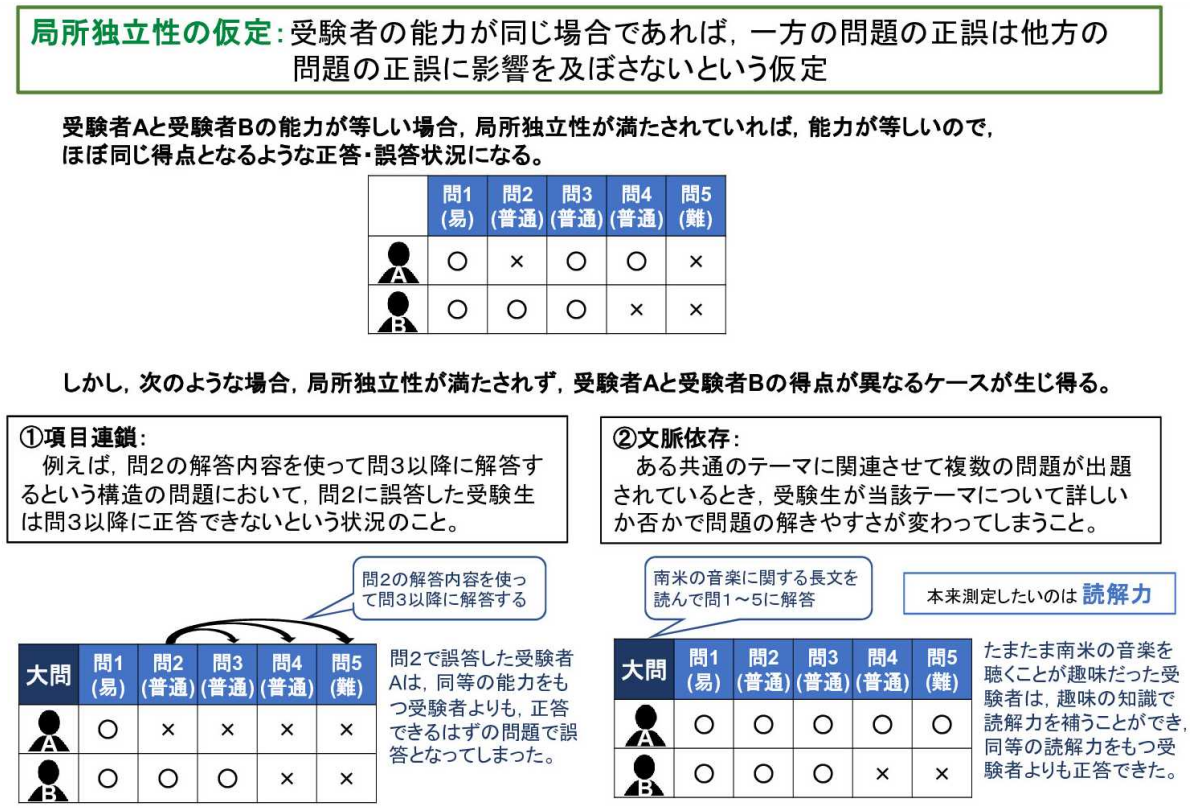
○局所独立性の仮定

局所独立性の仮定とは，受験者の能力が同じ場合であれば，一方の試験問題の正誤は他方の試験問題の正誤に影響を及ぼさないという仮定である。局所独立の「局所」とは，「受験者の能力が同じであれば」という意味である。また「独立」とは，解答結果に関連性がないという意味である。つまり，一次元として取り出した能力の高低のみが，試験問題間の正誤の間に結び付き（相関・連関）を生じさせるが，それ以外の要因は試験問題間の正誤に相関や連関を生じさせないことを意味する。IRT を使う場合には，局所独立性を仮定する必要があることが一般的である。

局所独立性は，①前の試験問題の答えを使って後の試験問題に解答するような形式を伴う問題がある場合（項目連鎖（item chaining)），②提示された文章内の特定のテーマなど，必ずしも当該の能力の測定とは直接関係がないとされる素材（問題に含まれている図表，文章など）が複数の試験問題間で共有されている場合（文脈依存（context dependence)），③測定を意図した能力とは別の能力次元が試験問題の出来不出来を反映してしまっている場合，に阻害される。

局所独立性の阻害が想定される場合に採られる対応は二つある。一つは，問題作成の段階でこういった構造となることを避ける対応が挙げられる。項目連鎖が生じないような問題や，ある程度一般的に受験者が知っているであろうテーマを題材とした問題を作成することになる。もう一つは，項目連鎖や文脈依存が生じている複数の試験問題を一つの大問（基本単位）と見なして，

多値型の IRT モデルを適用する対応である。



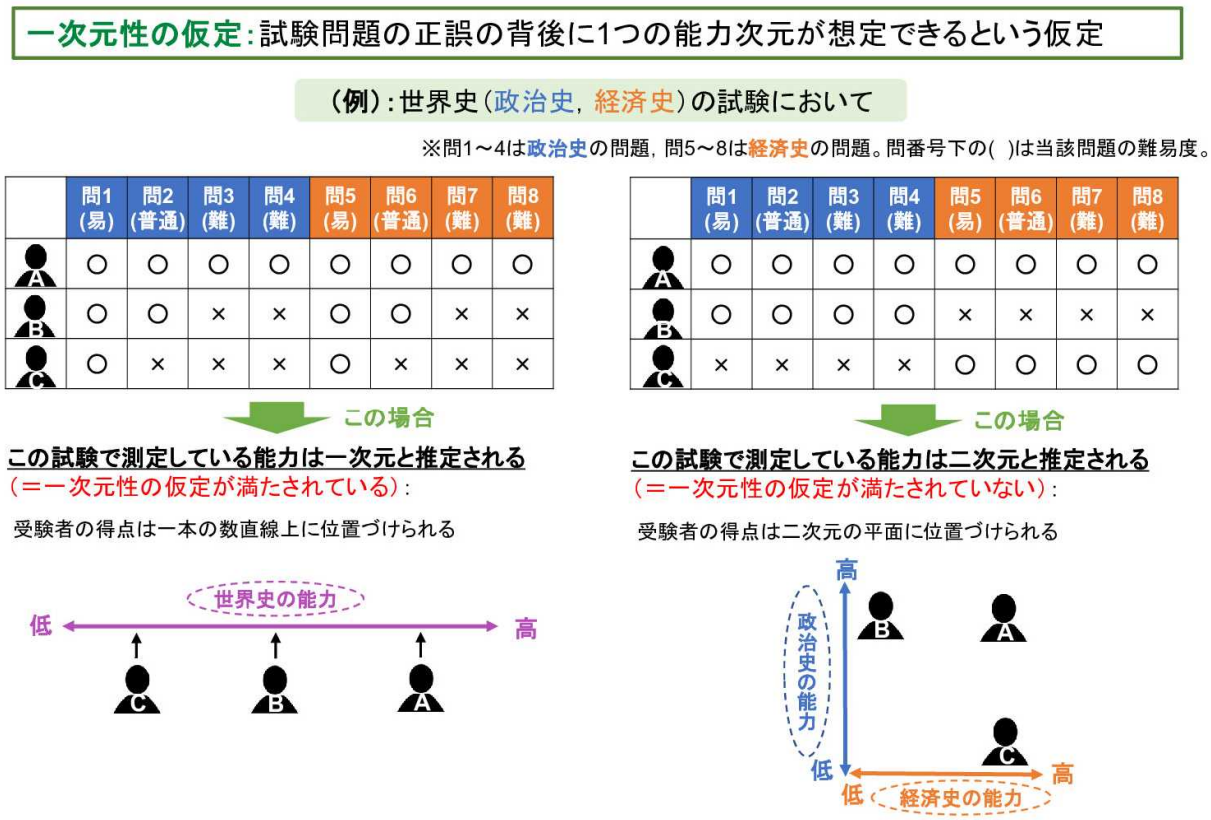
【図コラム⑬－１】局所独立性の仮定について

○一次元性の仮定

一次元性の仮定とは，一つの教科・科目の試験に含まれる問題の正誤データの背後に一つの能力次元が想定できるとする仮定のことである。すなわち，受験者 1 人につき測定される能力推定値は 1 種類に限定することを意味している。大規模試験において IRT を適用している事例の多くは，この一次元性を仮定している。

IRT モデルの中には多次元性に対応したものもある[55]が，選抜を目的とする試験の場合は，一つの試験で複数次元にまたがると運用が煩雑になること等から，一次元に縮約して得点を表示することが多い。

[55] 二つ以上の能力について考えるときには，補償型（一方の能力が低くても，他方の能力が補って正答できる構造）か非補償型（どちらも高くないと正答できない構造）か等，適切な IRT のモデルについて綿密な検討が必要となる。また，複数の下位能力が想定されていても，解答データを説明できる（より大枠な意味合いをもつ）単一次元の能力にまとめることが適切な場合もあるし，その逆も考えられる。



【図コラム⑬－２】一次元性の仮定について

(4) 成績の表示方法

○　現行の共通テストでは，正答した設問に割り振られた配点を足し上げた「素点」(raw score)が受験者の得点となり，大学入学者選抜で使用されている。このため，マークの誤り等がない限り，採点結果と自己採点は一致することになっている。

○　しかし，IRT に基づく試験の得点の表示に当たっては，一般に，１．(2)で述べたように，各試験問題の項目パラメタを用いた IRT の数式により受験者の能力値を推定し，それを基に得点を算出するという方法を採ることになる（**【コラム⑩】**）。

○　現行の共通テストや他の日本の試験の多くが素点に基づいて行われていることを踏まえると，共通テストの成績を IRT の数式により算出した得点を用いて表示することとした場合，自己採点結果と得点の関係が分かりにくく，例えば，従来のように受験者が自己採点により自身の成績を予測することが困難であったり，成績開示後に受験者が自分の得点に疑義を抱くという事案が生じたりする可能性がある。このため，受験者や保護者，高等学校関係者等への事前の周知や，成績に疑義が生じた場合の対応方法の検討などを行い，新しい成績の表示方法について受験者や保護者を含む社会全体の理解を十分に得ることが必要である。なお，IRT に基づく場合でも素点

で成績を表示できるような設計とする[56]ことも可能であり，こうした選択肢を視野に入れることも考えられる。

＜主な課題＞

● IRT の数式に基づいて算出した得点により成績を表示することが多いが，その場合，得点は現行の共通テストで用いられている「素点」とは異なる表示になる。

＜必要な対応＞

● 「素点」とは異なる表示とすることとした場合，受験者自身が自己採点結果と得点の関係を理解するのが困難になるため，新しい成績の表示方法について受験者や保護者を含む社会全体の理解を十分に得ることが必要。

【コラム⑭】日本的試験文化

近年では大学入学者選抜の在り方も多様化が進んでいるが，それでも，我が国における試験文化の形成に，共通１次と，それに続くセンター試験という大規模共通試験が果たしてきた役割は大きいだろう。その特徴としては，以下の点が指摘できる。

① 年に一度，同一の新作問題を用いて同一時刻に一斉で行われる。

・全ての試験場で同一の試験問題が出題される。

・試験問題について本番前の予備調査は行わない。

② 試験問題は直後に公開される。

・過去に出題された試験問題は教育・受験勉強に利用される。

③ 多肢選択式の大問形式が多く出題される傾向がある。

・小問形式より大問形式の方が高度の認知能力を測れるとされる。

・各試験問題がどの分野のどのような能力を測っているかが明確ではない。

④ 成績表示には，正答した設問の配点を足し上げた点数である「素点」が利用される。

・選抜においては素点が重視される。

このような特徴は，「日本的試験文化」として，大学入試の一つのスタンダードとして社会的な定着を見せている。すなわち，公平・公正な選抜を行うためには，上記の特徴を備えた方式で実施すべきという考え方が浸透しているといえるのではないだろうか。

一方，大学入学者選抜を IRT に基づいて行う場合，例えば，複数の試験問題セット，複数の試験日，試験問題の非公開等，こうした伝統的な日本的試験文化と相容れない部分が多い。

[56] １．(2)でも言及したとおり，難易度等のそろった等質な試験問題セットを使用する場合は，様々な条件を考慮した上で，現行の共通テストのように素点により成績を表示することも可能である。

新しい試験を導入する場合には，受験者や保護者を含む社会全体の理解は不可欠である。そのためにも，共通テストに IRT を導入する場合は，そのメリット及びデメリットについて丁寧に周知していくことが必要である。

<参考文献>

石井秀宗（2018）「大学入試における共通テストの複数回実施は実現可能か －日本のテスト文化やこれまで見送られてきた理由などからの検討－」，名古屋高等教育研究 第 18 号，23-38.

前川眞一（2014）試験の日本的風土 *大学入試センターシンポジウム 2014 大学入試の日本的風土は変えられるか*（URL: https://www.dnc.ac.jp/news/20141203-01.html）

3．IRT に基づく共通テストにより実現できること

○ 2．では，共通テストを IRT に基づいて実施する際にどのような課題があるかについて検討した結果をまとめたが，その中では，実現に向けて克服すべき課題も多いことが明らかになった。本節では，仮にそれらの課題が克服されて，共通テストを IRT に基づいて実施することとなった際に，どのようなことが実現可能になるかを整理する。

(1) 試験の年複数回実施

○ IRT に基づいて共通テストを実施すれば，異なる試験問題に解答した者同士の得点も比較できるようになる。そのため，現行では原則として年 1 回のみの試験であるところ，年に複数回実施することが可能となる[57]。これにより，年 1 回の同一時刻一斉実施に付随する，病気や事故等の事情により受験できないといったリスク，今般の新型コロナウイルス感染症などの流行性疾患の感染拡大や大規模な自然災害発生により実施不可能になるといったリスクを一定程度解消できる。

○ ただし，共通テストを年に複数回実施することとした場合，どの時期に試験を実施するかが問題となる。現行の共通テストの実施時期より早い時期，すなわち 1 月よりも前に共通テストを実施することについては，高等学校の授業の進め方や学校行事，課外活動等にも大きな影響を与える可能性がある。また，分割実施とする場合，試験日が早いか遅いかによる不公平（感）が生じたり，既に受験し終わった受験者と未受験の受験者が混在する時期に受験者の心理状況等に影響

[57] また，第 3 章 1．(9)で言及したように，受験環境の整備やトラブルへの対応を考慮すると，共通テストに CBT を導入する場合，同一時刻一斉実施ではなく分割実施（試験日時を複数設定）の方が実施しやすいことにも留意が必要である。

が生じたりすることなどが考えられる。このため，どのような時期・期間で試験を実施するかについては，高等学校教育への影響などを含めて慎重に検討していく必要がある。

(2) 一人の受験者による複数回受験

○　共通テストを IRT に基づいて実施し，試験日を複数設定することとした場合，一人の受験者が複数回受ける「複数回受験」を認める制度設計も可能となる。いわゆる「一発勝負」の試験の場合には，病気や事故等の事情によって受験できなくなる（あるいは，受験はできても，試験問題の内容や当日の体調等に影響される）というリスクがあるところ，複数回受験が可能になることで，受験者にとっても，また大学にとっても，より実力が反映された結果に基づいた選抜を行うことができると考えられる。

○　しかし，複数回受験について検討する際には，民間の英語資格・検定試験の活用に関しても議論のあった，経済格差や地域間格差の問題を避けて通ることはできない。仮に，１回受験する度に検定料を徴収する仕組みとした場合，当該受験者が置かれた経済的背景が受験できる回数に影響するおそれがある。また，試験場が近くに設置されている受験者にとっては複数回受験する際の移動・宿泊に係る物理的・経済的負担が小さい一方，離島・へき地在住の受験者をはじめ居住地の近くに試験場がない受験者にとってはそのような負担が大きいという，地域間格差が生じる可能性も想定される。一人の受験者による複数回受験について受験者や保護者を含む社会全体の理解を得るためには，このような格差の問題について検討する必要がある。

○　また，複数回受験を認める場合，PBT でも起こり得ることだが，テストの点数を上げるための小手先のテクニック（test-wiseness）や何度も受験をして試験慣れをしたことにより高得点を取るケース，繰り返し受験する中でたまたま高得点を取るケースなどが生じることも想定される。

○　さらに，複数回受験を認める場合には，各試験場が受け入れる必要のある「延べ受験者数」自体が大幅に増加することが想定されるため，それに対応できるよう試験日や座席数を確保する必要が生じることにも留意が必要である。

【コラム⑮】複数回受験を認める場合に生じる公平性の問題への対応

現行の共通テストの試験場は，原則として都道府県単位で大学が設定し，大学入試センターが志願者の分布や使用施設の収容数を考慮し指定している。現状では，約 700 の試験場のうち約 10 試験場を離島に設定しているが，それでもなお，試験場への移動に大きな負担を伴う受験者も少なくない。

共通テストで複数回受験を認める場合，試験場へのアクセスが容易な受験者と困難な受験者との間での公平性の問題が，これまで以上に拡大することが懸念される。特に，テストセンターを試験場として複数回受験を認める試験を実施する場合，その多くが都市部に立地していることに

留意する必要がある。

この問題を解消するためには，例えば，以下のような方策をとることが考えられる。なお，自宅・高校など受験者が自身で選択した場所で CBT 受験するという方策も考えられるが，本人確認・不正防止をどのように行うかが別途課題となる[58]ため，ここでの検討からは除外する。

○試験場を離島・へき地を含む地方部にも設置する

離島・へき地を含む地方部に試験場を設置するためには，高等学校などを活用することが考えられる。そのためには，

・パソコンの調達（輸送による持ち込み又は現地のパソコンを活用）

・サーバ，ネットワークの整備及び環境確認

・その他の必要機器の整備

が必要となる。また，試験場に配置するスタッフを確保する必要があるが，PBT と比較すると，機器の設置や操作への習熟など，スタッフにはより多くのことが求められると考えられる。

○受験可能な回数に上限を設ける

複数回受験を認める場合でも，一人の受験者が受験可能な回数に上限を設けることで，地域間格差の拡大を抑えられる。なお，無制限に受験を認めると，経済的格差の問題や，試験場の座席数の不足の問題を引き起こす可能性があるが，受験可能な回数に上限を設けることでこれらの問題にも対処できると考えられる。

一方，受験可能な回数に制限を設けることは，目標の得点に届くまで何度も受験したい，模試代わりに受験したいなど多くの回数を受けることを希望する受験者から受験する機会を奪うことにもなりかねないことから，慎重な判断が必要である。

(3) 受験者の能力の経時的な変化の把握

○　センター試験や現行の共通テストでは，異なる年度の試験結果を比較することは不可能だった。しかし，IRT に基づいて試験を実施することで，中長期的に試験結果を比較できるようになる。仮に，同一受験者が複数年にわたって受験した試験の結果から，当該受験者の能力推移を経年で比較することができるようになれば，例えば，ある受験者の大学入学前と大学入学後の能力を比較できるようになるなど，試験の結果をこれまで以上に様々な形で利用できるようになる可能性がある。また，異なる年度の受験者集団の能力推移の経年比較も可能になることから，さらに，高等学校・大学の教育改善や教育政策の検証・改善につなげることも考えられる。

[58] 自宅・高等学校など受験者が自身で選択した場所で CBT 受験する場合の本人確認・不正防止対策については，**【コラム⑤】**で詳述。

【コラム⑯】PISA 調査における得点の考え方～経年比較の観点～

OECD（経済協力開発機構）は，義務教育修了段階の 15 歳の生徒を対象に，読解力，数学的リテラシー，科学的リテラシーの三分野における学習到達度を国際的に調査する「生徒の学習到達度調査（Programme for International Student Assessment）」（以下「PISA 調査」という。）を実施している。PISA 調査は IRT に基づいて実施されている。2000 年調査から 2012 年調査までは PBT で実施されていたが，2015 年調査から CBT に移行し，直近の 2018 年調査においては，三分野のうち読解力についてアダプティブ方式[59]で実施された。

PISA の主な調査目的は，その国の教育制度の長所や短所を明らかにし，政策立案に資する基礎的データを提供することにある。このため，PISA 調査の結果は，生徒一人一人についてではなく，参加国・地域の生徒全体の平均得点により示される。IRT に基づいていることから，この調査結果は参加国･地域間で比較できる。【表コラム⑯－１】は，2018 年調査における 37 の OECD 加盟国における三分野の平均得点を，平均得点の高い順に上から並べたものである。

【表コラム⑯－１】PISA2018 の結果（OECD 加盟国（37 か国）における比較）

OECD加盟国(37か国)における比較　　は日本の平均得点と統計的な有意差がない国

	読解力	平均得点	数学的リテラシー	平均得点	科学的リテラシー	平均得点
1	エストニア	523	日本	527	エストニア	530
2	カナダ	520	韓国	526	日本	529
3	フィンランド	520	エストニア	523	フィンランド	522
4	アイルランド	518	オランダ	519	韓国	519
5	韓国	514	ポーランド	516	カナダ	518
6	ポーランド	512	スイス	515	ポーランド	511
7	スウェーデン	506	カナダ	512	ニュージーランド	508
8	ニュージーランド	506	デンマーク	509	スロベニア	507
9	アメリカ	505	スロベニア	509	イギリス	505
10	イギリス	504	ベルギー	508	オランダ	503
11	日本	504	フィンランド	507	ドイツ	503
12	オーストラリア	503	スウェーデン	502	オーストラリア	503
13	デンマーク	501	イギリス	502	アメリカ	502
14	ノルウェー	499	ノルウェー	501	スウェーデン	499
15	ドイツ	498	ドイツ	500	ベルギー	499
16	スロベニア	495	アイルランド	500	チェコ	497
17	ベルギー	493	チェコ	499	アイルランド	496
18	フランス	493	オーストリア	499	スイス	495
19	ポルトガル	492	ラトビア	496	フランス	493
20	チェコ	490	フランス	495	デンマーク	493
	ＯＥＣＤ平均	487	ＯＥＣＤ平均	489	ＯＥＣＤ平均	489
	信頼区間※(日本):499-509		信頼区間(日本):522-532		信頼区間(日本):524-534	

※信頼区間は調査対象者となる生徒全員(母集団)の平均値が存在すると考えられる得点の幅を表す。
PISA調査は標本調査であるため、一定の幅をもって平均値を考える必要がある。

また，PISA 調査は調査年ごとに対象集団が違うが，継続して出題している問題（アンカー問題（anchor test））の正誤等の解答情報を活用することで，各国・地域の平均得点を経年比較できるようになっている[60]。こうしたアンカー問題の正誤等の解答情報を手掛かりとして，異なる調

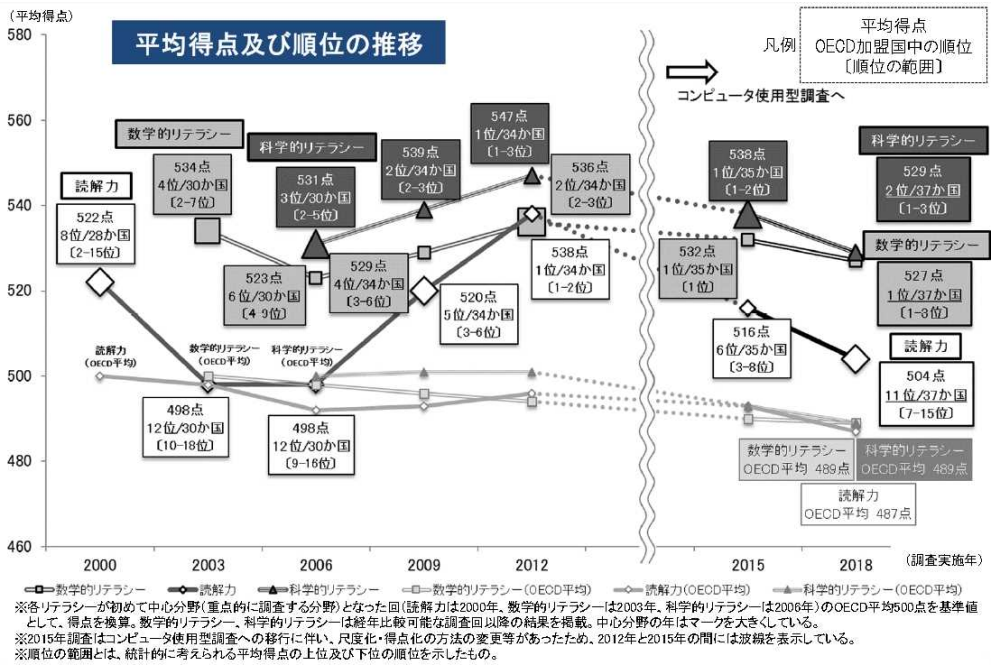
[59] アダプティブ方式については１．(1)で詳述。

[60] PISA 調査で出題される問題は，大きくアンカー問題と新規問題とに分けられる。例えば，2018 年調査の読解力では，全 245 小問のうち 72 小問がアンカー問題である。72 小問のうち，2000 年調査から出題されているアンカー問題が 28 小問，2009 年調査から出題されているアンカー問題が 44 小問である。

査年における比較をより正確なものにしている。

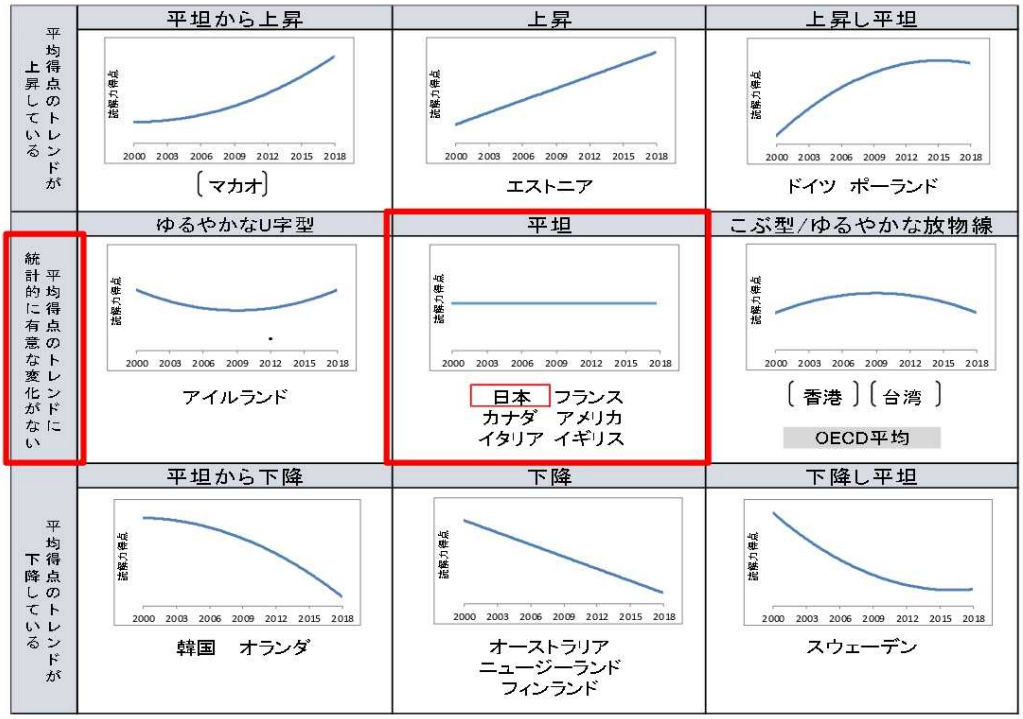
平均得点の表示も，各分野で得点の基準年を設定し[61]，その後の調査の平均得点を基準年のものに合わせることで，経年での比較が可能なようになっている。例えば，読解力は，最初に中心分野になった 2000 年調査時の OECD 加盟国の平均得点を 500 点，標準偏差 100 点となるように測定単位となる得点の基準が定められた。2003 年以降の調査においては，その基準と比較できるように統計処理し，得点化している。

実際に経年比較を行った資料を紹介する。2000 年に始まった PISA 調査は 3 年ごとに実施され，2018 年調査で 7 回の調査データが蓄積されているが，【図コラム⑯－１】は 2000 年から 2018 年までの日本の平均得点と順位の推移を示したものである。また，OECD は，こうしたデータの蓄積により中長期的な変化の傾向も公表するようになってきている。【図コラム⑯－２】は，読解力について 2000 年から 2018 年の 7 時点の各国の平均得点に統計モデルをあてはめ，九つのパターンに分類したものである。



【図コラム⑯－１】PISA 調査　日本の平均得点及び順位の推移

[61] 三分野とも，最初に中心分野として調査が実施された年が基準年として設定されている（読解力：2000 年，数学的リテラシー：2003 年，科学的リテラシー：2006 年）。



【図コラム⑯－２】PISA 調査　各国・地域の平均得点の長期トレンド（読解力）

ただ，PISA 調査の得点の経年変化を見る際の技術的な課題もある。例えば，PISA 調査は IRT に基づいて実施されているが，CBT が導入された 2015 年より前に開発された問題は 1 パラメタ・ロジスティックモデル，2015 年以降に新規開発された問題は 2 パラメタ・ロジスティックモデルに基づいて設計されている。このように問題によって基づく IRT のモデルが異なることにより，統計処理を施したとしても，平均得点を経年比較する際に精度が低下する可能性があると指摘されている。PISA 調査の平均得点の経年比較に当たっては，このような調査の設計や実施方法の変化にも留意する必要がある。

＜出典＞

文部科学省・国立教育政策研究所「OECD 生徒の学習到達度調査 2018 年調査（PISA2018）のポイント」（令和元年 12 月 3 日）

（執筆：大塚尚子（国立教育政策研究所国際研究・協力部総括研究官））