連載

心理社会的要因の測定⑹
「まとめ」

産業医科大学産業実務研修センター　**堤　　明純**

連載「心理社会的要因の測定」の最終回は，本稿で述べてきた尺度構成方法が準拠している古典的テスト理論の前提と限界に対比させながら，新しいテスト理論として尺度構成にその応用がなされている項目反応理論 Item Response Theory（IRT）について紹介し，心理社会的要因を測定する際の留意点についてまとめたいと思っている。

1.　古典的テスト理論の前提と限界

本連載では，心理社会的要因の測定について，主に古典的テスト理論 Classical Test Theory（CTT）に基づいて解説してきた。連載の中で何度か言及したが，古典的テスト理論で扱う尺度の特性は，その尺度を適用した対象のみに規定される。同じ構成概念を測定しようとするとき，違う対象・時期での比較は慎重であるべきだし，また，測定の精度も異なることに留意しなければならない。測定対象毎に測定の信頼性を確認すべき所以である[1)]。

1）尺度あるいは尺度項目の統計量はある特定の母集団に対して定義されている。

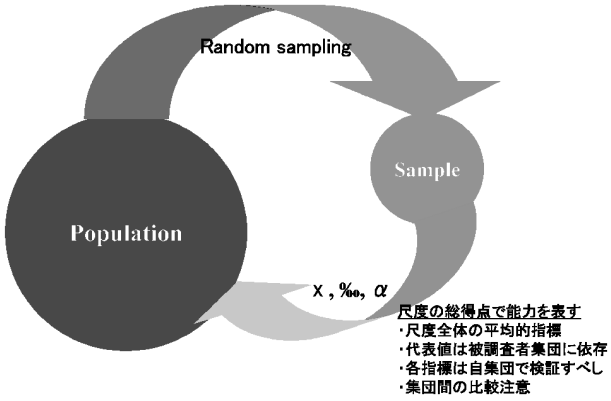
古典的テスト理論では，調査対象における正答率や項目間相関をもとに代表値や信頼性を推定し，標準化を行なっている。尺度の特性として表されるのは，尺度全体の平均的指標であり，尺度の特性のみならず，被調査者集団の特性の分布に依存している[2)]。

測定を行おうとする対象の母集団の特徴は，その母集団を代表すると考えられる標本における観測データを基に推測され，用いられた標本の統計量が尺度および尺度項目の特徴を決定づけることになる（図1）。その前提は，したがって，測定の対象とする集団を代表するサンプルにおける測定がなされているところにあるが，この前提は意外と認識されていない。

2）非平行テスト non-parallel test によって測定された異なる集団の特性を比較することはできない。

同一の構成概念を測定する尺度でも，異なる項目を有する尺度を用いて異なる集団，もしくは個人，の特性を比較することは許されない。読者にとって，このようなことは自明で想定したことはないかもしれない。しかし，後述するように，項目反応理論を応用することによって，異なる項目を用いて測定した得点を比較することが容易に可能となる。

図1　古典的テスト理論の前提
標本における正答率や項目間相関をもとに母集団の代表値や信頼性を推定



2.　項目反応理論の概要

潜在特性尺度値 θ（能力や熟達度と解釈できる）の高い個人ほどその項目に正答する確率が高くなる。言い換えると，テスト項目への正答確率が被調査者の潜在特性 θ の関数として表すことができる（図2）。この関係は正規累積分布に従うことが判明しており，被調査者集団（の得点分布）に依存せず，確率的に定まる。この関数を項目特性曲線 Item Characteristic Curve（ICC）または項目反応関数 Item Response Function（IRF）と呼ぶ。

上記関数はロジット変換を行うと数学的取り扱いが容易となり，識別度（a），困難度（b）といった項目のパラメータのみで曲線の形状を決定することができる（このように，ふたつのパラメータで項目の特性をすべて記述されるモデルを2パラメータ・ロジスティック・モデルという）（図3）。項目反応理

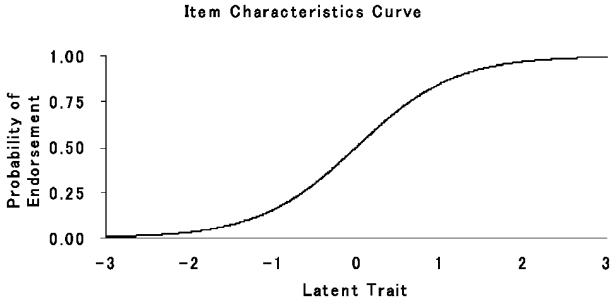
表１　古典的テスト理論と項目反応理論

	古典的テスト理論	項目反応理論
前提	観測得点が，真の得点と誤差の和であらわされる。 X＝T＋E 多数回繰り返し測定をした時の，個人の誤差の期待値は 0（1 回の測定について，多数の受験者について誤差の平均は 0)。 1 回の測定について，多数の被調査者について真の得点と誤差との相関は 0。	被調査者の，ある項目に対する反応は，他のいずれの項目に対する反応とも独立に生ずる（局所独立 local independence)。 測定しようとする潜在特性は一次元。 各項目の項目特性曲線の適合度（項目の特性を示すパラメータの数により，困難度だけで規定される Rash モデル，識別度，困難度，および当て推量パラメータを加えた 3 パラメータモデルなどがある)。
代表値	尺度得点は順序尺度水準	特性尺度値は間隔尺度水準
各項目の特性	以下の特性が対象毎にきまる。 難易度：当該被調査者集団における通過率（正答者の比率） 識別度：当該被調査者集団における点双列相関係数（項目得点とテスト得点との相関係数）	項目特性曲線のパラメータで表される。2 パラメータモデルでは，以下の二つの特性が，対象に依存せずに定まる 識別力：曲線の立ち上がりの強さ（傾き） 困難度：横軸上における曲線の位置（変曲点）
尺度の測定精度	当該被調査者集団における真の得点の分散と観測得点の分散との比（推定値：信頼性係数）で表す。 尺度全体としての精度を表すため，その尺度の被調査者に対する平均的な精度を示し，特定の個人についての測定精度は評価できない。	テスト情報量が特性尺度値 θ の関数として与えられ，以下の特徴がある： 被調査者集団から独立した尺度固有の性質として測定精度を表現できる。 異なる尺度値をもつ個人ごとにその尺度による測定の精度を評価できる。
尺度得点の解釈基準	対象の母集団が想定されている。 準拠集団（尺度が測定対象とする集団と同質の標本集団）における尺度得点の分布の中の相対的な位置に基づいて設定される。 平均的な値，一般的に代表値はすべての項目に回答することで算出される。 正答項目数を個人の得点として用いるため異なる項目に解答した受験者間の測定結果を比較することは不可能。	個人能力を潜在特性尺度上の値で表す。この値は当該被調査者の項目に対する正誤反応パタンから推定する。 実際に観測された項目反応パタンが得られる確率が最も大きい θ を推定する方法（最尤推定法）が用いられることが多い。 適応型テストのように解答する項目が被調査者間で異なる場合でも，同一特性尺度上の値で測定結果を表示することが可能。
標準化	項目を入れ替えると，基本的には標準化の手続きをやり直す必要がある。	尺度を構成する項目の一部を新しいものと入れ替えても，特性尺度値をもとに解釈基準が設定されるので尺度の標準化をやり直す必要がない。

図２　特性（能力）θ と項目正答確率の関係

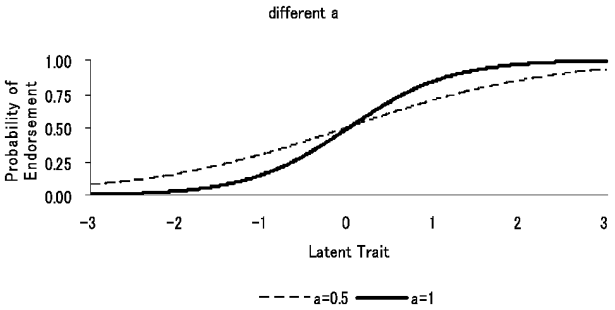
能力（θ）が高いほど，その能力を測定する項目に正答する確率が高くなる。
⇒この関係は正規累積分布に従う。すなわち，被調査者集団（の得点分布）に依存せず，確率的に定まる。
θ：直接測定できない潜在特性（項目を変数とした因子分析の第一因子）

図３　項目特性曲線
各項目の特性が項目特性曲線のパラメータで表される
$P_j(\theta)=\{1+\exp[-1.7a_j(\theta-b_j)]\}^{-1}$



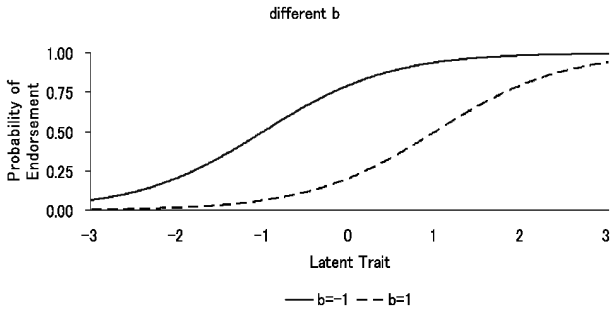
a_j：識別度パラメータ；曲線の立ち上がりの強さ（傾き）
b_j：困難度パラメータ；横軸上における曲線の位置（変曲点）

図4　識別度パラメータ



項目反応曲線の傾きが大きい項目では，高い能力を有する者と低い能力しか有しない者で正答率がかなり異なってくるため，能力の高低をよく識別できる。

図5　困難度パラメータ



曲線が右にシフトしている項目ほど正答が難しい。これくらいの能力の人は何％の確率でこの問題が解ける，という推定ができる。

論では，識別度や困難度といった項目の特性と，態度や能力といった被調査者の特性を決定するパラメータが確率的に推定されるという特徴を有している。すなわち，古典的テスト理論の通過率および点双列相関係数のように指標の値を求める集団が異なった場合に異なる値が得られるようなことがなく，被調査者集団によらず項目パラメータが定まることになる（項目パラメータの不変性 parameter invariance）。

2パラメータ・ロジスティック・モデルで定まるパラメータの特性を少し見てみよう。

識別度パラメータは，曲線の立ち上がりの強さ（傾き）をあらわす。識別度パラメータが大きい（傾きが大きい）項目は，能力の差をつけやすく，能力の高低をよく識別できる（曲線が寝ていると，差がつかない）（図4）。

一方，困難度パラメータは曲線の変曲点の位置であらわされる。曲線が右にシフトするほど，その項目は難しい（能力が高くなければ，その項目に正答できない）。ICCは，項目の特性と被調査者の特性を表現しているので，項目毎にどれくらいの能力の人がその項目に回答できるのか推定が可能となる（図5）。

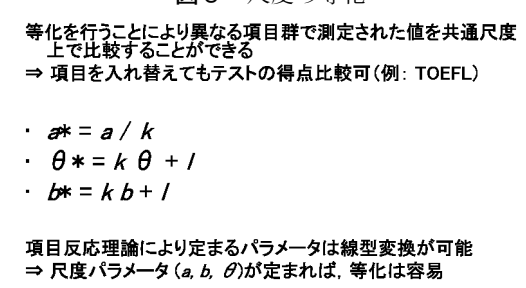
3.　項目反応理論の応用

上述したような項目反応理論の特徴を利用して，新たな尺度開発のみならず，項目分析による項目の取捨，反応尺度の形式（フォーマット）の変更などによる既存尺度の改良に応用されている。集団の能力でなく（集団に依存せず），個人の特性 θ が推定可能なことから，古典的テスト理論では不可能であった応用の可能性がある。代表的な例を以下に述べる。

1）　尺度の等化が容易

複数の異なる尺度による測定結果を共通に表示できる共通尺度を構成する手続きを等化 equating と呼ぶ。等化を行うことにより異なる項目群で測定された値を共通尺度上で比較することができる。項目反応理論のパラメータは線型変換が可能で（図6），項目反応理論により尺度パラメータが定まれば，等化は容易である。等化により項目を入れ替えても尺度の得点比較が可能となり（その項目に正答できる被調査者の能力が推定できているから！），異なる対象に対して，異なる試験問題を使用しても，同一の尺度（基準）で被調査者の能力を表すことができる。項目反応理論による等化の手続きは，TOEFL, TOEIC，大学入試センター試験の追試験などで応用されている。

図6　尺度の等化



2）　適応型テストへの応用

項目反応理論を適用すれば，被験者に最適な項目を選択し測定を行うことが可能になる。古典的テスト理論による従来の尺度では，尺度を構成する一定数の項目すべてに回答してもらう必要がある。コンピュータ適応型テスト Computerized-Adaptive Testing（CAT）では，項目の困難度が判明している利点を応用して，直前に実施した項目に対して被調査者が正答した場合には，次にはより難しい項目を抽出して実施し，逆に被調査者が誤答した後には，より容易な項目を実施するといった手続きを逐次繰り返すことによって被調査者の能力を推定する。被調査者に，その構成概念を測定する項目すべてに回答

してもらうことなく能力の推定は可能であり，調査項目数，ひいては，調査時間の節約をすることができる。また，同じ調査票を繰り返して使用することにより発生する，被調査者の調査に対する慣れ（記憶）の排除が可能となる。

3）　特異項目機能のチェック

特定のグループに関して系統的に異なる結果をもたらす項目を，バイアスのある項目という。項目反応理論の文脈では，同一の特性 θ を有する被調査者群で，その特性 θ を測定する同一項目を肯定する確率が異なっている（異なるICCが描かれる）[3]ことで同定される。テスト結果により社会的に不利を受けていたという歴史的な経緯から，バイアスという用語が使用されていたが，現在では純粋に統計的な差異が見出されるという観点から，特異項目機能もしくは差異項目機能 Differential Item Functioning（DIF）という用語が好まれて使用される。

異なるグループ間での尺度の比較妥当性を高められる可能性については，連載の5回で一部紹介した[4]。DIFを有する項目を除いたり，ワーディングを修正したりすることにより，尺度の改良を行なっている例がある。

4.　項目反応理論の応用のネック

上述のように，項目反応理論は古典的テスト理論を凌駕する多くの特質を有しており，好ましい応用例も多い。しかし，数学的に複雑で標準的な統計パッケージでは扱えない，パラメータ推定の手続きには多くの標本数が必要（TOEFLには，膨大な数のアイテム・プールがある），モデルの適合を含めて理論を適用する前提（表1）を満たすことは，実際はかなり厳しい，などといった応用上のネックがある。以上のような理由で，項目反応理論は一部専門家によって使用されているのが現状であるが，これから尺度開発を目指す研究者は勉強をしておきたい理論である。

5.　まとめ

最後に，項目のバイアスについて考えさせられる研究結果を挙げて，連載の締めくくりにつなげたいと思う。

Center for Epidemiologic Studies Depression Scale（CES-D）は，世界中で汎用されている抑うつ症状のスクリーニングツールで，その刺激性の少ない穏やかなワーディングのため，わが国でも疫学研究や地域調査でたいへんよく使用されている。Graysonらは，この尺度を75歳以上の高齢者に適用し，身体的な障害がCES-D尺度得点に影響することを見出している。すなわち，うつ状態のあるなしにかかわらず，身体障害を有する高齢者ではCES-D得点高値となり，CES-D得点の評価に留意する必要を述べている[5]。

われわれは，測定する対象と目標（構成概念）を定めて尺度を開発・活用している。開発の立場でも，利用者の立場でも，自らの測定の正確性―測ろうとしているものを測定できているのか，信頼できる測定ができているのか―ということを常に意識しておくことが必要であることをあらためて強調しておきたい[1,6]。

表2　まとめ

現在の心理社会的要因の測定は，一般に古典的テスト理論に基づいて行われており，いくつかの限界を有していることを認識しておく必要がある。
古典的テスト理論に基づいて表わされる尺度の特性は，尺度全体の平均的指標であり，尺度の特性のみならず，被調査者集団の特性の分布に依存している。
項目反応理論では，集団に依存せず個人の特性 θ が推定可能なため，応用範囲が広い。
何を測定しているのか？それは正確に測定されているのか？ということを常に意識しておく。

文　献

1）　堤　明純．心理社会的要因の測定(1)「心理特性Ⅰ信頼性」．日本公衛誌 2009; 56(4): 271-274.

2）　堤　明純．心理社会的要因の測定(4)「尺度の開発Ⅱ尺度の編集と標準化」．日本公衛誌 2009; 56(7): 485-488.

3）　Angoff WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H, editors, Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993; 3-23.

4）　堤　明純．心理社会的要因の測定(5)「外国で開発された尺度の応用と調査票の作成」．日本公衛誌　印刷中.

5）　Grayson DA, Mackinnon A, Jorn AF, et al. Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. J Gerontol B Psychol Sci Soc Sci 2000; 55: 273-282.

6）　堤　明純．心理社会的要因の測定(2)「心理特性Ⅱ妥当性」．日本公衛誌 2009; 56(5): 338-340.