都留文科大学研究紀要　第88集（2018年10月）
The Tsuru University Review, No.88（October, 2018）

＜研究ノート＞

心理測定におけるテスト理論
Test Theory for Psychological Measurements

市　原　　学

ICHIHARA Manabu

1　はじめに

心理学では、能力や性格等の人の様々な諸特性を構成概念（construct）と呼び、それらを何らかの方法（テスト、検査）によって測定し数値化する。主な例としては学力テストや性格検査などが挙げられる。しかしながら、自然科学の主な研究対象である物質とは異なり、構成概念は実際には目で捉えることも手に取ることもできない、いわば直接観測することが不可能なものである。つまり、構成概念とは説明の節約のために生み出された架空の存在でしかない。それでも、人間の様々な精神現象を説明するうえで非常に有益であるため、心理学ではこの構成概念を使用し続けるのである。

心理学では構成概念を定量的に数値化することによって研究を進めるのだが、この方法論に対して他の研究分野から批判を受けることも少なくない。その最たる例は「人のこころは数字で測れるものではない」というものだろう。たしかに、人のこころを完全に数字で捉えることは難しい。しかしながら、我々が日常接する学力を例にとれば、「この子は学力が高そうだ」と思える学生は実際に期末テストでも良い点を取っていることが多い。一方で、時々“できる”と思われる学生が思いのほか悪い点数を取ってしまうこともある。このように、心理テストはある程度構成概念を正しく捉えることができる反面、時には測定ミスを犯すこともあると考えるのが正しいと考えられる。

2　テストの平均点、分散

上記のように、心理テストは構成概念をある程度正しく測定する道具であるが、完全なものではない。使っているテストがほとんど使い物にならない代物であるなら、心理学における理論構築は到底不可能なものになってしまう。そこで心理学では、どの程度正しく構成概念を測定することができているかを把握し、その結果問題があればテストを改良することに腐心してきた。このようなテストを評価する分野を心理測定学（psychometrics）、その準拠する枠組みをテスト理論（test theory）と呼ぶ。

テスト理論において最も重要なのは、ある個人（i）がテストで獲得した得点（test score; x_i）は、そのままその者の実力、真値（true score; τ_i）を表すわけではないと考えることである。学力テストではよく起こりうることだが、テスト受験時にたまたま体調を崩したとか、張っていたヤマが当たったなど、実力以外の誤差（error; ϵ_i）の成分がテスト得点に影響を及ぼすことがある。

したがって、ある個人のテスト得点は下の (1) 式のように表現することができる。

$$x_i = \tau_i + \epsilon_i \tag{1}$$

そして、受験者数が十分に多いときには、誤差の期待値、および誤差の期待値と真値の期待値の積について

$$E[\epsilon] = 0 \tag{2}$$

$$E[\tau, \epsilon] = 0 \tag{3}$$

と仮定する。(2) 式からは、十分な数の受験者がいれば誤差はバランスよく出現し、(3) 式から真値と誤差の相関はないものと考えることができる[*1]。

2.1　平均

一般にテストを行った場合、受験者は自分の得点と、全体のできばえを示す指標である平均（mean）が気になるところである。誤差の期待値 ($E[\epsilon]$) が 0 であることを考慮しつつ、n 人のテスト得点の平均 ($\overline{x}$) を真値 (τ) および誤差 (ϵ) を使って表現してみると、

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (\tau_i + \epsilon_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \tau_i + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \tau_i \\
&= \overline{\tau}
\end{aligned} \tag{4}$$

となり、テスト得点の平均と真値の平均は一致する。

2.2　分散

ところで、入試のような受験者の選抜が目的となるテストにおいては、全体の出来映えよりもどの程度得点がばらついているかのほうが重要になってくる。この得点の散らばりの程度を分散（variance）と呼び、以下の (5) 式で表すことができる。

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5}$$

そしてこの分散を真値および誤差を使って書き換えると以下の (6) 式になる。

$$\begin{aligned}\sigma_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(\tau_i + \epsilon_i - \overline{\tau})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[(\tau_i - \overline{\tau})^2 + \epsilon_i^2 + 2(\tau_i - \overline{\tau})\epsilon_i]\end{aligned} \tag{6}$$

ここで、上の (2) 式、(3) 式から誤差の期待値および真値と誤差の積の期待値は 0 であることから、(6) 式右辺第三項の総和もやはり 0 となる。最終的にテスト得点の分散は、

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_\epsilon^2 \tag{7}$$

となる。結局のところ、テスト得点の分散は真値の分散と誤差の分散に分解される。

３　測定の精度、信頼性、古典的テスト理論

上の (7) 式から、テスト得点の分散は真値の分散と誤差の分散の総和であることがわかった。そこで今度は (7) 式の両辺を σ_x^2 で割ってみると、以下の (8) 式になる。

$$\begin{aligned}1.0 &= \frac{\sigma_\tau^2 + \sigma_\epsilon^2}{\sigma_x^2} \\ &= \frac{\sigma_\tau^2}{\sigma_x^2} + \frac{\sigma_\epsilon^2}{\sigma_x^2}\end{aligned} \tag{8}$$

右辺第一項はテスト得点の分散における真値の分散が占める割合を、第二項は誤差分散が占める割合を表している。つまり第一項はテストにおいてどれだけ構成概念の個人差が反映されているかを示しており、これを信頼性係数（reliability coefficient; ρ）と呼んで測定の精度を知る手がかりとする。反対に、第二項はテスト得点がどれだけ（構成概念の測定においては望ましくない）体調や運の良し悪しなどの誤差に左右されているかを示す指標といえる。ところで、心理学では現実に起こる精神現象を手がかりに理論構築するのであって、扱う数値（データ）は実数だけである。したがって、(8) 式における $0 \leq \sigma_x^2$、$0 \leq \sigma_\tau^2$、および $0 \leq \sigma_\epsilon^2$、それから $\sigma_\tau^2 \leq \sigma_x^2$、$\sigma_\epsilon^2 \leq \sigma_x^2$ という大小関係が導き出される。結果的に信頼性係数 (ρ) は

$$0 \leq \rho \leq 1.0 \tag{9}$$

の範囲を取ることになる。信頼性係数が1.0 に近づくほど（構成概念の個人差を反映する）精度の高いテストであるといえる。ただし実のところ、ある個人の得点から真値と誤差を切り分けることは不可能であり、信頼性係数の値を算出することはできない。そこで心理学では様々な代替方法（例 ; 再検査法、折半法、α 係数、ω 係数）を用いて信頼係数の推定値を算出し、テストの精度を把握してきた。

3.1　再検査法、折半法

再検査法（test-retest method）とは、一つのテストをある程度の時間間隔（例えば二週間）を空けて二回実施し、それらの回答結果間の相関係数をとる方法である。構成概念を強く有する（つまり真値の値が高い）受験生なら、何度テストを受けた場合でも（真値の低い受験生よりも）高い得点をとると考えられるので、再検査相関係数も強い正の値を示すはずである。二回のテスト結果をそれぞれ x_i、x'_i とすると、再検査法による信頼性係数は以下のように導出できる。

$$\begin{aligned} x_i &= \tau_i + \epsilon_i \\ x'_i &= \tau_i + \epsilon'_i \end{aligned} \tag{10}$$

$$\begin{aligned} r_{x,x'} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})(x'_i - \bar{x})}{\sigma_x \sigma_{x'}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(\tau_i - \bar{\tau} + \epsilon_i)(\tau_i - \bar{\tau} + \epsilon'_i)}{\sigma_x \sigma_{x'}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(\tau - \bar{\tau})^2 + (\tau - \bar{\tau})\epsilon_i + (\tau - \bar{\tau})\epsilon'_i + \epsilon_i \epsilon'_i}{\sigma_x \sigma_{x'}} \end{aligned} \tag{11}$$

誤差の平均と分散は任意なので、$\sigma_\epsilon = \sigma_{\epsilon'}$ とすると $\sigma_x = \sigma_{x'}$ となり、

$$r_{x,x'} = \frac{\sigma_\tau^2}{\sigma_x^2} \tag{12}$$

のように、信頼性の定義式に一致する。

再検査法は心理学における性格特性や態度などの質問紙開発においてよく用いられる手法である。しかし当然ながら、学力テストにおいては記憶の影響などによって真値以外の成分が混入し、再検査相関係数の値を歪めるので、再検査法以外の代替的な方法を使用することになる。また再検査法は、時間的に容易に変動しやすい構成概念（例；感情、生理的欲求など）を測定する場合の信頼性の指標には向いていない。

折半法（split half method）とは、複数の問題から構成されるテストを二つの部分テストに分けて、それらの相関係数をとる方法である。テストに含まれる各問題がある構成概念を反映したものなら、当然ながら二つの部分テストもその構成概念を反映したものとなり、両者には正の相関関係がみられるはずである。折半法は再検査法とは異なり、記憶の効果による反応歪曲や、時間的に変動しやすい構成概念に対しても頑健である。しかしながら折半法は、テストをどのように二つに分けるかという問題の組み合わせ次第で信頼性推定値がばらつくという欠点を持っている。なお、折半法の導出方法については再検査法と同じであるので省略する。

3.2 α係数

たとえば、数学の学力テストを実施する場合、通常複数の問題（以下、心理学における慣例にしたがい、“項目”と呼ぶ）から構成されることがほとんどである。それはテストがたった一つの項目から構成されている場合には、「たまたまその項目に出会ったことがある（または、ない）」といった偶然の要因（誤差）の影響を強く受けると考えるからである。つまり、項目数を多くしていくことでそういった（“たまたまその項目を知っている”といった）誤差の個人差の影響がどんどん薄まっていき、結果的により純粋な真値に近づくと考えられる[*2]。

そこで、i という個人が学力テストを受けて、j 番目の項目に回答したときの反応パタン（正答、誤答）を x_{ij} として、真値（τ_i）、および項目の誤差（ϵ_{ij}）との関係を(10)式のように表す。

$$x_{ij} = a\tau_i + \epsilon_{ij} \tag{13}$$

係数 a は、項目がどの程度真値を反映しているかを示す重みである。そして、k 個の項目から構成されるテストにおいてすべての項目は等しく真値を反映していると仮定すると[*3]、ある個人の総得点（x_i）は以下の(11)式で表現できる。

$$\begin{aligned} x_i &= \sum_{J=1}^{k}(a\tau_i + \epsilon_{ij}) \\ &= ka\tau_i + \sum_{J=1}^{k}\epsilon_{ij} \end{aligned} \tag{14}$$

(1)式との関係から、$a = 1/k$ である。

さらに、誤差はバランスよく出現し（$E[\epsilon_{ij}] = 0$）、誤差と真値の間には相関がなく（$E[\tau_{ij}, \epsilon_{ij}] = 0$）、かつ誤差間の相関がない（$E[\epsilon_{j,j'}] = 0$）と仮定したうえで、異なる二つの項目間の共分散（$\sigma_{J,J'}$）を求めると、

$$\begin{aligned} \sigma_{J,J'} &= E[(a\tau_i + \epsilon_{ij})(a\tau_i + \epsilon_{ij'})] \\ &= E[a^2\tau_i^2] \\ &= a^2\sigma_\tau^2 \end{aligned} \tag{15}$$

$a = 1/k$ であるから、項目間共分散と真値の分散の間には、

$$\sigma_\tau^2 = k^2\sigma_{J,J'} \tag{16}$$

という関係が得られる[*4]。

そしてタウ等価測定のもとでは添字によらず全ての項目間の共分散は等しいと考えられるが、現実にはそうはならないので、$\sigma_{j,j'}$ の推定値である $\hat{\sigma}_{j,j'}$ は共分散の平均値である、(14)式のように求める。

$$\hat{\sigma}_{j,j'} = \frac{1}{k(k-1)}\sum_{j=1}^{k}\sigma_{j,j'} \tag{17}$$

次に、(13)式の右辺 $\sigma_{J,J'}$ に(14)式の $\hat{\sigma}_{j,j'}$ を代入すると、以下の(15)式のようになる。

$$\sigma_{\tau}^2 = \frac{k}{(k-1)} \sum_{j=1}^{k} \sigma_{j,j'} \tag{18}$$

さらに信頼性係数の定義式 $\sigma_{\tau}^2/\sigma_x^2$ に(15)の右辺を代入すると、その推定値($\hat{\rho}$)は、

$$\begin{aligned} \hat{\rho} &= \frac{k}{k-1} \frac{\sum_{j=1}^{k} \sigma_{j,j'}}{\sigma_x^2} \\ &= \frac{k}{k-1} \left(1 - \sum_{j=1}^{k} \sigma_j^2/\sigma_x^2 \right) \\ &= \alpha \end{aligned} \tag{19}$$

となり、これを α 係数と呼ぶ。

α 係数は内的一貫性の指標ともいわれる。つまり、項目間の共分散（相関）が大きければ大きいほど、言い換えれば各項目が同一の上位概念を測っている度合いが強いほどその値が1.0 に近づく。厳密な基準はないが、心理学においては慣例的に0.7を超える程度であれば、信頼性が高いと判断する。

なお、(16)式にもあるように、 α 係数はテストに含まれるすべての項目の組み合わせの共分散を利用している。これは上記の折半法の考え方を拡張、一般化したものといえる。

3.3　ω係数

前節では信頼性推定値の指標である、 α 係数を紹介した。しかしながら α 係数はタウ等価測定という、やや（というよりも、かなり）無理のある前提のもとで導出された指標であることは否めない。現実にはテスト受験者の構成概念における個人差を敏感に検出できる項目もあれば、そうでない項目もある。それでも α 係数が使用されてきたのは、かつては計算機の処理能力が追いつかなかったという時代的制約によるところが大きかった。しかしながら、(比較的) 安価で高性能な計算機が容易に手に入るようになった現代においては、 α 係数よりも正確な信頼性係数を算出することも可能である。本節では、Mcdonald(1978) が提案した ω 係数を紹介する。

ω 係数はタウ等価測定の制約を外した指標であり、 α 係数と同様に内的一貫性の指標となる。 ω 係数は因子分析の立場から提案されたものであり、独自成分に基づいて算出される。前節で紹介したように、誤差はバランスよく出現し($E[\epsilon_{ij}] = 0$)、誤差と真値の間には相関がなく($E[\tau_{ij}, \epsilon_{ij}] = 0$)、かつ誤差間の相関がない($E[\epsilon_{j,j'}] = 0$)と仮定したうえで、誤差の分散(σ_e^2)は、各項目の独自分散の総和となる。

$$\sigma_e^2 = \sum_{j=1}^{k} \sigma_{\epsilon}^2 = \sum_{j=1}^{k} d_j^2 \tag{20}$$

(17) 式における d_j^2 は、因子分析[*5]をすると各項目について算出される独自性と呼ばれるものである。独自性は共通因子によって説明されない成分である。

(8) 式を変形したうえで (17) 式を代入すると、ω 係数が得られる。

$$\omega = 1 - \sum_{j=1}^{k} d_j^2 / \sigma_x^2 \tag{21}$$

繰り返しになるが、α 係数におけるタウ等価測定の制約を外した指標が ω 係数である。つまり、ω 係数は α 係数よりも正確な内的一貫性および信頼性の指標であり、反対に α 係数は内的一貫性の下限値であるともいえる。

3.4 信頼性の改善に向けて

信頼性係数の推定値である α 係数や ω 係数の目安が0.7であることは先ほど述べた通りである。これまで長い間、多くの心理学者が信頼性の高いテストを作ることに心血を注いできたといっても過言ではないが、それはなぜだろうか。

(8) の信頼性の定義式からも明らかなように、信頼性とはテスト得点の分散における真値の分散の割合である。裏を返せば、信頼性が低いということは誤差分散の割合（つまり運の良し悪しの個人差）が大きいことである。そして (3) 式で示したように、誤差は真値はおろか他の何物とも関連性を持たない（だからこそ、誤差なのである）[*6]。つまり信頼性が低いテストというのは、いわば（でたらめに当たり外れが決まる）サイコロや宝くじのようなものといえる。入試のような、ある意味受験者にとっては人生の方向性を決めるような重要なテスト（これを“ハイステークス・テスト”と呼ぶ）が、実はサイコロと同等の代物でしかなかったとしたら、（特に不合格になった）受験者はどのように思うだろうか。また、入試を実施する側も、サイコロで合否を決めてしまうようなずさんな試験を実施して社会的責任を果たせているといえるだろうか。こうした現実的な問題に対処するうえでも、信頼性を保障することはテストを作成、採点する人間にとって必要な責務といえる。

しかしながらこれらの値は、テスト実施後、受験者の回答パタンを分析することではじめて明らかになるのであって、時には0.68 というように、基準値を下回ってしまうこともあるだろう。このように低い信頼性推定値が得られた場合には、そのテストは誤差の影響が大きいということで全く使い物にならないのだろうか。

実は、信頼性の高いテストを作るためにテスト実施前後でできる工夫がある。

まず、テスト実施前の工夫について概説する。繰り返しになるが、テストには複数の項目が含まれる。そしてこれらの項目は数学の学力など、より上位にある構成概念を反映したものでなければならない。つまり、数学の学力が高いと思われる受験者ならば（学力が低い受験者よりも）、高い確率で正答できると思われる項目を作成、選択しておく必要がある。テスト作成者は常に上位の構成概念との関係を念頭に置きながら項目を作成選択するべきで、ただ漫然と項目を適当に選んだり、「こういうことを問うたら面白そうだ」という根拠のない興味や主観で選ぶなど言語道断である。極端な例ではあるが、数学のテス

トに国語の漢字書き取りの問題を含んではいけないということである。漢字の書き取りは国語の学力であって、数学の学力を測る指標ではないからである。この例のように、他とは異質な項目はテスト全体の信頼性を損なうので、あらかじめ項目を精選しておくことが重要である。

つづいて、あまりにも難しすぎる問題、簡単すぎる問題もテストには含めないようにする。難しすぎる問題は誰も解けない。反対に簡単すぎる問題は誰でも解けてしまう。こうした問題は構成概念の個人差を識別するといった目的からは何ら役に立たないからである。しかしながら、学力テストにおいては、難易度の観点から項目を取捨選択する場合には注意が必要である。それは受験者母集団の問題である。比較的学力水準の高いと考えられる受験生が想定される場合には、難易度の高い項目をそろえた方が、そこでの個人差を敏感に検出することができる。しかしながらそうした難しい項目というのは学力水準が低い受験者集団に実施すれば、難しすぎて誰も解けないために不良項目となってしまうものである。つまり学力水準が低いと思われる受験者母集団にとっては難しすぎるために不良項目になってしまうものでも、学力水準が高い母集団ではよい項目となる場合がある[*7]。このように、テスト実施前に内容や難易度といった観点から項目を作成、取捨選択することで信頼性を保障することが可能になる。

上記のように、作成者がテスト実施前に項目の内容や難易度を検討したからといっても、必ずしも受験者の回答パタンが予想通りになるとは限らない。受験者集団の能力水準が思いのほか高く（低く）、正答率が高く（低く）なってしまうことだってあるだろう。したがって、項目の良し悪しについては実施前の経験則や主観的判断だけではなく、得られた回答パタンからも検討する必要がある。

j 番目の項目における n 人の正誤データ (0 = 誤答、 1 = 正答) のテストの難易度の指標としては項目困難度（item difficulty）、通過率（passing rate）がある[*8]。項目 j の通過率は以下の (19) 式のように算出する。

$$p_{ij} = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{22}$$

右辺分子の $\sum_{i=1}^{n} x_{ij}$ は正答者数である。通過率はおおよそ0.2～0.8程度の範囲に収まっていることが望ましい。

通過率に基づいて項目の取捨選択を行ったら、次は項目識別力（item discrimination）を算出する。項目識別力とは、個々の項目がどれほど上位の構成概念を反映したものとなっているかを示す指標である。代表的なものとして、I-T 相関（item-total correlation）[*9][*10]があり、以下の (20) 式で算出する。

$$r(x_j, x) = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(x_{ij} - \overline{x_j})}{\sigma_y \sigma_{xj}} \tag{23}$$

I-T 相関の値が低かったり、負の値を示すようであれば、その項目は上位の構成概念を示す指標ではないことになるので、除外するべきである。I-T 相関に関しても厳密な基準はないが、おおよそ0.2 を超えるようであれば問題はないと考えて差し支えない。

また、使用するソフトウェアによっては当該項目を除いた場合の α 係数も算出してくれる。当該項目を除いた場合に飛躍的に α 係数が向上するようであれば、その項目は排除したほうがよい。

最後に、信頼性（主に内的一貫性）改善を目的とした場合の項目取捨選択における問題点を挙げておく。α 係数や ω 係数は内的一貫性の指標であり、項目間の共分散（相関）関係に基づいて算出されるということは、先ほど紹介した通りである。つまり、項目の内容が似通っていればいるほど、α 係数や ω 係数は高くなっていくということである。極端な例を挙げれば、テストが全く同一内容の項目から構成されていれば、信頼性推定値は 1.0 になる。しかしながら、そうやって信頼性を高めることだけに特化し、作り上げられたテストに何の意味もないことは自明である。数値計算ができるだけで数学の学力が高いとはいえないように、我々が想定する構成概念とは豊かな内容を包含した複雑で抽象的なものである。そしてテストは構成概念を反映する道具なのである（これを妥当性と呼ぶ）。しかしながら、テスト内容を豊かにしようとすれば、信頼性が多少犠牲になるのもまた事実である。このような信頼性と妥当性の背反関係を「帯域幅と忠実度のジレンマ」（bandwidth-fidelity dilemma）と呼び、心理テストを作ることの難しさを物語る例といえる。

4　一般化可能性理論

これまで紹介してきた α 係数、ω 係数などの信頼性推定値の算出方法は、原則として一人の採点者が採点するような客観テスト（例；空所補充型テスト、多肢選択式テスト）を想定してきた。客観テストは受験者の知識や記憶を測るのに適していると考えられ、基本的には誰が採点しても、だいたい一致した採点結果になる。しかしながら他方、客観テストは論理的思考力や文章構成能力など、より広い意味での学力を捉えるのには不向きとされ、そういった後者の学力を測る場合には論述形式のテストが用いられることが多い。論述形式テストを含む、論知的思考力や表現力、それからプレゼンテーションスキルなど知識や記憶力に限定されない、より広い学力を測定しようとするテストをパフォーマンス型評価と呼ぶ。

細部のフォーマットについては異なることもあるだろうが、パフォーマンス型評価においては受験者に何らかの素材（例；文章、統計資料など）を与え、それをどのように読み解くかを問うたり、その素材に基づいて受験者に意見を求めるというものが一般的であろう[*11]。そして、一つの項目に取り組む時間や負担が客観テストに比べて格段に増すため、項目数もせいぜい二つ、三つ程度の少数から構成されることがほとんどである。採点については、客観テストに比べて、採点者の主観が反映されてしまう恐れがあるため、一人の受験者について複数の採点者が採点することが多い。

一般化可能性理論（generalizability theory; Brennan, 2001）とは、こういったパフォーマンス型評価の精度を測るのに適した手法であり、古典的テスト理論と（変量効果の）分散分析（analysis of variance, ANOVA）の手法を組み合わせることで信頼性係数推定値を算出する。以下、一般可能性理論における信頼性係数の推定方法を導出していくが、その前に分散分析の手順について概説する。

例えば授業におけるある指導方法（例；発見学習、討論）の教授効果を確かめようという実験を計画する。当然ながら、そうした実験を意図した背景には従来の伝統的な講義法

よりも高い学習効果を生み出すのかどうかが知りたいという思いがある。したがって、この実験では三種の授業を実際にやってみて、その後修得度テストを実施して、その成績に影響を与えるかどうかを検討することになる。ここで成績に影響を与える原因となる授業方法を処遇、または処理（treatment）と呼ぶ[12]。そして、三種の授業方法を処遇内の水準（level）と呼ぶ[13]。

分散分析ではある個人の得点（x_{ij}）は外部環境の影響や処遇の効果（τ_j）と偶然の誤差の合算として表現される。

$$x_{ij} = \mu + \tau_j + \epsilon_{ij} \tag{24}$$

μは母集団平均とする。

実際に得られるデータを使って、(21)式を書き換えてみる。母集団の平均を得られたデータの全平均（$\overline{T}$）、処遇の効果を水準の平均と全平均の偏差（$\overline{G}_j - \overline{T}$）、誤差を個人得点と水準平均の偏差（$x_{ij} - \overline{G}_j$）と考えると、

$$x_{ij} = \overline{T} + (\overline{G}_j - \overline{T}) + (x_{ij} - \overline{G}_j) \tag{25}$$

となる。

右辺第一項である、全平均（$\overline{T}$）を左辺に移行し、個人得点と全平均の偏差平方和（sum square）を求める。

$$\begin{aligned} SS_{TOTAL} &= \sum_{i=1}^{n}\sum_{j=i}^{k}(x_{ij} - \overline{T})^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k}[(\overline{G}_j - \overline{T}) + (x_{ij} - \overline{G}_j)]^2 \\ &= n_j\sum_{j=1}^{k}(\overline{G}_j - \overline{T})^2 + \sum_{i=1}^{n}\sum_{j=i}^{k}(x_{ij} - \overline{G}_j)^2 \\ &+ 2\sum_{i=1}^{n}\sum_{j=1}^{k}(\overline{G}_j - \overline{T})(x_{ij} - \overline{G}_j) \\ &= SS_{TREAT} + SS_{ERROR} \end{aligned} \tag{26}$$

(23)式より、全体の平方和は処遇の平方和と誤差の平方和に分解できることが明らかである。

分散分析では、全体に占める処遇の平方和が十分に大きければ、処遇の効果あり（この実験例では“教え方の違いによって、成績が変動する”）と判断するが、実際には水準数に対して繰り返しの数が大きすぎるためにこのままでは判定できない。そこで水準数や繰り返し数を調整し、処遇および誤差の平均平方（Mean Square）を算出した上で判断することになるが、それは以下の式によって求めることができる。

$$MS_{TREAT} = \frac{SS_{TREAT}}{k-1} \tag{27}$$

$$MS_{ERROR} = \frac{SS_{ERROR}}{k(n-1)} \tag{28}$$

分散分析は、複数の水準を比較する上で大変有用な分析手法であるが、さらに組み合わせの妙といった現象を解きほぐすのにも威力を発揮する。例えば先ほどの指導方法ではどの方法がもっともよいのかということを検討したが、その効果は人によって変わってくるかもしれない。たとえば人間関係において積極的（外向的）な者にとっては討論法が適しているかもしれないが、消極的（内向的）な者にとっては従来の講義法のほうが向いているのかもしれない。これは指導法という処遇（A）と学習者の適性という処遇（B）の交互作用（interaction、AB）によって生み出された効果といえる。

このような実験計画を二要因分散分析（two-way ANOVA）と呼び、全体の平方和は以下のように分解できる。

$$SS_{TOTAL} = SS_A + SS_B + SS_{AB} + SS_{ERROR} \tag{29}$$

そして本節で紹介する一般化可能性理論は、三要因分散分析のを利用したものといえる。

一般化可能性理論ではテスト得点の散らばりを、被験者（$person, p$）、項目（$task, t$）、それから採点者（$rater, r$）の 3 つの相 (facet) に由来すると考える。そして、被験者は無限に存在する母集団（population）、項目と採点者もやはり無限に存在するユニバース（universe）の中からたまたま（無作為に）選ばれたという前提を置く。そして、テスト全体の分散は、被験者、項目、採点者それぞれ、およびこれらの交互作用の分散から構成されると考える[14]。

$$\sigma^2_X = \sigma^2_p + \sigma^2_t + \sigma^2_r + \sigma^2_{pt} + \sigma^2_{tr} + \sigma^2_{rp} + \sigma^2_{ptr} \tag{30}$$

上記 (27) 式のそれぞれの分散成分の推定値（$\hat{\sigma}^2$）は、分散分析における平均平方を利用して以下のように求める。

$$\hat{\sigma}^2_p = \frac{MS_p - MS_{pt} - MS_{rp} + MS_{ptr}}{n_t n_r} \tag{31}$$

$$\hat{\sigma}^2_t = \frac{MS_t - MS_{pt} - MS_{tr} + MS_{ptr}}{n_r n_p} \tag{32}$$

$$\hat{\sigma}^2_r = \frac{MS_r - MS_{tr} - MS_{rp} + MS_{ptr}}{n_p n_t} \tag{33}$$

$$\hat{\sigma}^2_{pt} = \frac{MS_{pt} - MS_{ptr}}{n_r} \tag{34}$$

$$\hat{\sigma}^2_{tr} = \frac{MS_{tr} - MS_{ptr}}{n_p} \tag{35}$$

$$\hat{\sigma}^2_{rp} = \frac{MS_{rp} - MS_{ptr}}{n_t} \tag{36}$$

$$\hat{\sigma}^2_{ptr} = MS_{ptr} \tag{37}$$

n_p、n_t、n_rはそれぞれ、被験者数、項目数、および採点者数を示す。

分散成分のうち被験者に由来する成分（σ_p^2）が大きいということは、そのテストが項目や採点者の違いにかかわらず個人差をうまく反映していることを意味している。他方、項目の分散成分（σ_t^2）が大きいときは項目間の難易度（平均点）が大きく異なることを、また採点者の分散成分（σ_r^2）が大きいときは、甘い基準で採点する者もいれば厳しく採点する者もいることを意味している。

被験者がかかわる交互作用のうち、σ_{pt}が大きい場合は課題間で被験者の順位が入れ替わってしまうことを、また、σ_{rp}は採点者間で被験者の順位が入れ替わってしまうことを意味している。さらに、σ_{tr}は課題によって採点者の採点基準が揺らいでしまうことを意味している。

一般化可能性係数（E_{ρ^2}）は以下の(35)式で求められる。

$$E_{\rho^2} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pt}^2}{n_t} + \frac{\hat{\sigma}_{rp}^2}{n_r} + \frac{\hat{\sigma}_{ptr,e}^2}{n_t n_r}} \tag{38}$$

(35)式のうち、n_t、n_rについては、自由に値を入れることにより、任意の信頼性係数を得るのに必要な項目数や採点者数をシミュレートすることができる。

また、客観テストは*ptr*デザインのうち、採点者による分散成分およびその交互作用を抜いたもの（*pt*デザイン）と考えられるので、その一般化可能性係数は以下の(36)式のように表現できる。

$$E_{\rho^2} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pt}^2}{n_t}} \tag{39}$$

(36)式は、実際に使用した項目数を代入すると α 係数に一致する。さらには、もしも客観テストの信頼性分析をしていて残念ながら基準値（>0.70）に満たなかった場合、あと何項目増やせばよいかという目安を得ることができる。

5　項目反応理論

次のような状況を考えてほしい。

・A、B中学校のそれぞれで身体測定があった。A中学校のa君の体重は55kg で、B中学校のb君の体重は57kgであった。どちらがどれだけ重いか。

・A、B中学校のそれぞれで定期学力（教師自作）テストがあった。A中学校のa君の数学の点数は55点で、B中学校のb君の点数は57点であった。どちらがどれだけ数学ができるだろうか。

一つ目の問いについては、多くの人が“B君のほうが少しばかり重い”と答えられるだろう。しかしながら二つ目の問いについては、“b君のほうが少し数学ができる”とはいわないだろう。なぜならば、B中学校のテストのほうが易しかった（A中学校のテストのほうが難しかった）という可能性も排除できないからである。体重（質量）については絶対的な原点があり、そして目盛りに意味づけがなされているため、異なる場所、時間において測定を行ってもそれらの比較が可能となる。しかしながら学力テストのような心理特

性を測る物差し（尺度）は原点がなく、目盛りの意味づけも任意であるため（尺度の不定性；arbitrariness of scale）、どちらがどれだけ優れているか（劣っているか）について結論が下せないのである。古典的テスト理論によってテストの信頼性や測定の標準誤差を推定することはできるが、テストにおける目盛りの等間隔性について知ることはできないし異なるテスト結果を比較することもできない。

また、（現実には起こりにくいが、）次のような例を挙げてみよう。

・中程度の学力の持ち主であると思われる中学生のｃさん、ｄさんが何かの手違いによって非常に難易度の高い高校の受験をしてしまったところ、二人ともテストで０点を取ってしまった。二人の実力は同じといえるだろうか？

この場合は、テストの難易度が高すぎて、二人の学力の個人差を識別することができなかったのだろう。しかしながら、このテストは当該高校を受験する高学力の受験者層の個人差を把握するのには有効なのだろう。つまり、テストと受験者の間には相性というものがあって、テストが敏感に個人差を識別できる範囲（守備範囲）というものがあると考えられる。古典的テスト理論に依拠すると信頼性係数（の推定値）は一つのテストにつき一個算出されだけであり、テストと受験者の相性、テストの守備範囲については不明のままである。

項目反応理論（item response theory; 以下、IRT）とは古典的テスト理論の限界を超えて、より実用的なテスト運用を可能にする新たなテスト作成の枠組みである。IRT の目的は、後述する項目特性曲線（item characteristic curve; ICC）、テスト特性曲線（test characteristic curve; TCC）、および情報関数（information function; IF）を描き、項目やテストの素性を明らかにすること、項目やテストの難易度とは独立に受験者の能力値や特性値の高さを推定すること、さらにはテストの守備範囲を明らかにすることである。

5.1　IRT の前提

最初に、IRT ではテストデータを分析するにあたって以下のような仮定が置かれている。

・一つのテストで測定される能力、特性（構成概念）は一つである（一次元性の仮定）。
・一つの構成概念を測定するために多数の項目が用意される。
・能力値、特性値の高いと思われる受験者は各項目への正反応率（正解確率）およびテスト得点が高い。
・ある項目に対する反応は、他の項目に対する反応に依存しない（局所独立；local independence）。

また、IRT によってテストデータを分析する以前に、先述のような項目分析や信頼性係数などの算出を行って、不良項目を除外しておくと、データのモデル適合度を高めるのに役立つ。

5.2　テスト特性曲線

テスト特性曲線（以下、TCC）とは二次元平面上の横軸に（直接観測することのできない）テスト受験者の能力値や特性値（以下、IRT の慣例にしたがい θ と表記する）、縦軸にテスト得点を置き、両者の関係をプロットしたものである。先述のように、テストはその全範囲において等しく個人差を識別できるわけではなく、極端な能力値、特性値の範囲では識別力が落ちると考えられ、Fig.1 のような形状を描く。IRT における θ はおおむね標準正規分布（standardized normal distribution）の標準得点 (z score) および偏差値 (Z score) に対応しており、$\theta = 2$ は偏差値70でテスト受験者全体の中で上位 2 ～ 3 % 程度の位置づけになる。同様に、$\theta = 0$ は偏差値50、$\theta = -2$ は偏差値30に対応する。ただし、TCC は項目の組み合わせ次第で、その位置や傾きが変化することもあり、その形状が必ずしも一義的に決まっているわけではない（Fig.2）。別の見方をすれば、異なるテストであっても同じ構成概念を測定しているのであれば、それらのテストを θ の次元で比較することが可能となる。たとえば、あるテスト60 点を取った場合、それが θ の次元上のどこに位置づけられ、同時にその θ が他のテストでは何点に相当するのかといったことがわかるようになる。

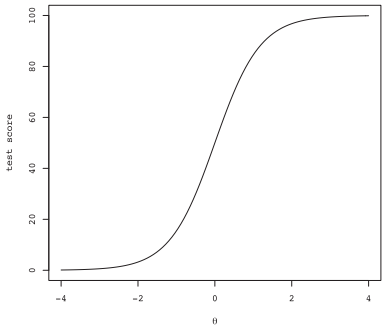

Fig. 1 Test characteristic curve(TCC)

また、Fig 1 、Fig.2をみてわかるように、どの TCC もその傾きが全範囲において一定ではない（一次関数のような直線ではない）。Fig.1 の曲線をみると、θ が - 4 から - 2 の間、2 から 4 の間では、θ が変化してもテスト得点はほとんど変わらない。一方で θ が - 2 から 2 の間ではテスト得点が急激に変化しているのがみてとれる。学力テストで考えれば、このテストは非常に学力の低い層、または高い層の学力の変化や個人差を検出するのにはあまり有用ではないことを意味している。反対にこのテストは中程度の学力層の受験者の変化や個人差を検出するのを得意としている。後述する情報関数と併用しながら TCC を吟味することによってテストの守備範囲を明らかにすることができ、想定する受験者の変化や個人差をより正確に検出することが可能となる。

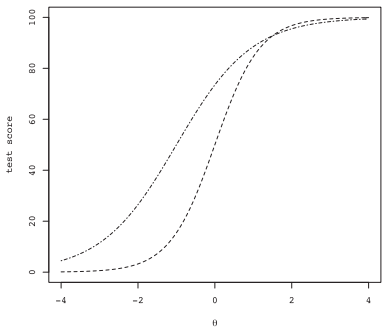

Fig. 2 Comparing different TCC's

5.3 項目特性曲線

項目特性曲線（以下、ICC）とは2次元平面上の横軸をテスト受験者の θ、縦軸を当該項目への正反応確率[*15]として、両者の関係をプロットしたものである（Fig.3）。ICCは θ が高い受験者ほど正反応確率が高いという単調増加関係を示しているものの、その増加率（傾き）は θ の全範囲において一定ではない。IRT 黎明期にはICC を正規累積モデル（normal ogive model）で表現していたものの、積分計算を含み計算が煩雑になるため、現在ではロジスティックモデル (（logistic model）を利用するようになった。

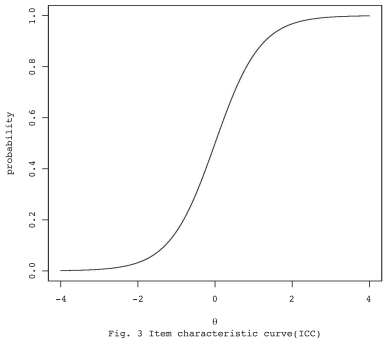

Fig. 3 Item characteristic curve(ICC)

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{(-1.7 \times a_j \times (\theta - b_j))}} \tag{40}$$

(40) 式のうち、a_j はj番目の項目の識別力（ロジスティック曲線の傾斜の度合い）を、b_j は項目jの困難度または位置（ロジスティック曲線の位置）[*16]を、さらに c_j は当て推量確率（ロジスティック曲線の下方漸近線）を意味する。IRTでは必ずしもこれら3つの母数を推定しなければならないということはない。テストの形式やテスト受験者数、およびテスト作成者（分析者）の仮説などを考慮し、b_j だけを自由推定する一母数ロジスティックモデル（one parameter logistic model; 1 PLM）、b_j と a_j を自由推定する二母数ロジスティックモデル（two parameter logistic model; 2 PLM）、および b_j、a_j、c_j すべてを推定する三母数ロジスティックモデル（three parameter logistic model; 3 PLM）がある。

先述のように、1 PLMでは困難度母数 b_j だけを自由推定する。困難度母数とはそれぞれのロジスティック曲線において正反応確率=0.5 となるときの θ の値である。1 PLMの例をFig.4 に示した。一番左側の曲線は b_j = - 1、真ん中は b_j = 0、右側は b_j = 1である。これら三つの曲線は形状は全く同じであるが、並行移動したものであり、交わることはない。すなわち、θ の全範囲において、すべての受験者は右側よりも真ん中や左側（および真ん中よりも左側）の曲線の項目への正解確率が高いということになる。また、1 PLM では正反応数と θ の間に一対一の対応関係があり、どの問題に対して正解したかという反応パタンは θ の推定に影響を及ぼさない。

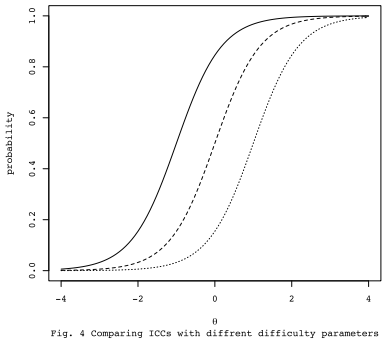

Fig. 4 Comparing ICCs with diffrent difficulty parameters

つづいて、2 PLMでは各項目について困難度母数 b_j に加えて、識別力母数 a_j も自由推定する。識別力母数とは各ロジスティック曲線の困難度 b_j = θ における接戦の傾きを示す指標で、その近辺の θ の範囲でその項目がどれだけ個人差を敏感に検出できるかということを意味している。Fig.5 にあるように傾斜のけわしいロジスティック曲線（a_j = 1.5、b_j = 0）は、緩やかな曲線（a_j = 0.5、b_j

= 0 ）よりも θ = 0 近辺においてより敏感に個人差を検出できている（ 同じ θ の範囲でも正解確率が大きく変動している）様子が見て取れる。

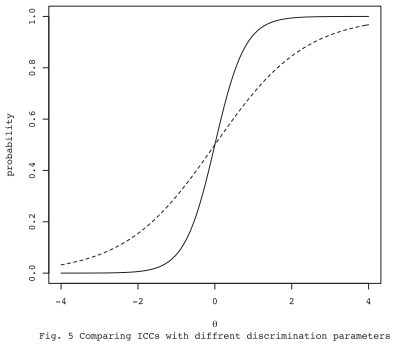

Fig. 5 Comparing ICCs with diffrent discrimination parameters

ただし、 2 PLM（および 3 PLM）においては必ずしも識別力母数の値が大きな項目がよいというわけではない。Fig.5において θ = 0 を離れて、 θ = 2 （または - 2 ）付近になると、むしろ識別力母数の低い $a_{0.5}$ 項目のほうが検出力が高い（傾きがけわしい）という逆の現象が生じていることがわかる。このことから、学力テストにおいて受験者の学力層が一定の狭い範囲内に収まっていることがあらかじめ予想される場合には識別力母数の高い項目を出題するのがよく、反対に受験者の学力層が不明である場合や、非常に広範囲にまたがると予想される場合には、識別力母数が（ある程度低い）項目も出題するのがよいと考えられる。なお、 2 PLM（や 3 PLM）では、同じ正答数でも、反応パタンの違いによって、異なる θ が推定されることがある。

最後に、 3 PLM では困難度母数 b_j や、識別力母数 a_j に加えて、当て推量確率 c_j も自由推定する [*17]。当て推量確率 c_j とはロジスティック曲線の下方漸近線を示し、どんなに θ の低い受験者でも（偶然） 正反応してしまう確率を意味している。 したがって、 3 PLM および c_j の推定は、マークシート形式に代表される多肢選択型のテストにおいて利用される。Fig.6 に c_j = 0 （ b_j = 0 、 a_j = 1 ）のロジスティック曲線、および c_j = 0.25 （ b_j = 0 、 a_j = 1 ）の曲線を示す。Fig.6に示されている通り、 θ の全範囲にわたって c_j = 0.25の曲線は c_j = 0 の曲線の上に位置している。つまり学力テストにおいては c_j = 0.25の曲線のほうが容易に正解しやすいといえる。 なお、 3 PLM では困難度母数 b_j や識別力母数 a_j も正解確率 =0.5の場所に位置しているわけではないことには注意が必要である。

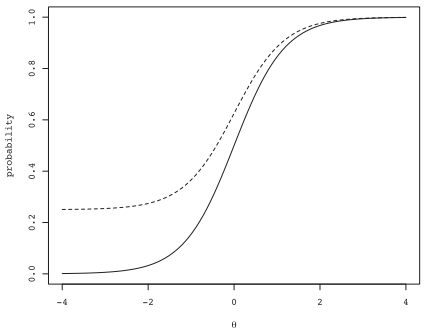

Fig. 6 Comparing ICCs with different pseudo-chance level parameters

5.4　TCC と ICC の関係

TCC はテストに含まれる項目の ICC を積み重ねたものである [*18]。

$$T(\theta) = \sum_{j=1}^{J} P_j(\theta) \tag{41}$$

5.5 母数推定

母数推定においては以下のような状況が考えられる。

1. 受験者の能力値や特性値、および項目母数が未知の状況ですべてを推定する（新しいテストの開発）。
2. 項目母数が既知の状況で受験者の能力値や特性値を推定する（テストの運用）。
3. 受験者の能力値や特性値が既知の状況で項目母数を推定する（項目の追加、精錬）。

いずれの状況においても最尤推定法（maximum likelihood estimation; MLE）、ベイズ推定（Bayesian estimation）が利用される。重要なことに、IRT で推定される能力値や特性値の推定は項目の表面的な特徴（e.g., 難易度）に依存しないし、項目母数の推定も標本の偏りの影響を受けにくい。このことから IRT では古典的テスト理論のような厳密な標本調査を行わずとも項目母数の推定ができるため、容易なテスト開発が可能となる。

しかしながら IRT における母数推定法に問題がないわけではない。MLE を使った場合、全問正解者（不正解者）の θ が推定されないとか、全員正反応（誤反応）の項目については項目母数が推定されないといった問題が生じる。一方ベイズ推定によって能力値や特性値を推定する場合、その推定値が事前分布の影響を受けるため、結果が安定しないという問題が生じる。

5.6 情報関数

古典的テスト理論では、テストの測定精度は信頼性係数という形で、一つのテストにつき一個算出された。たとえば、α =0.8という値であれば、そのテストは誤差の分散が小さく、精度が高いと評価される。しかしながら、これまで繰り返し述べてきたように、テストの精度は、θ の全範囲に渡って一定というのは考えにくく、精度よく個人差を識別できる範囲（守備範囲）というものがある。IRT では情報関数（information function; 項目情報関数（item information function; IIF）を $I_j(\theta)$、テスト情報関数（test information function; TIF）を $I(\theta)$ と表す）という形でテストの測定精度を表し、項目およびテストの測定精度および守備範囲を示す。

IIF は項目の測定精度を示す指標であり、以下の式で求められる。

$$I_j(\theta) = \frac{P_j'(\theta)}{P_j(\theta)Q_j(\theta)} \tag{42}$$

ここで、$P_j'(\theta)$ は $P_j(\theta)$ を θ で微分した導関数である。また、$Q_j(\theta)$= 1 - $P_j(\theta)$ であり、項目 j に対する誤反応確率を表している。IIF は $\theta = b_j$、つまり θ が項目困難度パラメーターに一致するときに最大となる[19]。

TIF はテストの測定精度を示す指標であり、IIF を足し上げることで求められる[20]。

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta) = \sum_{j=1}^{J} \frac{P_j'(\theta)}{P_j(\theta)Q_j(\theta)} \tag{43}$$

TIF の具体例を Fig.7に示す。

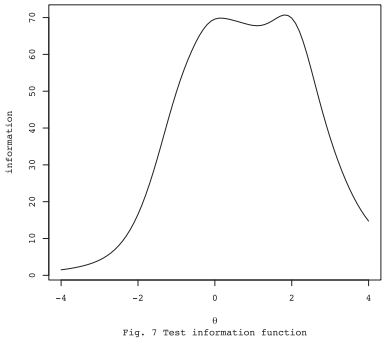

Fig.7をみてわかるように、テストの精度（情報量）は全範囲にわたって一定ではない。この例では、θ が 0 から 2 あたりで測定精度が最大となり、その範囲を越えると急激に情報量が減少していく。学力テストであれば、平均的な学力層から上位層あたりの個人差を識別するのに有用なテストであるといえよう。

なお、テストの誤差分散（error variance; $V(\theta)$）および標準誤差（standard error; $SE(\theta)$）は、

$$V(\theta) = \frac{1}{I(\theta)} \tag{44}$$

および、

$$SE(\theta) = \sqrt{V(\theta)} = \sqrt{\frac{1}{I(\theta)}} \tag{45}$$

で求めることができる。したがって、θ の95% 信頼区間は以下の通りである。

$$\theta - 1.96 \times SE(\theta) \leq \theta \leq \theta + 1.96 \times SE(\theta) \tag{46}$$

あらかじめ IRT によって情報関数や標準誤差を明らかにしておくことにより、想定される学力層の受験者の個人差を識別するのに適したテストを構成することができる。

5.7　IRT によるテスト開発のメリットとデメリット

これまで述べてきたように、IRT によってテストを開発することで異なる複数のテスト結果を比較することができるようになる。したがって、テストの表面的な特徴（e.g., 難易度）にとらわれず、異なるテストを受けた二人の受験者の能力を比較することや、時期をずらして異なるテストを受けた一人の受験者の能力の変化を追うことが可能となる。また、二人の受験者が異なるテストを受けても能力の比較が可能ということは、受験者のカンニング（cheating）に代表される学業不正（academic dishonesty）の防止にも役立つ。さらに、コンピュータを利用すれば、受験者の回答に応じて項目を適宜出題することもできるので、θ の値が収束したところでテストを打ち切ることが可能となり（コンピュータ適応型テスト ; computer adapted testing)、時間の節約にもつながる。

しかしながら古典的テスト理論と比べると、IRT は構成概念の一次元性や項目間の局所独立など条件（制約）が多く、分析の適用範囲が狭まってしまうのも事実である[*21]。また、基本的に IRT は検定試験や筆記試験などのテストデータに利用されており、面接や小論文などのパフォーマンス評価の分析には向いていない[*22]。

6　結語

テスト理論の発展により、今まで漫然と実施されてきたテストの性能や、評価された能力の精度が批判的に検証できるようになってきたのは事実である。テストを実施する者は想定される受験者の学力層に配慮しながら、もっとも精度よく受験者の能力や個人差を測定できるよう項目を作成、選択すべきである。そうすることで、たとえば学力テストにおいては受験者の評価や選抜に一定の理論的、技術的根拠を与えることが可能となり、説明責任を果たすことにもつながる。

しかしながら、テスト理論の技術的な面に注意を向けて、測定精度の高いテスト開発にばかり心血を注いでいるばかりでは本末転倒といわざるをえない。テストの精度の改善に取り組むことに夢中になりすぎると構成概念の本質を見失うおそれがある（忠実度と帯域幅のジレンマ)。テストによって何（構成概念）を測りたいのか、その構成概念からどのような現象が導かれるのか（予測)、テスト項目は数的にも内容的にも構成概念を適切に反映しているのかなど、テスト開発者の理論的アプローチが必要なことはいうまでもない。テスト理論による分析と構成概念についての理論的アプローチは心理測定、評価においてどちらも欠くことのできない車の両輪のような関係にある。

7　引用文献

Brennan, R.B. (2001). *Generalizability theory*. New York: Springer.

McDonald, R.P. (1978). Generalizability in factorable domains. “Domain validity and generalizability”. *Educational and Psychological Measurement*, 38, 81-105.

注

*1 これらの条件を満たすことを、弱同族測定と呼ぶ。

*2 当然ながら、この例において各項目は数学の学力という上位の構成概念を反映するものである。つまり数学の学力を高く有する者ならば、(低い者に比べて）どの項目にも正答する確率が高く、必然的に項目間に正の相関関係が生まれるということが前提にある。

*3 これをタウ等価測定と呼ぶ。

*4 ただし $j \neq j'$ である。

*5 因子分析を実行するには、有償または無償の専門的な解析用ソフトウェアが必要となる。

*6 これを信頼性の低下による相関の希薄化という。

*7 残念ながら従来の古典的テスト理論の枠組みでは、項目の取捨選択における標本依存性の問題を解決することはできない。この議論については“項目反応理論”を参照さ

れたい。

* 8 この指標が値が高ければ高いほど簡単な項目であり、困難度というよりは容易度を示しているので、以下では通過率と呼称を統一する。

* 9 相関係数の値は -1.0～1.0 の範囲をとる。-1.0 に近づくほど負の相関係数強くなり、二つのデータは負の比例関係、つまり一方が増加すれば他方が減少するようになる。反対に1.0 に近づくほど正の相関関係が強くなり、二つのデータは一方が増加すればもう一方も増加するようになる。無相関である0.0 付近では、二つのデータ間に関連性が見られないことを示す。

*10 この例では 2 値の正誤データである項目と多値データであるテスト得点との相関をとっており、厳密には点双列相関係数（point biserial correlation coefficient）という。

*11 教員採用試験における集団面接のように、受験者にはフリートークをさせておいて、その様子を見た採点者が複数の観点から評価するといった方法もある。

*12 実験計画法の文脈では独立変数と呼ぶ。

*13 水準が決定したらいよいよ実験を行うことになるが、各水準内における被験者数（分散分析の文脈では繰り返しの数という）を数多く確保しておくことが重要である。被験者数が少ない場合には、被験者自身の能力の違いや運の良し悪しなどの影響が強く表れ、処遇の影響を検出できなくなる恐れが生じるからである。

*14 このように、被験者、項目、および採点者の三相を含むテスト形式を *ptr* デザインと呼ぶ。

*15 正反応 = 1 、誤反応 = 0 とする。

*16 性格検査などでは位置母数と呼ぶほうがふさわしい。

*17 ただし、当て推量確率については選択肢の数に応じて分析者があらかじめ固定する場合もある。

*18 ただし、項目間の局所独立が前提である。

*19 3 PLM では一致しない。

*20 ただし、項目間の局所独立が前提である。

*21 近年では多次元 IRT や段階反応モデル（graded response model）が開発されており、少しづつ状況が改善しつつある。

*22 パフォーマンス評価の分析に IRT を利用しようという試みもみられる。

Received : May, 10, 2018
Accepted : June, 13, 2018