項目反応理論についての説明書(学生版)　Ver1.1

共用試験事後評価解析委員会
試験信頼性向上専門部会

はじめに

共用試験を受験される皆さん。この説明文書は共用試験 CBT (Computer Based Testing) で用いられている項目反応理論(Item Response Theory, 略して IRT)を応用した評価の概要について, 分かりやすく解説したものです。

まず, 項目反応理論という用語がわかりにくいかもしれません。この「項目(Item)」とは試験を構成する1つ1つの問題のことです(以下個々の試験問題については項目と表記します)。「反応(Response)」とはその項目に正答するか誤答するかの状況を意味するものです。項目反応理論は, 項目の特性(項目の難しさや識別力(後述))が判明している場合, その項目に対する反応(解答状況)を用いて, 当該試験の結果から測定できる能力を推測するものです。大規模試験の項目作成・実施・評価・運用のための優れた実践モデルとして世界的に定着しています。

この文書では, まず, 共用試験でどうして項目反応理論を用いるようになったかについて述べます。次に, この理論の基本となる成績の求め方について, 簡単な実例を用いて解説します。特にただ正答した項目の数を(時に重み付けて)足し合わせて採点する通常の素点とは異なる方法であることに注目してください。そして最後に共用試験で項目反応理論を用いるメリットと大学に返却している成績指標について述べます。更に, この文書の付録として, 項目反応理論の考え方の基本となる「尤度(ゆうど)」について詳細に述べることにします。

１．共用試験で項目反応理論が用いられる背景

皆さんご存じのように, 共用試験は, 医学生・歯学生が患者さんに接する臨床実習を開始する前に, 必要不可欠な医学・医療・歯学・歯科医療の基本知識・技能・態度を客観的に評価するために実施されます。臨床実習の開始時期は各大学のカリキュラムによって異なるので, 同一の方法を用いて同一日時に全国一斉に試験を実施することは困難です。そこで, 異なる時期に, 異なる場所で, 異なる項目の試験を実施しても公平な評価が得られる試験方式を採用する必要があります。医学・医療・歯学・歯科医療の知識及びその応用や理解力についての評価を行う CBT では, これまでの試行期間を含めて過去に出題した項目について, 複数回の内容のチェックと解答状況の解析を行い, 良質かつ適切(臨床実習開始前の難易度として)と考えられる項目を継続的に蓄積(プール化)してきています※1。CBT 実施時には, このプールした項目バンクの中から, 受験生ごとに異なった項目がランダムに抽出されて出題されます。異なる項目が出題されるための不公平をなくすために, 皆さん一人一人に出題される項目セット間の難しさの差をできるだけ小さくなるように設定しています。

※1 後の受験生の成績を正確に判断するためには, 受験生が解答した問題の守秘が必要です.

しかし，この方法だけでは，一人一人のテストの難しさの差を完全になくすことはできません。そこで，皆さんに出題された項目の特性を考慮して項目セット間の難しさの差に影響されない成績評価の方法として項目反応理論を用いることとなったのです。

２．項目反応理論の原理解説

２－１．項目の難易差と人の能力差を切り離して考える理論

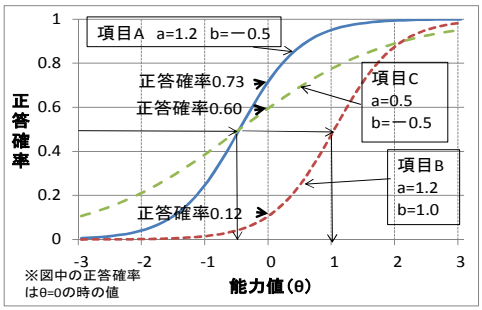
通常，物の長さを測るときは物差しを使います。身長 180cm といえば，それは誰が測っても同じ長さを意味します。体重計であればどの体重計に乗っても同じ針を示すはずです。そうでなければその物差しや体重計は役に立ちません。しかし能力を測るテストではそう簡単ではありません。あるテスト X で 80 点取った人と別のテスト Y で 70 点取った人とでは，80 点取った人の能力の方が高いとは言い切れません。それはテスト Y の項目が平均してテスト X の項目よりも難しかった可能性があるからです。このように，出される項目が違うテストの得点（素点）では，項目の難易差と人の能力差の二つが互いに錯綜し見分けがつかないのです。

この問題を解決するには，項目の難しさや識別力（後述）といった測定に使用した項目側が持つ固有の特性と受験生の能力や学力を示す値とを切り離して考える工夫が必要です。そうした求めに応じて生まれたのが項目反応理論です。それは従来のように単に正答した項目の数を数えて受験生の能力を推定するのではなく，全く別の，しかしより科学的な角度からアプローチしていく方法です。項目反応理論は，近年コンピュータの性能アップに伴い欧米を中心に一つの標準的な方法として広く利用されるようになったテスト理論です。TOEFL，日本の情報処理技術者試験の一つである「ITパスポート試験」などにも用いられています。

２－２．曲線を用いての項目特性の評価

次に，その原理を簡単な例で分かりやすく説明してみましょう。薬効検定などで使われる bioassay の手法などを連想しながら考えれば分かりやすいかもしれません。図 1 で横軸は能力を示す尺度 θ（以下能力値(θ)と表記）を表すことにします（薬効検定で言えばたとえば殺虫剤の濃度に相当）。左から右に行くにしたがって高い値を示します。ただ，ここでは便宜的に 0 点から左右に無限に伸びた尺度を考えています（注:図１は－3～＋3の範囲に限定して表示しています）。また縦軸は能力値が θ の人のその項目に正答できる割合を示すものとします（殺虫剤ならその濃度で何%の虫を駆除できるか，駆除率は濃度が高くなるほど 1 に近づきます。耐性の強い虫は同じ薬なら濃度の高いものが必要です。また同じ虫でも効きの悪い薬では濃度を高くしないと効きません）。

図1．3つの項目の項目特性曲線



ここで３つの項目 A，B，C があるとしましょう。答える人の能力が低い間は，どの項目も正答で

きる人の割合は低いですが，能力が高くなるにしたがって正答できる人の割合は増え，だんだん 1 に近づいていきます（殺虫剤なら濃度が増えるにしたがって駆除率も高まります）。しかしその様子は出される項目（薬剤の種類）によって異なります。項目 A では早い段階で正答できる人の割合が増えていきますが（つまりやさしい項目），項目 B では能力がだいぶ高い人でないと正答できません（難しい項目）。また，項目 A と B では能力がある段階まで達すれば急に正答できる人の割合が増えますが，項目 C では，はじめから正答できる人が 1 割くらいいるのですが，その伸びは緩やかでゆっくりとしか増加しません。こうした特徴は項目ごとに異なり，それをその項目の特性曲線とよびます。曲線の形は項目ごとに違いますが，共通する特徴として，多くの場合 S 字型曲線（Sigmoid curve）をしています。

２－３．項目の特性を表す２種類の特性値（ａとｂ）

共用試験 CBT では，2 つの項目特性値 a と b で形の決まるロジスティック曲線が項目特性曲線として使われています。図 1 には項目 A, B, C ごとに項目特性値 a と b の値が示されています。それらの曲線の増え方を見ると，いずれも始めは緩やかですが，だんだん傾きが大きくなり，ある点から後は逆に増加率が減少に転じ，やがて正答率が1に近づくことが分かります。その曲線の変わり目（変曲点）の位置を決めるのが b であり，そこで増加率は最大となります。また，そこでの傾きを示すのが a となります。b が大きいということは能力値（θ）が高くないと正答できないという意味であり，項目の難しさを決める項目困難度と呼ばれます。a が大きいということは，b を挟む能力値（θ）の前後で正誤が比較的はっきり見分けやすいという意味で項目識別力とも呼ばれます。各項目の a と b の値が分かれば，それぞれの項目の特性曲線の形が決まります。それが図 1 に示される 3 つの曲線です※2。a と b の値がわかっていれば，各受験生がどの項目に正答でき，どの項目に正答できなかったかの解答パタンを調べ，その情報を基に受験生の能力を表すのにもっともふさわしい能力値（θ）を推定することができるようになります。

２－４．解答パタンの利用

ここでは A, B, C の 3 つの項目を例として考えましたが，従来の採点法では 1 問もできなかった人，1 問だけできた人，2 問だけできた人，3 問ともできた人と 4 通りの解答しか取り上げませんでした。しかし，○と×で示した各項目への正誤の解答すべての組み合わせを考えると全部で $2^3=8$ 通りもの異なる解答パタンがあります。もし n 問のテストであれば，その解答パタンは 2^n 通りが可能です。そこで起こり得るすべてのパタンを考慮すればその情報量は天文学的に増えていきます。これを利用しない手はありません。

※2　これを数式で表すと　$P(\theta)=1/[1+\exp\{-1.7a(\theta-b)\}]$　となります.
なお，項目反応理論には，項目特性値 2 個以外に，項目困難度のみの特性値1個，当て推量を加えた特性値 3 個のもの，あるいは別の関数形を使うものなどありますが，詳しいことは参考図書など参照ください.

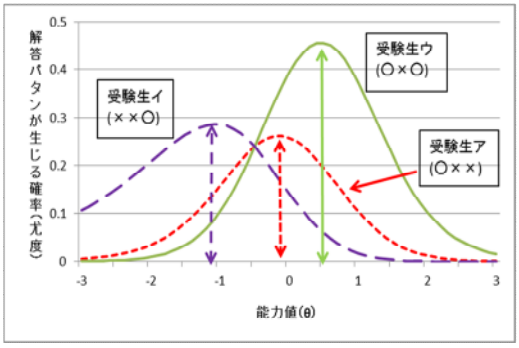
２－５．最も可能性の高い能力値（θ）を推測する方法（尤度関数の利用）

さて，ここで，図 1 に示す 3 つの項目を受けたア，イ，ウという 3 人の受験生の成績について調べてみることにしましょう。受験生アはやさしい項目 A のみ正答(○××)，受験生イは項目 A と同じ難しさの項目 C のみを正答しました(××○)。受験生ウはこれらのやさしい項目 A と C の両方に正答しましたが，難しい項目 B は誤答しました(○×○)。そこでまず受験生アから得られる尤度（ゆうど，付録に詳細を示す）を計算してみることにします。この受験生が平均的能力を持っているとしたら（能力値（θ）＝0），項目特性曲線から項目 A に正答する確率は図 1 に見るとおり 0.73，項目 B は 0.12，項目 C は 0.60 ですので，解答パタン（○××）が得られる確率の計算は，項目反応の独立を仮定すると，(0.73)・(1－0.12)・(1－0.60)＝0.26 となります。同様に，もう少し能力が高い（能力値（θ）＝1)場合の確率は 0.10，もう少し能力が低い(能力値（θ）＝－1)場合の確率は 0.16 となり，受験生アは，この 3 つの能力値（θ）のなかでは 0 が最も高い確率を示すことになります。これをもっと細かく連続した能力値(θ)で計算し，それを能力値（θ）の関数と見たもの，すなわち尤度関数を求めたのが，図 2 です。

２－６．素点との違い

これを見ると，受験生アは能力値（θ）が 0 より多少低いレベルの－0.1（図中両向き矢印）のところに最も大きな尤度(以下最大尤度)がみられます。同様に難しさは同じですが，識別力の低い項目 C のみに正答した受験生イ（××○)が最大尤度をもつ値（能力値（θ）の最尤推定値）は，－1.1 と低い値となります。一方，受験生ウ(○×○)は難しさの程度が同じレベルの項目 A と C には正答できましたが，難しい項目 B は誤答しており，最大尤度は 0.46 で最も高くなります。しかし，推定される値（能力値（θ）＝0.5）は受験生ア（能力値（θ）＝－0.1）とそれほど大きな差は見られません。このように受験生アと受験生イでは，正答数だけ数える従来の素点方式では同じ1点として「差無し」と判定されますが，項目ごとの解答パタンの違いを考えて求めた能力値（θ）では，実は1の開きがあり，受験生アの能力値（θ）はむしろ 2 問正答できた受験生ウに近いことが分かります。

図２．解答パターンと尤度関数



ただ注意しなくてはならないのは，ここではテストの項目が僅か 3 問しかありませんので，能力値（θ）の推定値として最尤推定値を用いるとしても，正確さは乏しく，それがその人の真の能力値（θ）を表すとは限りません。図 2 を見ても分かるように，尤度関数の広がりは大きく，推定の誤差は大きいかもしれません。しかし，項目数が増えてくれば数多くの解答パタンが生まれ，その中で特定のパタンを得る人の能力値（θ）の範囲は狭くなり，その中でいちばん可能性の高い（尤もらしい）能力値（θ）をその解答者の能力値とすることができます。しかし，その計算は大変な作業で，今日のように高速計算機が発達して初めてできるようになったわけです。

３．項目反応理論を共用試験に用いる利点

３－１．共用試験CBTの成績評価に用いる利点と新作問題をプール問題とする方法

このように項目特性値（項目識別力 a と項目困難度 b）の定められた項目を採点の基礎におくことで，今までのように受験生全員に同じ項目を出さなくても，その解答パタンからその受験生に一番ふさわしいと考えられる能力値（θ）を推定することが可能になります。そのような特徴は共用試験のように出される項目が受験生によって違う場合には大変好都合です。また，一度ある集団に実施したデータからそのときに利用した項目の特性値a，b を計算しておけば，以後項目特性値の分かっていない新しい項目を使用するときも，すでにそれが分かっている既知の項目と組み合わせて出題することにより，既知項目の解答状況から推定される受験生の能力値（θ）を使って，逆に新しい項目特性値 a と b を求めることもできるようになります。

３－２．全く初めての状態で項目特性値を求める方法

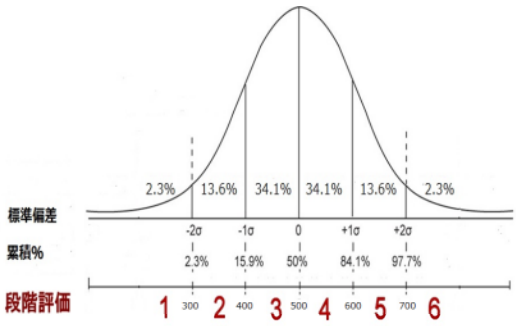
では，項目特性値 a，b も受験生の能力値（θ）も分かっていないまったく初めの状態のとき（第1回～4回のトライアルの時期も同様であった）どうするか。それにもいくつかの方法が考えられますが，詳しい手続きはここでは省略するとして（参考図書5，6，7，8，9など参照），簡単に一つの考え方を説明すると，一度多数の大規模受験者集団（これを基準集団といいます）を対象に実施した大量のデータに対して，正答数などをもとにした仮の能力値（θ）を決め，その能力値（θ）を使って最も当てはまると思われる項目特性値 a と b を推定します。今度はその a と b を使って所与のデータに一番当てはまる能力値（θ）を求めます。その次はさらにその能力値（θ）を使って，それに一番当てはまる a と b を決める。こうした作業を次第に安定した値に近づくまで繰り返すことによって，最終的な a，b，能力値（θ）のセットを手に入れます。

このようにして，一度基準となる大集団に実施した項目群とそれに対する受験生群の解答から得られたデータを基礎に，受験生の能力値（θ）と項目特性値を定義しておけば，それを元に以後の受験生の能力値（θ）と新作項目の特性値を関連させて順次定義付けていくことが可能となります。

３－３．基準集団を用いた成績評価の方法

そうした基準集団として医学系共用試験では 2012～2014 年度の 3 年間の受験集団を，また歯学系共用試験では 2013 年度の受験集団を基準集団として設定しています。そして能力値を便宜的に定めた平均 0，標準偏差 1，の尺度上に定義付けています（この分布が正規分布すると仮定しています）が，共用試験では平均 500，標準偏差 100 の尺度に変換して表すことにしています。これを「IRT 標準スコア」と呼ぶことにしています。そして，それ以後の年度のデータは，経年的変化も分かるように，この基準集団で定められた能力値（IRT 標準スコア）に合わせる形で表示するようにしています。

３－４．返却されている成績指標



最後に，各大学に返却されている IRT 標準スコアの表記について触れておきましょう。基準集団をもとにした IRT 標準スコアと IRT 標準スコア 6 段階評価の 2 種類が掲載されています。IRT 標準スコア 6 段階評価は正規分布を仮定した基準集団の能力値の分布より標準偏差を用いて 6 段階に区分けしたものです（図 3）。段階 1 は IRT 標準スコアが 300 未満，段階 2 は 300 以上～400 未満，段階 3 が 400～500，段階 4 が 500～600，段階 5 が 600～700，段階 6 が 700 以上としています。この段階区分により，受験生の基準集団での概略的な位置が分かります。

４．まとめ

項目反応理論は試験項目ごとの難易度と識別力等の項目特性を考慮して，いつでも，どこでも同等の評価を可能とする方法であり，共用試験の評価には最も適する方法といえます。

５．参考図書　（発行年度順）

1）R. K. Hambleton, H. Swainathan, and H. J. Rogers（1991）. *Fundamentals of Item Response Theory*. SAGE.

2）池田 央（1994）. 現代テスト理論 朝倉書店

3）大友賢二（1996）. 項目応答理論入門 大修館書店

4）W. J. van der Linden, and R. K. Hambleton (1997) . *Handbook of Modern Item Response Theory*. Springer.

5）F. B. Baker and Seock-Ho Kim(2004). *Item Response Theory:Parameter Estimation Techniques 2nd edition, Revised and Expanded*. Marcel Dekker.

6）de Ayala, R. J. (2009) . *The theory and practice of item response theory*. The Guilford Press.

7)村木英治（2011）. 項目反応理論（シリーズ〈行動計量の科学〉） 朝倉書店

8）豊田秀樹（2012）. 項目反応理論[入門編]【第 2 版】 朝倉書店

9）加藤健太郎・山田剛史・川端一光（2014）. Ｒによる項目反応理論 オーム社

６．付録＜尤度について＞

コインの表と裏の関係と各項目の正答誤答の関係は非常によく似ています。例えば 2 回コインを投げる(2 つの項目のテストを行う)と両方とも表(2 問とも正答)，両方とも裏(2 問とも誤答)，一方が正答他方が誤答(正・誤と誤・正)のパタンができます。両方とも表の確率は 0.5×0.5＝0.25，どちらか一方が正答する確率は 0.5×(1－0.5)＋(1－0.5)×0.5＝0.5 になります．この確率の計算ではコインの表がでる確率が 0.5 であるとして計算しています．項目反応理論では，このコインの表裏の確率に相当するものとして項目の識別力と困難度を 1 問 1 問計算して，解答パタン毎の確率を計算します．更に能力値毎に解答パタン別の確率を求めるより複雑な計算を行っています．

このように項目特性値(a(項目識別力)と b(項目困難度))が分かっていれば，能力値が θ である人がこの項目を解いたとき，それぞれの解答パタンが得られる可能性(確率)を計算することができます。項目反応理論では，この確率を能力値(θ)の関数としてみているのでこの確率を尤度(ゆうど)と表現します。たとえば能力値(θ)=－1 の人が ABC という 3 つの項目(本文中の図 2)に○○×の解答パタンで解答する確率は，各反応が独立であるとすると，項目特性曲線からおよそ(0.27)・(0.02)・(1－0.40)=0.003 となり，こういう状態はほとんどあり得ないことが分かります。一方××○と答える確率は(1－0.27)・(1－0.02)・(0.40)=0.286 であり，8 つの解答パタンの中では，×××を除き，最も可能性が高い解答パタンとなります。他の能力値(θ)の人についても同様に計算ができます(表 1)。実際の試験の場合には能力値が分かっていることはまれでありますので，逆に，解答パタンから能力値を推定する事になります。各解答パタンの尤度関数を図 2 に示してあります。

この図より，表 2 に示すように各解答パタン別に最も可能性の高い能力値が推定できます．これを能力値(θ)の最尤推定値としております．IRT 標準スコアは，この尤度を利用して計算しています．

表 1. 能力値(θ)別 8 種の解答パタン(図 2 の 3 問の項目特性曲線から作られる)と反応確率(尤度) (黄色はその中でいちばん高い値で, 最大尤度の近似値)

θ	×××	○××	×○×	××○	○○×	○×○	×○○	○○○
−3.0	0.888	0.005	0.000	0.106	0.000	0.001	0.000	0.000
−2.5	0.831	0.014	0.001	0.152	0.000	0.003	0.000	0.000
−2.0	0.745	0.035	0.002	0.208	0.000	0.010	0.000	0.000
−1.5	0.616	0.080	0.004	0.263	0.000	0.034	0.002	0.000
−1.0	0.437	0.158	0.007	0.286	0.003	0.103	0.005	0.002
−0.5	0.239	0.239	0.011	0.239	0.011	0.239	0.011	0.011
0.0	0.093	0.257	0.012	0.142	0.033	0.393	0.018	0.051
0.5	0.025	0.195	0.009	0.059	0.070	0.456	0.021	0.164
1.0	0.005	0.104	0.005	0.018	0.104	0.373	0.018	0.373
1.5	0.001	0.040	0.002	0.004	0.112	0.220	0.010	0.611
2.0	0.000	0.012	0.001	0.001	0.094	0.102	0.005	0.786
2.5	0.000	0.003	0.000	0.000	0.069	0.041	0.002	0.884
3.0	0.000	0.001	0.000	0.000	0.048	0.016	0.001	0.935

注 : 表の能力値 θ は分かりやすいように 0.5 刻みと大まかに設定してある. 従って, 実際の最大尤度とは異なるが, 方法の理解のために示してある.

表 2. 8 種の解答パタン別正答数, 最大尤度, 対応する能力値(θ), IRT 標準スコア

	×××	○××	×○×	××○	○○×	○×○	×○○	○○○
正答数	0	1	1	1	2	2	2	3
最大尤度	—	0.26	0.01	0.29	0.11	0.46	0.02	—
対応するθ	−3.0未満	−0.1	−0.1	−1.1	1.4	0.5	0.5	3.0以上
IRT標準スコア	200未満	490	490	390	640	550	550	800以上

医療系大学間共用試験実施評価機構　事後評価解析委員会
試験信頼性向上専門部会　委員（平成 26 年度）

荒木　孝二　（東京医科歯科大学医歯学教育システム研究センター教授，公益社団法人医療系大学間共用試験実施評価機構理事，歯学系 CBT プール化専門部会部会長，共用試験制度・システム開発委員会委員長）
石田　達樹　（公益社団法人医療系大学間共用試験実施評価機構事業部長）
池田　央　（日本テスト学会元理事長）
植野　真臣　（電気通信大学大学院情報システム学研究科教授）
大久保　智哉　（大学入試センター研究開発部助教）
大西　弘高　（東京大学医学教育国際研究センター講師，公益社団法人医療系大学間共用試験実施評価機構医学系 CBT 事後評価解析小委員会事後評価専門部会委員）
後藤　英司　（横浜市立大学名誉教授，公益社団法人医療系大学間共用試験実施評価機構理事，医学系 CBT 事後評価解析小委員会委員長）
齋藤　宣彦　（公益社団法人医療系大学間共用試験実施評価機構副理事長，医学系 CBT 実施小委員会委員長）
嶋田　昌彦　（東京医科歯科大学歯学部教授，公益社団法人医療系大学間共用試験実施評価機構歯学系 CBT 実施小委員会委員長）
高木　康　（昭和大学医学部教授，公益社団法人医療系大学間共用試験実施評価機構理事，医学系 CBT 事後評価解析小委員会副委員長，医学系 CBT 事後評価解析小委員会問題プール化専門部会部会長）
前川　眞一　（東京工業大学大学院社会理工学研究科教授）
三谷　昌平　（東京女子医科大学医学部教授，公益社団法人医療系大学間共用試験実施評価機構医学系 CBT 実施小委員会副委員長）
仁田　善雄　（公益社団法人医療系大学間共用試験実施評価機構研究部長）