ニューラルテスト理論－資格試験のためのテスト標準化理論

荘島宏二郎（大学入試センター研究開発部）

1．テストの解像度

いま，ある体重計を用いて A_1 さんの体重を測定したら 73kg だったとしよう．A_1 さんのまさにそのときの体重は，72kg でもなく 74kg でもなく，まさに 73kg であると我々は通常疑わない．一方で，あるテストを用いて B_1 さんの学力を測定して 73 点だったとき，その 73 点は，72 点でもなく 74 点でもなく，まさに 73 点であると信じることができるだろうか．このことは，測定道具としてのテストが，その測定精度について，体重計に及ばないことを意味している．

また，体重計を用いて A_2 さんの体重を測定したら 75kg であったとする．そのとき，我々は，A_2 さんのほうが，わずかに「でも確実に」A_2 さんのほうが，A_1 さんよりも重いと信じることができる．一方で，先ほどと同じテストを用いて B_2 さんの学力を測定したら 75 点だったとき，B_2 さんのほうが B_1 さんのほうが，わずかでも確実に学力が高いと信じることができるだろうか．これは，測定道具としてのテストが，異なる 2 つ以上の対象の差の検出力（識別力）において，体重計に及ばないことを意味している．

以上のことは，テストの解像度について考えることに等しい．体重計を用いてたくさんの被験者の体重を測定したら，グラム尺度上で最も重い人から最も軽い人までほぼ間違えることなく並び替えることができる．一方で，テストを用いてたくさんの被験者の学力を測定したとき，学力が最も高い受験者から最も低い受験者まで，間違いなく並び替えることは不可能である．学力が高くても低い得点をとることがあるし，また，その逆もあるからである．これは，家庭の体重計ですら 0.1kg のレベルでほとんど誤差なく測定できるのに対し，テストは 1 点のレベルで誤差なく測定することが難しいことを意味している．つまり，測定道具としてのテストは，その解像度において，体重計に遠く及ばないことを意味している．

古典的テスト理論における信頼性研究が明らかにしてきたように，多くのテストで信頼性が 0.9 を超えることは珍しい．乱暴な言い方をすれば，テストで測定されているもののうち 10%は誤差ということである．したがって，テストは，1 点刻みで十分な測定をできるような測定道具ではなく，受検者を 5〜20 レベルに段階づけるくらいがせいぜいである．

言うまでもなく，テストは，我々の社会に必要な公具である．しかし，テストで測定された結果が，あたかも体重や身長を測定したかのような客観的な指標として，受検者の性質を示す物理指標であるかのように扱う風潮がないとはいえない．テストの身の丈にあった使用を行うためにも，テストは学力を段階評価するための道具として捉えることは 1 つの有効なアプローチである．ニューラルテスト理論（neural test theory, NTT; Shojima, 2008a, 2008b, 2009）は，学力を段階評価するためのテスト標準化理論である．

2．ニューラルテスト理論の概要

2.1 潜在ランク

潜在ランク（latent rank）とは，学力段階を示すものである．したがって，潜在ランク

数はあらかじめ自明ではないため，データを分析する前に，潜在ランク数を決める必要がある．その際に，各種の適合度指標が参考になる．適合度指標は，NTT モデルがデータにどれくらい当てはまっているかを見るものである．仮に，潜在ランク数を 5 のもとで分析した結果，適合度指標が良くなければ，潜在ランク数を 6，7 と増やして分析することが考えられる．一般的に，潜在ランク数を増やすほど，適合度指標はよくなる．

しかしながら, テストの実施者は, テストを実施する上で, 何レベルに分割したらよいかについて, 理論的なターゲットがある場合も少なくない. そのようなときは, 適合度指標が最適な値を示さない潜在ランク数でも, 実質的に分割して意味のある潜在ランク数を設定したほうがテストの目的に適うと思われる.

2.1 項目参照プロファイル

項目参照プロファイル（item reference profile, IRP）は，各潜在ランクに所属する受験者が，各項目にどれくらいの確率で正答できるかを示したものである．図 1 は，ある 35 項目からなるテストを分析したときの，最初の 8 項目の IRP である．また，潜在ランク数が 10 の下での結果を示す.

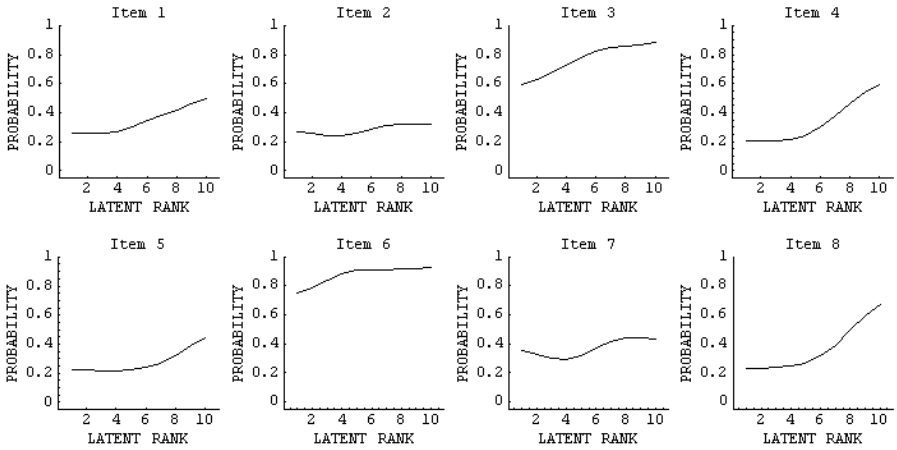


図1: 項目参照プロファイル（item reference profile, IRP）

たとえば，項目 1 は，最も学力が高い潜在ランク 10 に所属する受験者ですら 50%くらいの確率でしか正答できないため，難しい項目と言える．一方で，項目 3 は，最も学力が低い集団である潜在ランク 1 に所属する受検者でも 60%ほどの確率で正答できるため，易しい項目と言える．

また，項目 8 は，潜在ランク 7 あたりで，カーブの勾配が急になっている．このような項目は，潜在ランク 8 以上の受検者を正答させ，潜在ランク 6 以下の受検者を誤答にさせる力を持っている．言い換えれば，この項目に正答できれば，潜在ランク 8 以上に所属する可能性が高く，誤答すれば潜在ランク 6 以下に所属する可能性が高い．したがって，項目 8 は，潜在ランク 7 付近で識別力がた高いと言える．一方で，項目 2 や項目 7 は識別力が低いと言える.

2.2 テスト参照プロファイル

テスト参照プロファイル（test reference profile, TRP）は，IRP の重み付き和であり，

各潜在ランクに所属する受検者が，当該テストにおいて，何点くらいとることができるのかについての期待得点である．図 2 は，上述の 35 項目からなるテストを分析した際の TRP である．

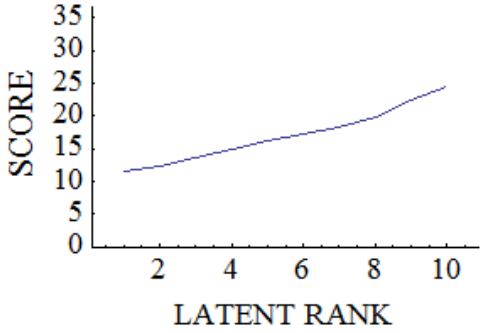


図2: テスト参照プロファイル（test reference profile, TRP）

図 2 では，すべての項目の重みが 1 であるため，縦軸の期待得点は，期待正答数と見ることができる．たとえば，潜在ランク 6 に所属する受検者は，このテストにおいて，およそ 17 問くらいに正答すると期待できる．

TRP は，得られた潜在ランク尺度が，順序尺度であるための根拠として重要である．弱順序配置条件（weakly ordinal alignment condition, WOAC）は，すべての項目が単調増加ではなくても，TRP が単調増加である場合を指す．また，強順序配置条件（strongly ordinal alignment condition, SOAC）は，すべての項目が単調増加であり，必然的に TRP も単調増加である場合である．少なくとも弱い条件を満たさなければ，得られた尺度が順序尺度である根拠が薄弱である．したがって，弱い条件を満たさなければ，異なるパラメタを用いて再分析する必要がある．

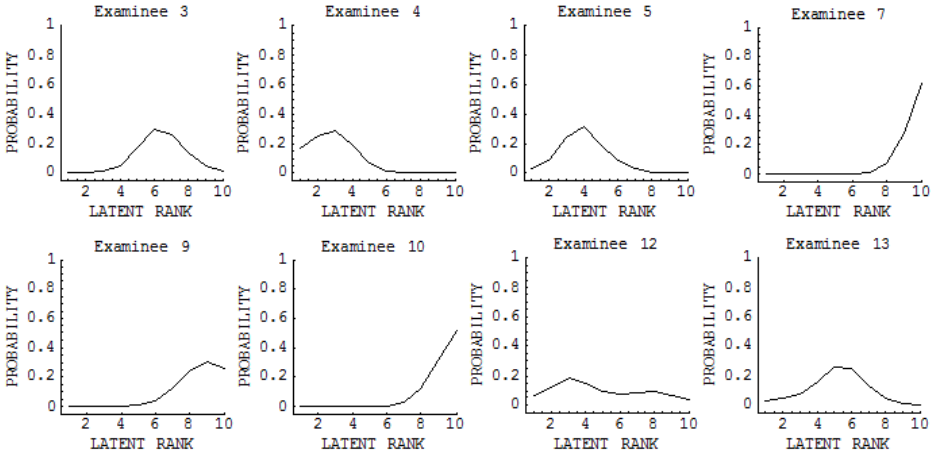
2.3　ランク･メンバーシップ･プロファイル

ランク･メンバーシップ･プロファイル（rank membership profile, RMP）は，各受検者がどの潜在ランクに所属するかについての事後分布である．図 3 では，8 人の受検者の RMP を抜粋した．

たとえば，受検者 3 は，潜在ランク 6 に所属する確率が最も高いので，彼／彼女の所属潜在ランクは，潜在ランク 6 と推定される．また，受検者 7 と 10 は，ともに最高ランクに所属すると推定されたが，よくみると，受験者 7 のほうが潜在ランク 10 に所属する確率が高い．受検者 10 は，潜在ランク 9 に所属する確率も高いので，潜在ランク 9 に落ちそうな潜在ランク 10 と言える．したがって，受検者 7 には「この調子でがんばれ」，受検者 10 には「がんばらないと下のレベルに落ちるよ」などと異なるアドバイスをすることができる．

同様な論理で，受験者9は，潜在ランク9に所属する確率が最も高いが，潜在ランク10に所属する確率も低くない．したがって，このような受検者には「もう少し頑張れば上の段階にレベルアップできる」などとアドバイスすることができる．

さらに，受験者12のように，潜在ランク3に所属する確率が最も高いために，この受検者の潜在ランクの推定値は潜在ランク3であるが，この受検者のRMPの形状は二峰性をもっている．このような受検者は，比較的難しい問題に正答できたわりに，やさしい問題に



誤答した受検者である．したがって，NTTモデルが，彼／彼女の学力が高い可能性と低い可能性の両方を示唆している状況である．このような受検者は，基礎を訓練を十分行う前に応用問題に取り組んだ可能性がある．

このように，RMPは，各個人の教育診断情報として，各受検者やテストの実施者にフィードバックすることができる．

2.4　その他

その他にも，潜在ランク分布，ランク･メンバーシップ分布，IRP指標（熊谷，2007）などの出力がある．

3.　まとめ

NTTのもっとも重要な役割は，can-do chartの作成である．潜在尺度を順序変数にしたことで，各潜在ランクがどのような能力に対応するのかについての記述文（can-do statement）を書くことができる．Can-do chartは，そのような記述文がうまくまとめられ，テストの達成目標やそれに至る道程が要約された図表である．

テストは，ただ実施すればよいものではない．テストの実施者は，各受検者には何ができて何ができないかについての能力カタログと，各学力段階が達成に向けてどのような途上にあるのかについての説明責任がある．Can-do chartを作成することによって，テストの説明責任を向上させることができ，十分に説明されたテストは資格試験と言ってよいだろう．NTTは，can-do chart作成支援ツールと言ってもよい．

References

熊谷龍一（2007）ニューラルテスト理論を離散変数型 IRT とみなしたとき項目特徴を示す指標について　第 1 回ニューラルテスト理論ワークショップ

Shojima, K. (2008a) Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, 08-01.

Shojima, K. (2008b) The batch-type neural test model: A latent rank model with the mechanism of generative topographic mapping. *DNC Research Note*, 08-06.

Shojima, K. (2009) Neural test theory. K. Shigemasu et al. (Eds.) New Trends in Psychometrics, Universal Academy Press, Inc. (pp.417-426).