認知的診断のための測定論的情報を抽出する IRT 分析

○坂本　佑太朗
株式会社リクルートマネジメントソリューションズ

1. はじめに

テストデータの分析へ項目反応理論（item response theory, IRT）（Lord, 1952）を適用することが普及しつつある中で，受検者の反応データから心理測定論にもとづく認知的診断を行うための情報を抽出し，受検者に対して有益なフィードバックを行うことが求められつつある．歴史的には，Carroll & Maxwell（1979）以降，認知的診断を目的とした測定論的研究が行われてきた（Embretson, 1994）．しかしながら，わが国ではそれほど研究の蓄積は十分になく，最近では認知診断モデル（cognitive diagnostic model）（e.g. Leighton & Girel, 2007；Nichols, 1994；Rupp, Templin & Henson, 2010）や認知的 IRT モデル（cognitive IRT model）（van der Linden & Hambleton(Ed.), 1997；Embretson, 1998）を活用した研究（倉元・スコット・笠居, 2003；坂本・柴山, 2014；鈴木・豊田・山口・孫, 2015）がなされているというのが現状であろう．

Embretson（1998）が指摘するように，認知的 IRT モデルの代表例としては線形ロジスティックテストモデル（linear logistic test model, LLTM）（Fischer, 1973）がある．IRT の場合，テスト項目が測りたい構成概念を測るように設計されている際には，項目困難度パラメータがその項目に回答するのに意図された認知的操作（cognitive operations）（認知診断モデル的にいえばアトリビュートやコンポーネント）で説明できると言われている（Embretson, 1994）．LLTM はラッシュモデルにおける項目困難度パラメータを認知的操作ごとの困難度の線形和に分解する．言い換えれば，従来は推定される項目困難度パラメータに関する困難度の「要因」は分析者の定性的な判断や解釈が必要であったが，LLTM ではその「要因」を計量的に明らかにできる点に特徴がある．

しかしながら，LLTM は 1 次元性の仮定を置く心理測定モデルである．この 1 次元性の仮定について，実際には，「1 次元性を満たす」ことを確認する方法自体は多種多様であり（Hattie,1985），Embretson and Reise（2000）は「十分な 1 次元性」の判断基準は明確ではないことを指摘している．また，多くのテスト項目への回答の背後には，多次元性が内在されていると考えられている（Ackerman, Gierl & Walker, 2005；Yao, Boughton, 2009）．そのため，1 次元性を仮定する IRT モデルを多次元に拡張した多次元 IRT（multidimensional IRT）（Reckase, 2009）[1]による分析が期待されているところである．最近では，Reise, Cook and Moore（2015）が 1 次元性を仮定した IRT 分析結果を報告する際，多次元 IRT を使ったテスト項目に関する精緻な分析結果も同時にするべきであると指摘している．特に，ハイステイクスな場面で用いられるテストに関わるテスト開発者/実務家こそ，テストの品質保証

[1] Embretson（1998）は，Embretson（1984），Mislevy & Verhelst（1990），Whitely（1980）を「多次元認知的 IRT モデル（multidimensional cognitive IRT model）」だと指摘しているが，一般にこれらのモデルは多次元 IRT モデルとして包含できる．

という観点から「そもそもそのテストは何を測っているのか」という基本的な問いを常に検証し続ける必要がある．そのため，多次元 IRT を用いた精緻な項目分析は，認知的な情報をテストデータから引き出し，受講者へフィードバックを行う前提として必要であると判断できる．

そこで本発表では，受検者がテスト項目に反応する際に求められる認知的操作に関する情報を得るための線形ロジスティックテストモデル（linear logistic test model, LLTM）（Fischer, 1973）を用いた分析を坂本・柴山（2014）を紹介する．また，多次元 IRT を使った分析として，テストを構成する下位領域（subscales）がもつそれ特有の測定論的な情報を計量的に明らかにする分析（坂本，2015，2016）を紹介する．本稿ではそれらのモデルと分析手法を紹介し，実際の分析例は当日示す．

2. LLTM

古典的テスト理論にもとづいて得られる項目の通過率や1次元性を仮定するIRTモデルでの項目困難度パラメータにおいて，その困難度の要因がどこにあるのかはテスト分析者の推測に頼らざるを得ない．つまり，テスト項目に反応する際に必要となる認知的操作に関する情報は，計量的に表現することができなかった．LLTM は，受検者iの項目jに対する反応をu_{ij}，項目困難度パラメータをb_jとすると，ラッシュモデル（1PL モデル）

$$P(u_{ij}=1|\theta_i, b_j) = \frac{exp(\theta_i - b_j)}{1+exp(\theta_i - b_j)}$$

における項目困難度パラメータb_jを，

$$b_j = \sum_{k=1}^{p} w_{jl}\alpha_k + c$$

と線形和を仮定する．このとき，α_kは基本母数（basic parameters），w_{jl}はそれぞれの項目の基本母数に対する重み（行列），そしてcは基準化のための定数である．基本母数α_kの設定に当たっては，線形構造を仮定している以上，それぞれの基本母数同士が互いに独立であるという仮定を置いていることに注意が必要である．また，LLTM における重み行列$\boldsymbol{W}$が対角行列である場合が通常のラッシュモデルとして位置づけられる（Fischer, 1995）．従来，ラッシュモデルや IRT におけるロジスティックモデルなどでは項目困難度パラメータとして「項目」固有の推定値を得ていたが，LLTM では認知的操作固有の困難度パラメータを得ることができる．

つまり，テスト開発の過程で「その項目がどのような能力（認知的操作）を測るか」を定義している場合には，理論的にはそれを重み行列$\boldsymbol{W}$に落とし込み，それ固有の困難度を推定できる．言い換えれば，受検者の学習上のつまづきが，従来の項目レベルではなく測定領域ごとの困難度として計量的に示すことができる．その結果を踏まえて，今後の学習指導や受検者へのフィードバックをより詳細に行うことや，受検者集団の特徴を測定領域別に把握できる．しかしながら，その適用にあたってはいくつかの条件がある．たとえば，基本母数α_kの数は項目数mを越えてはならず（$p < m$），また$m \times (p+1)$の重み行列$\boldsymbol{W}^+ = (\boldsymbol{W}, \boldsymbol{1})$が階数（rank）$p+1$のフルランクを持つ必要がある（Fischer, 1983, 1995）．つまり，集団統計量を得ることが目的となる大規模学力テストでは，出題範囲が広く設定されるため，項目数mに対して基本母数pが多くなることが予想され，重み行列$\boldsymbol{W}$が設定しづらい．そもそも，LLTM による分析を前提に設計されるテストではあれば，その点に注意して出題範囲等を決定すればよいが，テストデータの二次分析の際には注意が必要である．また，基本母数の推定によって

測定領域固有の困難度を計量的に示すことができるのはLLTMの特徴であるが, 学力テストの場合では，重み行列$\boldsymbol{W}$の設定自体には教科の専門家によるいわば主観的な判断が必要となる．つまり，「テストの専門家」（木村，2006）としての，実際の分析を行う「教育測定（テスト理論）の専門家」とテストの測定領域についての定性的な判断を行う「教科の専門家」の協働が重要となる．

なお, 発表当日にLLTMを実際のわが国で行われた学力調査データに適用した例を示す.

3．多次元IRT

多次元IRTは一般に補償型（compensatory）モデルと非補償型（noncompensatory）モデルに大別される[2]．補償型モデルは，複数の能力を測定するテストにおいて，ある能力が低い場合でも他の能力が十分高ければ当該の項目には正答しやすいという仮定を置くモデルである．つまり，数学的にはそれぞれの次元同士は和の関係にある．補償型多次元2値2PLモデル（以下，多次元2PLモデル）は数式で表現すれば

$$P\left(u_{ij}=1\middle|\boldsymbol{\theta_i},\boldsymbol{a_j},d_j\right)=\frac{exp\left(\boldsymbol{a_j\theta'_i}+d_j\right)}{1+exp\left(\boldsymbol{a_j\theta'_i}+d_j\right)} \quad (1)$$

と表される．このとき次元数をmとすると$\boldsymbol{a_j}$は$1\times m$の項目jの識別力パラメータベクトル，$\boldsymbol{\theta_i}$は$1\times m$の受検者iの潜在特性尺度値ベクトル，d_jは困難度に関連するパラメータ（スカラー）を示している．このとき，d_jは1次元性を仮定した項目困難度パラメータb_jと同じ解釈はできないことには注意が必要である．そこで，多次元IRTの場合には多次元困難度（multidimensional difficulty, MDIFF）

$$MDIFF_j=-\frac{d_j}{\sqrt{\Sigma_{v=1}^{V}a_{jv}^2}}$$

によって同じ解釈が可能となる（Reckase, 2009）．ただし，数学的には，多次元IRTはカテゴリカル因子分析と同値であり，探索的な場合には項目識別力パラメータの単純構造への回転が可能となる（荘島，2003；Takane & De Leeuw, 1987）．

一方，非補償型モデルは当該の項目に正答するにはある能力だけが高いだけでは達成されないことをモデル化しており，数学的にはそれぞれの次元同士の積によって正答確率を定義していることに特徴がある. その代表例としては, Whitely（1980）の複合潜在特性モデル（multicomponent latent trait model，MLTM）がある[3]．MLTMは認知心理学と心理測定学とをつなぐ数理モデルとして提案されてきた背景をもっている（Whitely, 1980；Embretson, 1997）．

測定領域に関する事前の仮説がある場合，確認的な多次元 IRT 分析を実行することも可能である．その代表例としての双因子モデル（bifactor model）（Gibbons & Hedeker, 1992）における項目識別力パラメータは行列で表記すると

$$\boldsymbol{a}=\begin{bmatrix}a_{11} & a_{12} & 0\\ a_{21} & a_{22} & 0\\ a_{31} & 0 & a_{33}\\ a_{41} & 0 & a_{43}\end{bmatrix}$$

となる．これから明らかなように，双因子モデルは全項目へ影響を与える一般因子（general factor）と，それぞれの下位領域からの影響としてグルー

[2]最近では，岡田（2014）が補償型と非補償型を包含する多次元IRTモデルを提案している．

[3] MLTMの発展形として，項目困難度についてLLTMと同じ構造をもつ一般化複合潜在特性モデル（general multicomponent latent trait model, GLTM）（Embretson, 1984）がある．

プ因子（group factor）を定めるモデルである．このとき，一般因子を統制することによってグループ因子による影響を検討できる点がこのモデルの特徴である（Reise, et al., 2015）．

Reise, Morizot and Hays (2007), 坂本 (2015, 2016) はテスト全体が測定する構成概念を一般因子，テストに含まれる下位領域を「次元」とみなし（グループ因子），下位領域特有の影響が一般因子と比較してどの程度存在しているかを検証した．具体的には，IRT における項目情報量の観点からその比較を行っている．通常の IRT モデル（2PL モデル）における項目情報量は，項目jに正答する確率と誤答する確率をそれぞれ$P_j(\theta)$，$Q_j(\theta)$とすれば

$$I_j(\theta) = \frac{\{a_j P_j(\theta) Q_j(\theta)\}^2}{P_j(\theta) Q_j(\theta)} = a_j^2 P_j(\theta) Q_j(\theta) \quad (2)$$

と定義される．(2) 式から，2PL モデルにおける項目情報量は，項目識別力パラメータの値によって規定されることがわかる．多次元 IRT においてはこれを多次元空間に拡張し，空間上の一点であるθ_iとv次元目のθとのなす角をα_{iv}とすると

$$I_j(\boldsymbol{\theta}) = \frac{\{\boldsymbol{a_j} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})\}^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})} = \frac{(P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) \sum_{v=1}^{V} a_{jv} cos\alpha_{iv})^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})} = P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) (\sum_{v=1}^{V} a_{jv} cos\alpha_{iv})^2 \quad (3)$$

として定義できる（Reckase & Mckinley, 1991；Reckase, 2009）．(3) 式から，多次元の場合でも項目識別力パラメータの値によって項目情報量が規定されることがわかる．Reise, et al. (2007) と坂本（2015，2016）は，双因子モデルによって推定した一般因子とグループ因子の項目パラメータの 2 乗値の比較によって，それぞれの構成概念に対してもつ測定論的な情報の比較を項目レベルで行った．これを通して，たとえば一般因子を統制した後にもなお，グループ因子のほうがより多くの情報量を保有している項目については，テスト出題時の項目選択の際に注意すべきであることを示唆している．つまり，通常の IRT モデルが仮定する「1 次元」の構成概念以外の影響が強いことを意味するため，テストが測りたいものを測るという前提からそれてしまっている可能性がある．

なお，発表当日は坂本（2015，2016）の分析例を示す．

4. おわりに

本発表では, 認知的 IRT モデルとしての LLTM, また多次元 IRT を使った分析を紹介した. 認知的診断を行う心理測定モデルの重要性が指摘されている（Wang & Gierl, 2011）が，その前提として「そもそもそのテストは何を測っているのか」という素朴な問いは重要である．つまり，最後はテストの妥当性という観点に回帰するのである．テストの信頼性は妥当性の必要条件となるが，特にテスト開発者/実務家こそテストの妥当性検証を不断に行う必要がある．その際には，宇佐美（2015）が指摘する通り，構成概念妥当性という単一の概念に収斂するそのさまざまな側面（Messick, 1989）のどの側面について議論しているかを明らかにする必要があろう．そのような測定論的前提を踏まえてはじめて，認知診断を目的とした統計的分析を行うことができると意識し続けることが大切となる．

坂本 佑太朗（yutaro_sakamoto@recruit-ms.co.jp）