厚生労働科学研究費補助金(政策科学総合研究事業(臨床研究等 ICT 基盤構築・人工知能実装研究事業))
令和 5 年度　分担研究報告書

医師国家試験トライアル試験 CBT 問題への IRT モデル適用に関する研究

研究分担者　久保　沙織（東北大学高度教養教育・学生支援機構　准教授）

研究要旨

現在の医師国家試験は、毎年新たな問題を準備し、マークシート方式の PBT(paper-based testing)で実施されている。試験後には全ての問題と正答が厚生労働省の Web ページで公開され、回答データの分析、合格基準の設定など、試験の運用は古典的テスト理論に基づいて行われている。古典的テスト理論では、得られる結果がその時々の受験者集団の性質に依存してしまうという問題を抱えている。すなわち、毎年異なる問題で構成される試験を、それぞれ異なる集団が受験している状況では、各年の結果の変動について、試験に含まれる問題の難易度や識別力の違いによるものなのか、受験者集団の能力分布の違いによるものなのか、区別することができない。

IRT(Item Response Theory)は上述のような古典的テスト理論の問題を克服するものである。IRT では、試験を構成する各問題の難易度や識別力といった項目の持つ特性と、受験者の持つ能力とを分離して考えることができる。将来的に、IRT の枠組みで等化された項目プールを用いてテストを運用することにより、事前に難易度や測定精度を制御して一定の質を満たす問題セットを継続的に作成し、実施日程や会場、受験者集団が異なる場合でも、共通尺度上のスコアとして表現することが可能となる。本研究では、昨年度に引き続き、2023 年度 CBT トライアルの回答データに対して IRT による分析を適用した。

A. 研究目的

CBT 化に伴い、IRT による試験運用を目標とした場合、項目の難易度と識別力を表すパラメタ(項目母数)の推定値が得られていて、かつそれらが同じ尺度上に等化されている数多くの項目をあらかじめ用意しておく必要がある。このような項目を集めたものを項目プール(item pool)あるいは項目バンク(item bank)と呼ぶ。IRT による試験運用においては、この項目プールの構築が極めて重要である。本研究では、将来的な医師国家試験の CBT 化を見据え、2023 年度の CBT トライアルで出題された 200 問について、IRT による項目母数の推定を行うことを目的とする。

B. 研究方法

2023 年度の CBT トライアルで出題された 200 問について、A 問題(75 問)、B 問題(50 問)、C 問題(75 問)のそれぞれで、古典的テスト理論に基づく項目分析及び、IRT による項目母数の推定を実行した。まず、各項目の要約統計量を確認した上で、古典的テスト理論に基づき項目困難度(通過率)と項目識別力(item-total correlation: IT 相関)を求めた。その後、IRT の 2 母数ロジスティックモデル(two parameter logistic model: 2PLM)を適用して項目母数(困難度母数と識別力母数)を推定した。なお、IRT の分析においては、項目分析の結果を踏まえて識別力が極端に低い項目を除き、IT 相関が 0.2 以上の項目のみを用いた。

C. 研究結果

1. A 問題（*N*=1356）

全 75 項目の通過率の平均は 0.549 であった。通過率が極端に低かった項目は A62（0.077）、A50（0.085）、A15（0.092）、A31（0.097）、高かった項目は A48（0.956）、A27（0.945）、A64（0.920）であった。IT 相関の最大値は A59 の 0.498 であり、IT 相関が 0.2 を下回った項目は A33（0.032）、A50（0.047）、A12（0.066）など 11 項目であった。信頼性を表すクロンバックのα係数は 0.86 であった。

2PLM による A 問題の項目母数の推定値を図 1 に示した。プロットされている文字は項目番号を表す。一般的に、識別力母数の値は概ね 0.3～2.0 の間で推定されるとされるが、その値が 1.5 以上となったのは A59 のみであり、概して識別力が低めの項目が多かった。困難度母数の推定値は、通過率が低かった A15、A31、A62 などで大きな値となっているものの、全体としては正の項目よりも負の項目の方が多く、易しい項目がやや多く含まれていたことがわかる。なお、IT 相関が 0.2 未満のため IRT 分析の際に除外された 11 項目中 7 項目が、画像や音声、動画を使用した項目であった。

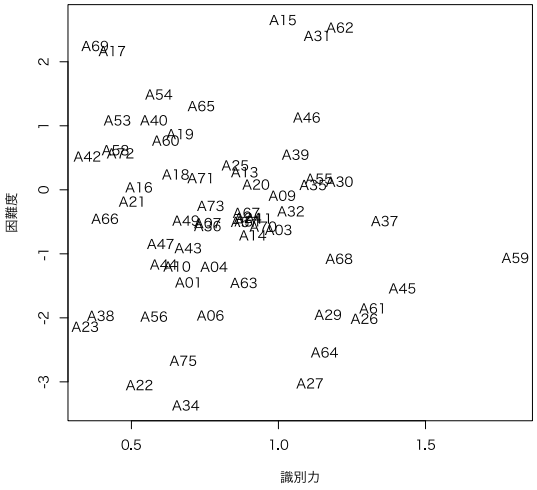

図 1　2023 年度 A 問題の項目母数
（64 項目）

2. B 問題（*N*=1351）

全 50 項目の通過率の平均は 0.660 であり、通過率が 0.2 を下回る項目はなかった。相対的に通過率が低かった項目は B39（0.209）、B42（0.258）、B34（0.295）、高かった項目は B25（0.967）、B31（0.967）、B26（0.929）、B28（0.925）、B02（0.916）であった。項目識別力については、B42 の IT 相関が-0.197 で負の値となっていた。B42 は連問の 2 問目に相当する項目であったが、当該項目は、B 問題全体で測定しようとしている能力とは異なる能力を測定している可能性が高いことが示唆された。他に、B07（0.056）、B04（0.065）で IT 相関が 0.1 を下回っており、識別力が低かった。このうち B04 は選択肢が画像の問題であった。一方で IT 相関が高く識別力が高いと判断された項目は、B24（0.546）、B46（0.498）、B44（0.491）であった。B44 と B46 はいずれも連問の 2 問目に相当する項目であり、B24 は動画を使用した項目、B46 は画像・音声・動画の全てを使用した項目であった。また、B 問題のクロンバックのα係数は 0.75 であった。

2PLM による B 問題の項目母数の推定値を図 2 に示した。古典的テスト理論に基づく項目分析において通過率が最も低かった B39 の困難度母数が最も大きな値となり、IT 相関の高かった B24、B46、B44 の識別力母数の値が大きな値となるなど、一貫性のある結果が得られている。困難度母数が正の値となった項目は 11、負の値となった項目は 29 あり、IRT の分析結果からも B 問題では易しい項目が多かったことが示された。

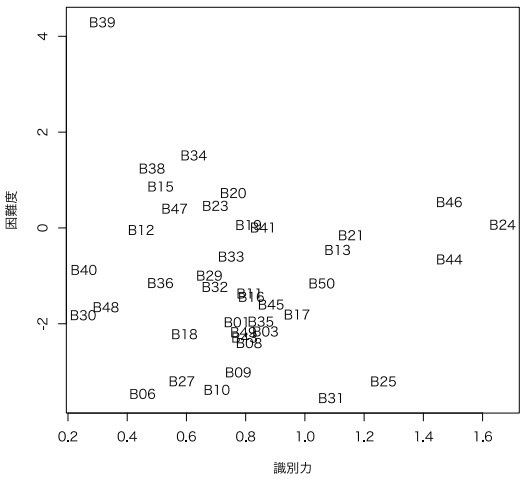

図 2　2023 年度 B 問題の項目母数
（40 項目）

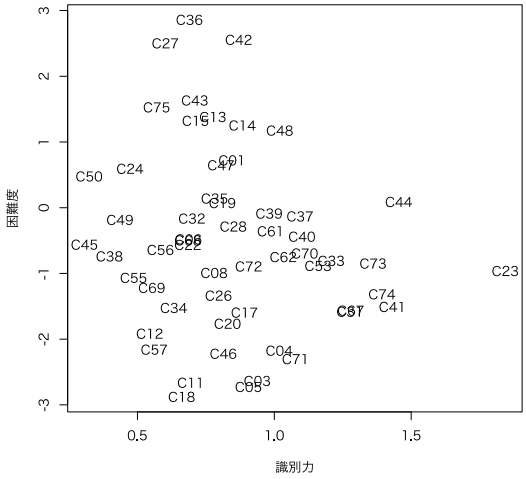

図 3　2023 年度 C 問題の項目母数
（56 項目）

3. C 問題（*N*=1345）

全 75 項目の通過率の平均は 0.593 であった。通過率が低かった項目は C51（0.062）、C42（0.125）、C60（0.140）、C36（0.141）、通過率が高かった項目は C52（0.970）、C66（0.958）、C65（0.936）、C16（0.926）であった。項目識別力については、C09 の IT 相関が-0.009 とほぼ 0 であった。その問題に正答したか否かと合計得点とが無相関ということであり、C09 は、C 問題全体で測定しようとしている能力とは関連のない能力を測定している可能性が高いことが示唆された。他に IT 相関が低かった項目として C54（0.022）、C64（0.069）などがあった。IT 相関が高かった項目は、C44（0.520）、C23（0.469）、C40（0.442）などであった。

2PLM による C 問題の項目母数の推定値を図 3 に示した。先に述べた識別力母数の範囲の目安（0.3～2.0）に照らして、識別力が十分に高いと言える項目は少ないが、困難度母数はおおよそ-3～3 の広い範囲で推定されていることから、易しい問題から難しい問題まで満遍なく含まれていたことがわかる。なお、IT 相関が 0.2 未満のため IRT 分析の前に削除された 19 項目中 10 項目が画像や音声、動画を使用した項目であった。

D. 考察

2023 年度に実施した CBT トライアルの回答データに対し、古典的テスト理論に基づく項目分析と、IRT の 2PLM による分析を行った。本年度は 3 年間の研究事業の最終年度にあたるため、2021 年度及び 2022 年度の CBT トライアルの分析結果も併せて総括する。

まず、古典的テスト理論に基づく項目分析の結果から、識別力を表す IT 相関が 0.2 を下回る項目が毎年いずれの問題でも 2～3 割程度含まれていた。IT 相関が極端に低い（目安は 0.3 以下）項目は、テスト全体で測定を意図している能力を適切に測定できていない可能性が示唆されるため、通常、項目プールには含めない。CBT による IRT に基づく試験の運用においては、項目プールの構築と維持が肝要であり、かつ最も労力のかかる作業である。項目作成の段階では、このように識別力の低い項目が含まれることは一般的であるものの、医師国家試験の CBT 化を目指すにあたっては、質の高い項目を効率的に作成するための体制作りも重要となるであろう。

トライアルの受験者数は 2021 年度が約 320 名、2022 年度が約 450 名だったのに対し、2023 年度は 1350 名程度まで大幅に増加したことで、IRT の項目母数の推

定値が安定した。

IRT による分析結果は、古典的テスト理論による分析結果と整合性を保ちつつも、豊かな解釈を与える。3 年間のトライアルに共通して、A 問題は、困難度、識別力ともに幅広い推定値が得られる傾向が見られ、必修を扱う B 問題は易しい項目が多く、C 問題は困難度のばらつきは大きいが高い識別力が得られにくい傾向が示された。それぞれ、各論、必修、総論を扱っているという内容に鑑みて、妥当な結果と言える。

現時点では、各年度のデータに対し IRT モデルを適用し、個別に困難度母数と識別力母数を求めているが、このままでは項目母数や受験者の能力値に関する年度間比較を共通尺度上で行うことはできない。IRT に基づく試験運用では、作成された項目は共通尺度への等化のプロセスを経てはじめて、項目プールに格納される。2023 年度のトライアル試験は、共通項目計画による等化を企図し、過去 2 年間に出題した問題の一部を含めて実施された。今後、共通項目を精査し、等化を試みる予定である。

E. 結論

3 年間の CBT トライアルの回答データに対し、IRT による分析を行ったことで、IRT に基づく CBT の運用を実現させるための課題及び要件が明確となった。

第一に、多くの良問を効率的に作成するための体制作りの必要性である。問題作成者に分析結果のフィードバックを行い、識別力が極端に低い項目や難しすぎる項目、易しすぎる項目は、なぜそのような結果となったのか、誤答分析等の詳細な分析結果も参照しながら、問題の内容に遡って検証することが求められる。項目プールを充実させるためには、問題作成者と分析者の協働が不可欠である。

第二に、データが IRT モデル適用のための前提条件を満たしているかの検証と、等化の実行である。本来、IRT モデルの適用のためには、データが局所独立の仮定と一次元性の仮定を満たしていることが前提となる。出題領域や問題形式等、医師国家試験という特殊な実施条件の下でどれだけ頑健性が担保できるのか、検証の必要がある。また、等化にあたっては、基準集団の設定等の検討事項も生じる。これらの課題を一つひとつ解決していくことで、医師国家試験の IRT 基づく CBT 化の実現が確かなものになると考える。

F.健康危険情報

特になし

G. 研究発表

1. 論文発表
   特になし
2. 学会発表
   特になし

H.知的財産権の出願・登録状況

1.特許取得
   なし
2. 実用新案登録
   なし
3. その他
   なし