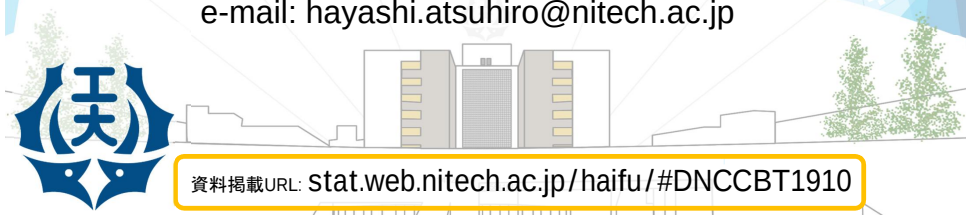
CBT活用有識者会議(40分)
10/25/19 @DNC

数理的見地から観た選抜試験における
CBT・IRTの利活用に関する考察
--- 顕在変量と潜在変量とのハザマで ---

林 篤裕
(名古屋工業大学 社会工学専攻
& 副アドミッションオフィス長)
e-mail: hayashi.atsuhiro@nitech.ac.jp



資料掲載URL: stat.web.nitech.ac.jp/haifu/#DNCCBT1910

1. CBT(Computer-Based Testing): 利点

- 出題にマルチメディアが利用可能
  <===> PBT(Paper-Based Testing)
- 受験者への設問のデリバリーが容易・安価
- 解答の電子化が容易、採点も容易
- 障害者対応も容易かも(文字の拡大、音量の調整等)
- 解答過程のログが取れる(解析するかは別であるが)
- 動的出題(Adaptive, CAT)も実装可能
  <=== 短時間で測定できる。逐次検定的(統計学)。
- e-Learningでの利用が想定される
- (任意の時刻・場所で試験実施が可能)
  <=== 選抜試験には無用

1. CBT(Computer-Based Testing): 欠点

- 膨大なItem Bankが必要
- Item Bankの管理(統計量、履歴、プレテスト等)も必要
- Itemは原則非公開
- 装置不具合の危惧 ＝ DNCリスニング機器(少部品点数)でさえ発生・対応してきた過去(2006年～)
  - 部品点数に比例して故障が発生
  - 大学入試センター試験における英語リスニング試験の試験監督を担当したことのある者なら危惧するはず
  - 紙の頑健性・利便性を凌駕する環境が提供できるのか?
- アジアでの失敗例: TOEFLのItem窃盗(2000年前後)
  - Item自動生成の研究が過去にはあったがその後聞かない

2. Item Bank(Item Pool、項目/設問銀行)

- 利点
  - Itemの的確な管理の下では、質の高い出題が可能
  - 受験者の評価がより正確にできるであろう
  - 教育課程の整備をベースに
    <=== { 医学部共用試験等(コアカリキュラムの存在) / 資格試験や達成度試験 }
- 欠点(前半)
  - 開始時の設問数の確保・その方策。膨大な項目数の必要性。
  - 事前評価(プレテスト)が必須
  - 溜めた設問の秘匿が継続的に保持できるのか疑問。
    曝露の危険性 ＝ アジア諸国の国民性(中国での例)
  - 初出設問(Virgin Item)ではない
    <===> 日本では初出設問だけで実施してきた過去

2. Item Bank(Item Pool、項目/設問銀行)

- 欠点(後半)
  - 日本の特性: 学習指導要領改訂(概ね10年間隔)への対応
  - Itemの自動生成の徒労
  - 一度出題すると同じスキルは対策される(過去問分析の経験)
  - アジアでの選抜試験(ハイステークステスト)の
    導入事例はあるのか? 韓国に構想はあったが
- 共通試験の年複数回実施
  - Item Bank(やCBT)が救世主とは成り得ない。PBTでさえ。
  - Item供給量の圧倒的不足をどう補うつもりなのか?
    - 問題作成の労力は膨大であり、Item Bankを構成できるほどの
      作成(質・量とも)は不可能 ===> 統括官にお問い合わせを
  - 日本で事前評価が可能か? 曝露を危惧しなくて良いのか?
  - 道具立てが在っても、中身が無ければ機能しない・運用不可

3. 顕在変量と潜在変量とのハザマで

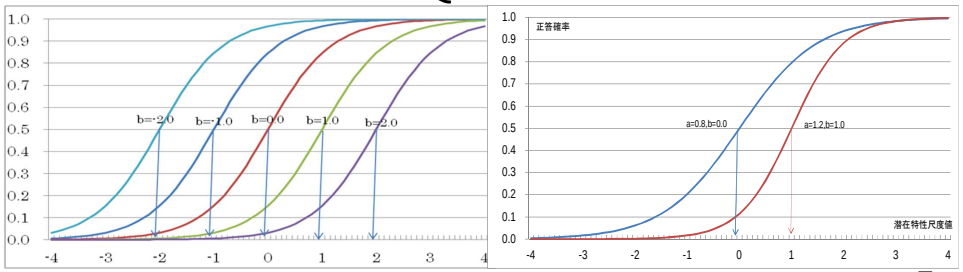
- 顕在変量(Manifest Variable)
  - 回帰分析(RA), 主成分分析(PCA), ...
    (Regression Analysis) (Principal Component Analysis)
  - 線形計算、行列計算で実現: 解は一意に確定
- 潜在変量(Latent Variable)
  - 因子分析(FA), IRT, 潜在クラス分析(LCA), ...
    (Factor Analysis) (Latent Class Analysis)
  - 測定できない。解は複数(無限個)。
    確認の手段がない。恣意性の入り込む余地。
  - 推定したパラメータの真偽は誰ができるのか?
- 「真値＝計算結果」に対する
  考え方・捉え方の違いのように思われる

4. IRT(Item Response Theory、項目反応/応答理論)

- モデル: 非常にシンプル。理解も容易?
- 適用条件(前提条件、強い条件)
  - 単答式の出題を対象としている
  - 個々のItem(設問)の独立性

  } 日本の出題文化としての大問形式は対象外

  - 能力の1次元性 <===
    - 多様性を測ろうとしている時代に?
    - 複数単元が包含された科目には?



(野口 裕之(2017)から引用)

4. IRT(Item Response Theory、項目反応/応答理論)

- [危惧]
  - 単答式の出題＝知識を問う設問になりがちではないか?
  - 思考力を測るのに大問形式は有効な手段(日本の文化)
- 変換点(Scaled Score):
  - 選抜試験における変換点の受容が進むか? : 疑問
  - 資格試験(TOEFL、TOEIC等)での普及で理解が進んだのか??
- パラメータの推定
  - 目的関数(尤度関数)が複雑(多次元多峰性)で最適解が求まる保証がない。擬似的に求まってように振る舞っているだけ。

    ===>
    - モデルの特性
    - 計算技術の問題ではない(EM、MCMC、ベイス、...)
    - 計算手法が改良されたら数値が変わる可能性

4. IRT(Item Response Theory、項目反応/応答理論)

- パラメータの推定
  - 尤度関数を最大にするパラメータθを求める。
  - 解析的には解けず、繰り返し計算で求める必要。
  - Newton-Raphson法等の最適化手法を用いて。初期値依存。
  - 計算機精度にも依存。
  - Local Maxは求まるが、Global Maxが求まる保証はない。
    - θの組み合わせは無限組存在する ===> 一種の違和感
  - IRT: $L(\theta|\boldsymbol{u}) = f(\boldsymbol{u}|\theta) = P(u_1, u_2, \cdots, u_m|\theta) = \Pi_j P_j(\theta)^{u_j} Q_j(\theta)^{(1-u_j)}$

    $\ln L(\theta|\boldsymbol{u}) = \sum_j (u_j \ln P_j(\theta) + (1-u_j) \ln Q_j(\theta))$
  - LCA: $L = f(y;\phi) = \prod_{i=1}^{n} \left\{ \sum_{c=1}^{C} \pi_c f_c(y_i|\theta_c; x) \right\}$

    $\log L = \log f(y;\phi) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c f_c(y_i|\theta_c; x) \right\}$

4. IRT(Item Response Theory、項目反応/応答理論)

- パラメータの推定
  - FA
    - $\underset{(pxn)}{Z} = \underset{(pxk)}{A}\underset{(kxn)}{F} + \underset{(pxn)}{E}$
    - Z: 測定値。pxn行列。顕在変量。
    - A: 因子負荷量。pxk行列。
    - F: 共通因子。kxn行列。
    - E: 独自因子。pxn行列。

    （A, F, E: 潜在変量）
    - k: 縮約次元数。陽に利用者が指定する。
    - $I=TT^{-1}$ を用意すると、$Z=AF+E=AIF+E=ATT^{-1}F+E=A^*F^*+E$
    - 回転の不定性
      - AとFの組み合わせが無限個数存在する ===> 一種の違和感

5. 潜在変量を扱う手法の背景

- IRT: Lord(1952), Lord & Novick(1968)
  - Lord, Frederic M. (1912-2000)
  - Novick, Melvin R. (1932-1986)
- LCA: Lazarsfeld(1968)
  - Lazarsfeld, Paul Felix (1901-1976)
- シンプルなモデル(素朴?)
  - 計算パワーの不十分な時代。モデルの提案として。
  - 今日の学力を表現するにはシンプル過ぎるモデル
  - 実験室に留めておくべきでは? <=== 実用前段階
  - 計算パワーが上がり無理やり計算できる時代だが、真値が求まる保証がなく、また、モデル自身が改善されたわけでもない

- FA: Spearman(1904)
  - Spearman Charles E. (1863-1945)

6. IRT再考

- BILOG-MGやRのパッケージ
  - 真値を提示できているのか? できるのか?
  - ブラックボックス化の怖さ、統計ソフト黎明期(1980年代)
  - 確認の手段があるのか?
- 十分なサンプルサイズがあれば良いという問題ではない: 無いのはそれ以前だが
  - 膨大なパラメータ数があることに加えて
- 一種の鞍点を求めているに過ぎないのでは?
  - 繰り返し計算の宿命

7. IRTの活きる道

- 測定・評価用ツール(教育、学校現場)として
  - 修正が効く。教員がそばで指導できるから。
    多人数にも対応化。教員の支援にも使える。
  - 達成度試験、資格試験: Item Bankも有用だろう
  - 初等中等教育の現場での利活用はすぐにでも
    - 学習指導要領が完備された国＝日本
- 選抜用ツール(入試、合否)としては不完全
  - パラメータの確度が保証できない以上、
    受験生の進路選択を誤る可能性をはらんでいる
  - 誰が責任を取れるのか?
  - IRT(やCBT)は向かない領域と、私は考える

7. IRTの活きる道

- 全くの別物:同一視されがち
  - ローステークステストで実現できていること
  - ハイステークステストでの実現可能性

  ⇔ 独立
- 適材適所で利用できる領域はあるはず
- 「定期試験、教育評価」と「入試、選抜」は
  明確に分離して考えるべき
- 実社会への送り出し方:
  誤解のない利用方法を確立した上で
- モデルの特性や課題を熟知した上での利用
- (法科大学院適性試験の轍を踏まないためにも)

8. まとめに代えて:数理的見地から

- 西欧諸外国: 出口管理の国
  - その国で使えるものが
    日本でもそのまま使えるとは限らない。文化に依存。
  - 入試における共通試験の重みは低い。粒度が低くても利用可能。
- 日本: 入口管理の国
  - 選抜試験: 複数の単元がまとまって一つの科目を構成。大問主義。
  - 日本の文化に根ざした技術の確立・利活用を。
- 峻別: 選抜試験と資格試験は個別に検討
- 潜在変量を影響力の大きい社会に出す恐ろしさ
  - そこまで危険を犯して導入するメリットはあるのか?
  - 危ないと判っていて船出をするのか? 子供たちを危険に曝すのか?
- CBT・Item Bank・IRTとも、将来に向けてさらなる研究を

【参考文献】

- 丘本 正(1986)、「因子分析の基礎」、日科技連出版社。
- 芝 祐順編(1991)、「項目反応理論—基礎と応用」、東京大学出版会。
- 浅野 長一郎 & 江島 伸興(1993)、「潜在構造分析論の現状」、日本統計学会誌 第22巻 第3号、PP357-373。
- 池田 央(1994)、「現代テスト理論」、朝倉書店。
- @saltcooky(2018)、「潜在クラス分析についてまとめて、Rでお試し」、https://qiita.com/saltcooky/items/dc48ca3cefa9c1dfc010 (2019/10/23 閲覧)。
- 宮澤 芳光(2019)、「Computer Based Testing(CBT)を用いたテストの出題」、大学入試センター・アドミッションリーダー研修「入試問題の作成・分析とCBT入門」配布資料、2019年7月20日。
- 寺尾 尚大(2019)、「CBT・CATとは何か」、日本テスト学会第17回大会、公開シンポジウム2「多面的総合的評価・CBT・アクティブ・ラーニング」配布資料、2019年8月29日。