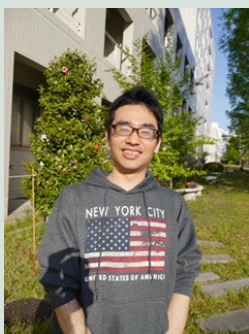




項目パラメタドリフトの検出と 原因の検討・共通尺度への影響 —そのドリフトはどこから来てどこへ行くのか—



寺尾 尚大 / Takahiro Terao

大学入試センター 研究開発部

(terao@rd.dnc.ac.jp)

シンポジウムのテーマと測定理論との接点

- シンポジウムのテーマ

テスト項目の著作権および問題公開／非公開を巡る問題

- さて 教育測定学からどのようなアプローチが可能か？
- **わが国では，以下のような重みづけのように思える**
 - 試験問題公開 → 内容の周知を通じた「身につけるべき力」の共有【媒体】
 - ＞ 測定理論・統計的品質管理・得点の比較可能性
- **試験問題を非公開にして，得点を比較可能にしたとしても，さまざまな「受験者の権利」をめぐる問題が……ある**
 - 項目パラメタドリフトという「測定論的な現象」の話題を経由して，問題非公開時に認めるべき受検者の権利について，多面的に捉えたい

本日の話題提供内容

- **項目反応理論とテストの等化**【初めての方向け】
- **項目パラメタドリフトの概要と検出方法**
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- **ドリフトの原因と方向性**—そのドリフトはどこから？
- **ドリフトの影響範囲**—そのドリフトはどこへ？
- **分析事例の紹介**
- **ドリフト検出時の対応**
- **まとめ**

本日の話題提供内容

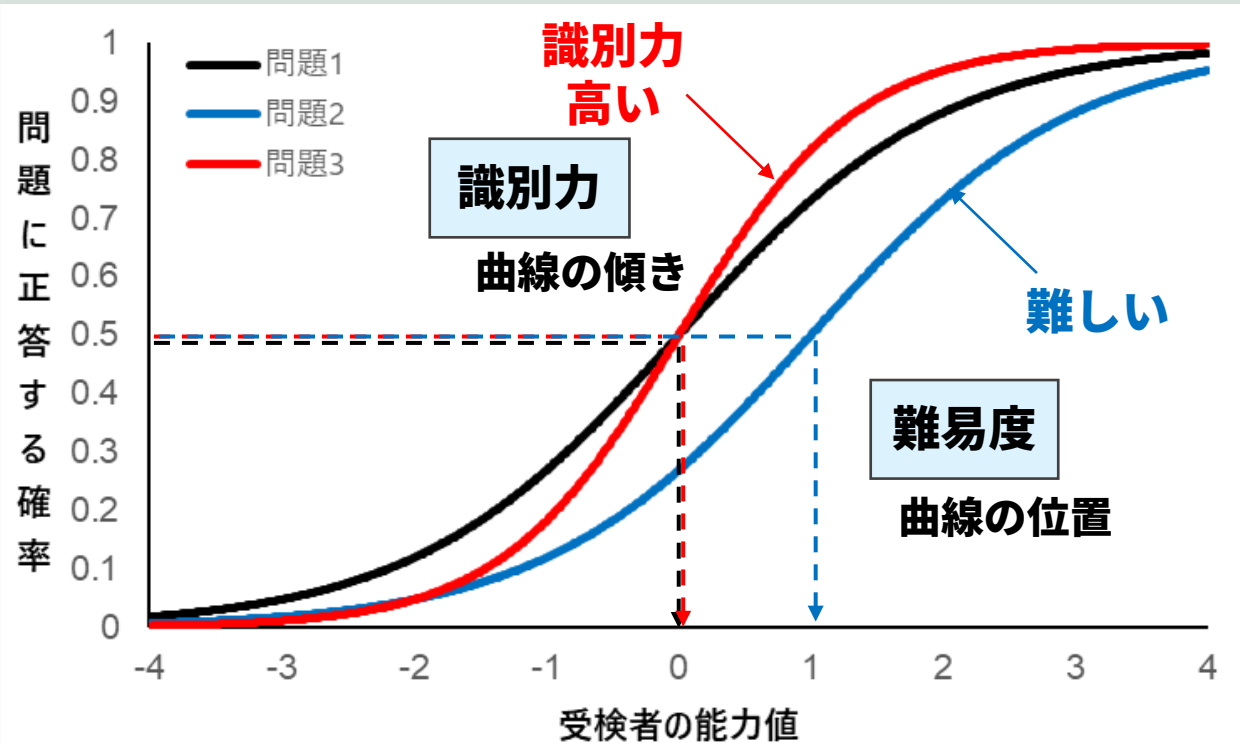
- **項目反応理論とテストの等化**【初めての方向け】
- **項目パラメタドリフトの概要と検出方法**
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- **ドリフトの原因と方向性**—そのドリフトはどこから？
- **ドリフトの影響範囲**—そのドリフトはどこへ？
- **分析事例の紹介**
- **ドリフト検出時の対応**
- **まとめ**

項目反応理論（項目反応モデル）

- **項目＝テスト問題，反応＝解答**
- **項目への正誤（得点の高低）の背景要因をパラメタ化**
 - 受検者の能力の高低（能力高 → 正答，能力低 → 誤答）
 - 項目特性
 - 難易度（困難度）低 → 正答，難易度高 → 誤答
 - 識別力 低 → 能力が高くて正答できるとは限らない
能力が低くても誤答するとは限らない
 - 識別力 高 → 能力が高いと，正答しやすい
能力が低いと，誤答しやすい
 - （記述式やパフォーマンス評価の場合）評価者特性，採点基準など
- **混合物をいくつかの単体に分解して推論するためのモデル**

項目特性曲線

受検者の能力値と正答確率の関係を 項目ごとに描いた曲線



難易度
(b)

正答確率が半々（50%）となる能力値はどこか？

- 曲線が左側にある (問題1) = 易しい
→能力値が低いところで50%の正答確率になる問題
- 曲線が右側にある (問題2) = 難しい
→能力値が高いところで50%の正答確率になる問題

識別力
(a)

難易度付近での正答確率の変化幅はどの程度か？

- 傾きが緩やか (問題1) = 識別力低い
→能力値が少し高い値に変化しても、正答確率はあまり変わらない問題
- 傾きが急 (問題3) = 識別力高い
→能力値が少し高い値に変化したとき、正答確率も高い問題

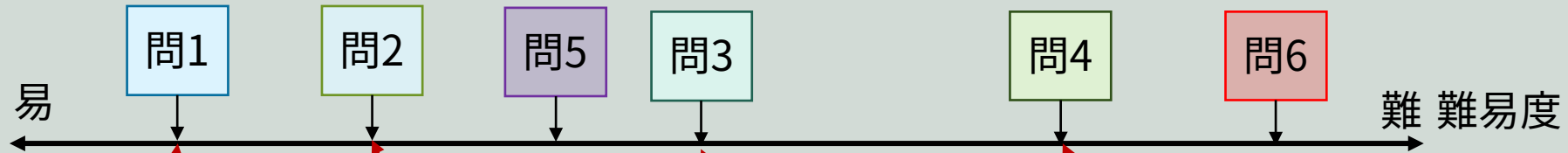
$$P(\theta_i | a_j, b_j) = \frac{1}{1 + \exp(-Da_j(\theta_i - b_j))} \quad D = 1.702 \text{ (定数)}$$

※ 能力値については（間隔尺度と考えて）
受検者にフィードバックする際、受け入れられやすいよう線形変換を行うことが可能

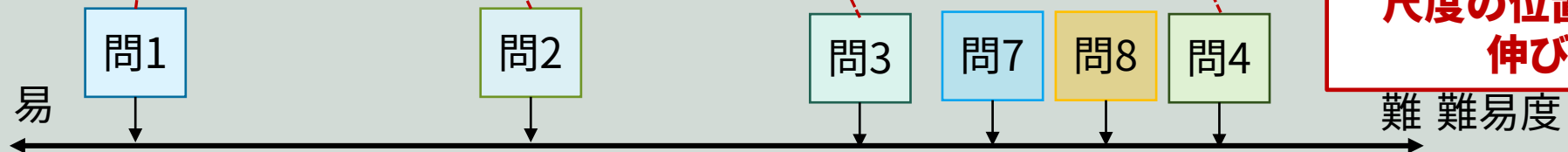
等化・対応づけ

多数の項目があると、ピッタリ一致した等化係数を求めることは難しくなる → できるだけ誤差が小さくなるように求める

第1回の試験 (202Y年6月) ※基準集団



第2回の試験 (202Z年6月)



等化

$$b_{2j}^* = Ab_{2j} + B$$

$$a_{2j}^* = \frac{a_{2j}}{A}$$

尺度の位置の移動と
伸び縮み

※ いったんこの試験内に閉じた項目パラメタが得られる

※ 共通項目デザインの例

項目パラメタを活用した等化

- **基準集団を決める方法**

- 特定の実施回を基準のものさしとして，以降は基準へ等化
- 学力調査などで使用されている

- **アイテムバンクを構築する方法**

- 難易度・識別力のパラメタと，テスト項目をセットにして，大量にデータベースに蓄積する → 巨大な基準尺度ができる
- プレテストで得られた項目パラメタを使用し，出題時に能力値を推定する

アイテムバンクの活用可能性

- **複数等質テスト**

- テスト全体の測定精度が等しくなるように項目を選んでテストをつくる (シャドーテスト法, 最大クリーク法, 整数計画法, ...)

- **適応型テスト**

- 項目の正誤に応じて, 次に出題する問題を変える
- 能力値の更新幅が一定程度小さくなったところで出題をやめる

- **マルチステージテスト**

- 大問単位で適応的に出題する ; 大問内では固定項目セット
- あらかじめ決めたルーティンルールに沿って出題

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- **項目パラメタドリフトの概要と検出方法**
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- **ドリフトの原因と方向性**—そのドリフトはどこから？
- **ドリフトの影響範囲**—そのドリフトはどこへ？
- 分析事例の紹介
- **ドリフト検出時の対応**
- **まとめ**

項目パラメタドリフトとは何か

- 基準集団を決める方法・アイテムバンクを用いる方法のどちらでも，項目パラメタは固定の値 (所与) として得点[※]を算出

※ 能力値 θ のこと



時間が経過していく中で，難易度が徐々に動いていくことはないのかな？
また，試験問題が秘密に漏れていたとしたら，能力の高くない人も正答できることになり，知らないうちに識別力が低くなりそうだな。

- 項目パラメタの経時的変化 = **項目パラメタドリフト**

- 時間経過による難易度・識別力の変化の大きさ
- ただし，標本変動の大きさを考慮する

(同一の特徴をもつ別の集団に解答を求めたとき，困難度や識別力の推定値は多少揺らぐ。
そのときの揺らぎ (= 標準誤差) 以上の大きさを，項目パラメタドリフトとして検出する。)

項目パラメタの不変性 (item parameter invariance)

- 項目パラメタを用いた等化の前提

項目パラメタは，推定時点と使用時点で値に大きな変化がない。

- 項目特性関数から考えても，項目パラメタの値を所与にしている

$$P(\theta_i | a_j, b_j) = \frac{1}{1 + \exp(-Da_j(\theta_i - b_j))} \quad D = 1.702 \text{ (定数)}$$

難易度 (b) と識別力 (a) の値がそれぞれ与えられた下で，正答確率を能力値 θ の関数として描く

- もちろん，ピタリ一致は難しい

- あまりに変化量が大きいと，得られる能力値に影響するかも？

項目パラメタドリフト研究の黎明

- **新しい考え方ではない；1980年代から指摘**
 - **Lord (1980)：**
項目パラメタの不変性に言及
 - **Mislevy (1982)：**
アイテムバンクの管理と項目露出がドリフトの原因になりうることを指摘
 - **Goldstein (1983)：** IPD研究の冒頭でよく引用される

If we consider the two times as part of a single time period, then **the Rasch model, or any other model that assigns parameter values to items, requires that the characteristics of all the items during that period remain constant**, otherwise the assumptions of the model are not met. Yet by hypothesis, **the A and C items, but not the B items, change their characteristics or parameter values and, hence for these items, there can be no unique calibration and the model cannot therefore be true, over the whole time period.** (Goldstein, 1983; p.373)

ドリフト検出手法に関する二つの潮流

- 項目パラメタドリフトに特化した分析
 - CUSUM (cumulative sum) 法
- 特異項目機能 (Differential Item Functioning, DIF) の分析の応用
 - デルタプロット (delta-plot) 法
 - Mantel-Haenszel 法 (分割表に基づく分析)
 - 尤度比検定
 - Lordのカイ二乗統計量
 - 項目特性曲線の差 (の和) を面積として考えるもの
 - Rajuの方法, Index-K (熊谷, 2012)

同一項目を同一の受験者が
2回以上解答することはないと仮定



実施回の違いを
独立な集団と見なした分析の
適用が可能

CUSUM法

- 項目パラメタのズレの大きさの和を累積的に評価

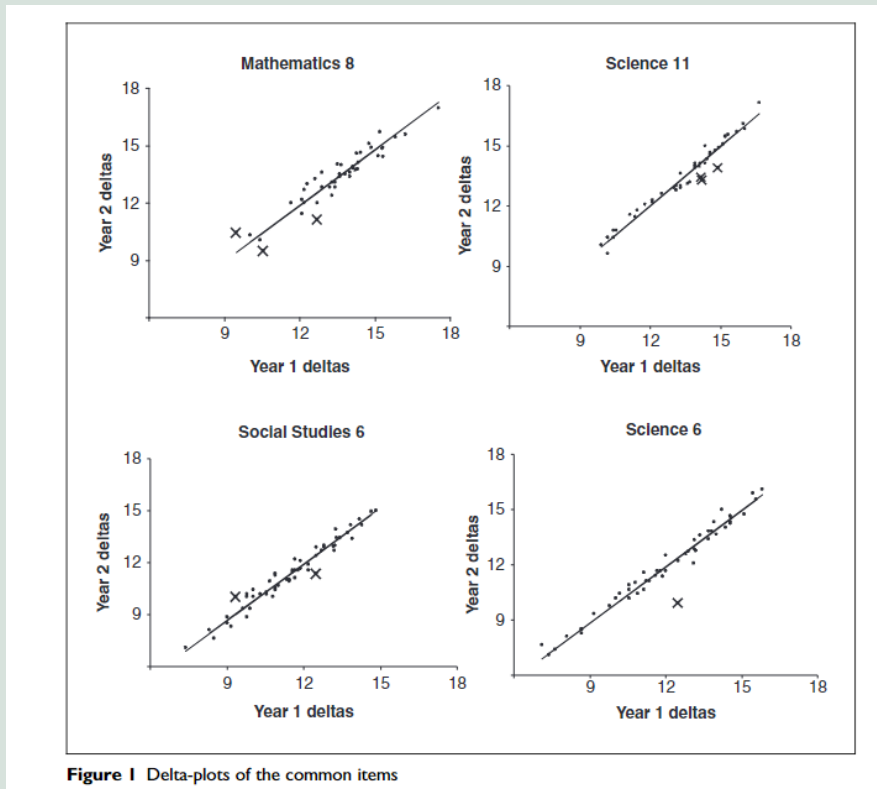
$$S_j(t) = \max \left(\underbrace{S_j(t-1)}_{\text{これまでの累積的な“ズレ”}} + \underbrace{\frac{a_{j0} - a_{jt}}{SE(a_{j0} - a_{jt})}}_{\text{推定時点 (t=0) からの識別力パラメタのズレ}} + \underbrace{\frac{b_{j0} - b_{jt}}{SE(b_{j0} - b_{jt} | a_{j0} - a_{jt})}}_{\text{推定時点 (t=0) からの難易度パラメタのズレ}} - k, 0 \right)$$

Glas (2001) の式を参考

- 推定時点 (t=0) の難易度 b_{j0} , 識別力 a_{j0}
- ズレの合計が k 以下であれば対応しない (次へ持ち越さない)
- $S(t)$ が h 以上の値になったら, ドリフトが生じたと見なす
- Veerkamp (1996) では, $k = 1/2, h = 5$ として評価
- 統計的検定の枠組みの範疇, unidirectionalなドリフトの検出

デルタプロット法

- ETSが独自に使用する簡便な難易度指標 Δ を，各回について算出して同時プロットを描く



$$\Delta(p_j) = 13 - 4\Phi^{-1}(p_j)$$

- ドリフトが起きていなければ，デルタも大きくは変わらない想定
- Figure中の×印は，フラグのついた項目（要確認）

DIF分析でよく用いられるものを応用

• Mantel-Haenzel 法

- 得点に基づいて K 群に分割
- 各群で，推定時点と今回の正誤人数のズレを検討し，総合して評価

第 k 群における分割表

	正答	誤答
推定時点 (Reference)	n_{1R}	n_{0R}
今回 (Focal)	n_{1F}	n_{0F}

$$\alpha_{MH} = \frac{\sum_{k=1}^K \frac{n_{0F}^{(k)} n_{1R}^{(k)}}{n^{(k)}}}{\sum_{k=1}^K \frac{n_{1F}^{(k)} n_{0R}^{(k)}}{n^{(k)}}}$$

• 尤度比検定

- モデル0:
推定時点と今回で
項目jの項目パラメタは等しい
- モデル1:
推定時点と今回では
項目jの項目パラメタは等しくない
(別々の値として推定する)

• Lordのカイ二乗統計量

$$Q_j = \frac{(b_{jR} - b_{jF})^2}{\hat{\sigma}_{jR}^2 + \hat{\sigma}_{jF}^2}$$

難易度のズレ評価
2PLM以上でも発想は
同じ

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- 項目パラメタドリフトの概要と検出方法
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- **ドリフトの原因と方向性**—そのドリフトはどこから？
- **ドリフトの影響範囲**—そのドリフトはどこへ？
- 分析事例の紹介
- ドリフト検出時の対応
- まとめ

原因として考えられるものは何か

- **カリキュラムの改訂や項目内容の変化**
 - 特定の内容を重点的に教えるようになった → 易しくなる
※ 能力分布のほうにも変化が生じる可能性はある
 - 項目内容が常識化/陳腐化 → 易しくなる／難しくなる
 - ニュースなどの影響を一時的に受けることも
- **項目漏洩・露出回数の増加**
 - 項目漏洩 (問題を非公開とする試験なら重大インシデント)
→ 異常に易しくなる, 識別力が下がる
 - 項目露出回数の増加 → 徐々に易しくなる

項目内容の変化の例：漢字テスト

- **昔と今で項目パラメタが変化していそうな項目** (あくまで予想)
 - 漢字の読み
 - 忖度（そんたく）：昔は難易度高，今はそれほど
 - 逼迫（ひっぱく）：コロナの影響で難易度低下？
 - 未曾有（みぞう）：首相の言い間違いで話題になったが，今は難易度戻った？
- **能力分布の形とは関係ない** (きわめて僅少な影響にすぎない) **が，個々の項目単位では変化しうる**
 - 一時的な変化なのか，長期にわたる安定的な変化なのかの見極めが必要
 - ただ，一時的な変化は（実務上は）見過ごされがち

一方向性 vs. 双方向性

- **一方向性のある場合 (unidirectional)**
 - 易しく変化する
 - 項目漏洩や露出の影響？重点的に教えるようになった？
 - 難しく変化する
 - あまり重きを置いて教えなくなった？
 - 識別力の低下
 - 能力値が低くてもわかるようになった／能力値が高くてもピンとこない
- **双方向性のある場合 (bidirectional)**
 - 易しい方向の変化と難しい方向の変化がテスト内で入り混じる

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- 項目パラメタドリフトの概要と検出方法
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- ドリフトの原因と方向性ーそのドリフトはどこから？
- **ドリフトの影響範囲**ーそのドリフトはどこへ？
- 分析事例の紹介
- ドリフト検出時の対応
- まとめ

能力推定への影響

- **意外なことに，能力推定への影響は小さい**
 - **Wells et al. (2002)**
一方向のドリフトが生じていた場合にも，人数・ドリフト項目の割合などによらず能力推定値への影響は小さかった
 - **Witt et al. (2003)**
項目パラメタの分布・能力分布が多少歪んでいても，ドリフト項目ばかりでなければ能力推定値への大きな影響は見られなかった
 - **Huang et al. (2003)**
ドリフトが起こっていた項目を取り除いても，能力推定値への影響は僅か
- **共通項目でドリフトが起こった場合には，影響が出る**
 - **Kopp & Jones (2020) :**
プレテストのサンプルサイズが小さい場合には能力値への影響がある

等化係数への影響

• Han et al. (2012)

- 4種類の項目パラメタドリフトが，等化係数に与える影響を検討

$$b_{2j}^* = Ab_{2j} + B \text{ のとき}$$

mean-mean法

$$B = \bar{b}_1 - A\bar{b}_2 \quad A = \bar{a}_2 / \bar{a}_1$$

mean-sigma法

$$B = \bar{b}_1 - A\bar{b}_2 \quad A = s_{b_2} / s_{b_1}$$

Stocking-Lord法

テスト特性曲線 (TCCs) の差が小さくなるAとBを求める

Inward IPD (内向き)



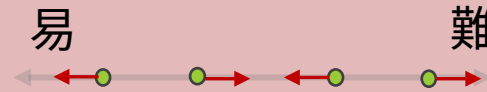
平均変わらない
標準偏差小さくなる

Outward IPD (外向き)



平均変わらない
標準偏差大きくなる

Even IPD



平均変わらない
標準偏差やや大きくなる

Local IPD



平均変わらない
標準偏差
あまり変わらない

• 結果

- 最も簡便なmean-mean法で，どの方向のドリフトの場合であっても等化係数を精度よく推定できた
- 双方向のドリフトがあることで，影響が相殺されるケースも見られた

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- 項目パラメタドリフトの概要と検出方法
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- ドリフトの原因と方向性ーそのドリフトはどこから？
- ドリフトの影響範囲ーそのドリフトはどこへ？
- 分析事例の紹介
- ドリフト検出時の対応
- まとめ

Bock, Muraki, & Pfeiffenberger (1989)

- 物理テストで難易度パラメタにドリフトが検出された
 - 力学分野において項目が易くなる方向に変化
力学の基本概念に，指導の重点が置かれたためと考察している

- 単位系の変化

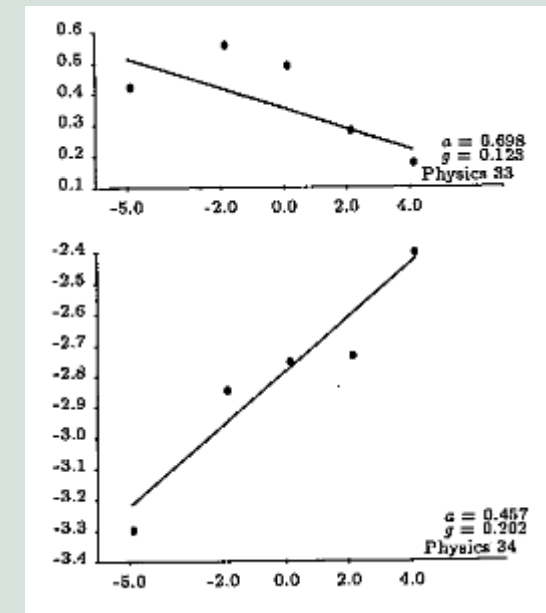
- 1975年 メートル法転換法 (Metric Conversion Act)
→ 1977年前後から，教科書にメートル単位系
- Physics 33：キログラムを用いた問題【易化】
Physics 34：ポンドを用いた問題【難化】



1975年
フォード大統領署名



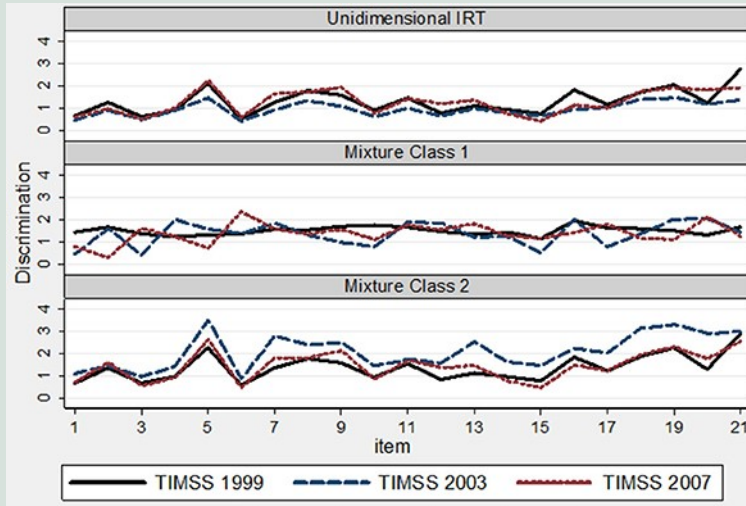
1991年
ブッシュ (父) 大統領後押し



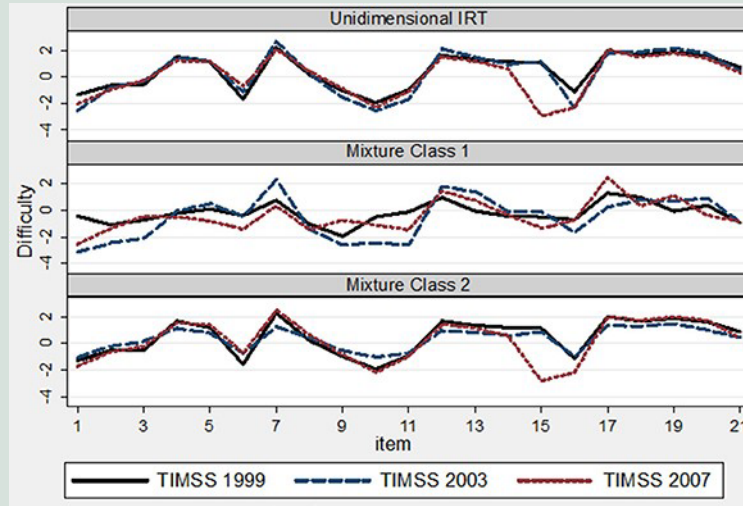
Park, Lee, & Xing (2016)

• TIMSSにおける3時点の項目パラメタドリフトの検討

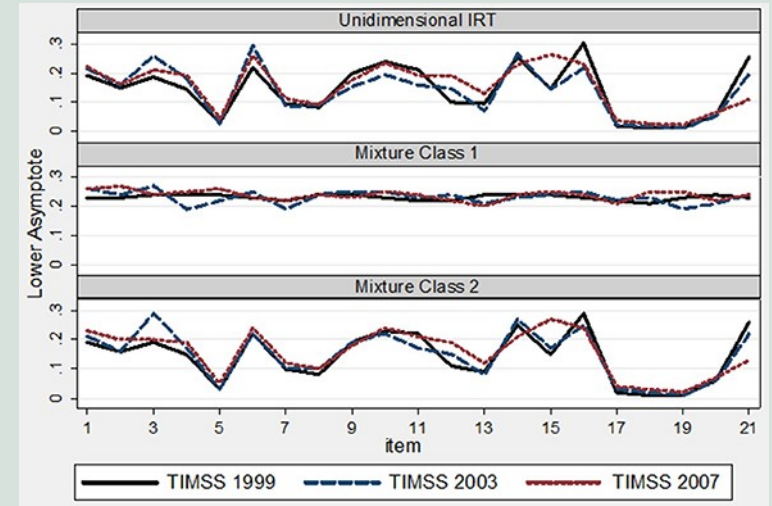
識別力パラメタ



難易度パラメタ



当て推量パラメタ



- 全体としてはドリフトがないように見えても、
潜在クラスに分けてみると見えてくるドリフトがある
- 系統的なドリフトというよりは，“乱高下”と言ったほうが適切

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- 項目パラメタドリフトの概要と検出方法
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- ドリフトの原因と方向性ーそのドリフトはどこから？
- ドリフトの影響範囲ーそのドリフトはどこへ？
- 分析事例の紹介
- **ドリフト検出時の対応**
- まとめ

出題の見合わせと精度のはざままで…

- **ドリフト検出時の対応の理想は“出題見合わせ”**
 - その問題を出題の第一線から退場させる
→ 統計量の詳細分析，漏洩の可能性の検討，カリキュラムとの関係の分析等を尽くす
- **ただし，そう一筋縄でいかない；ドリフト対応の要否を検討**
 - どれぐらいの大きさを超えたら，対応を検討しなければならないか？
 - (テストデザインにもよるが) 共通項目でドリフトが生じていたら，どう対応するか？ その項目を取り除いて分析するか，含めたままにするか？
 - 得点の解釈基準の変更を検討するレベルか？ (standard settingしている場合)

... the perplexing issue in evaluating item parameter drift is not just **how to identify it**, **but how to handle it** (Han et al.,2012; p.116)

項目パラメタのOnline Calibration

- **Online calibrationの前提**

- 本物の受検者に対して，推定済みの問題と未推定 (新作) 問題の両方を解答する
- 「推定済み問題が大半」「ある1時点を基準に推定」が前提

- **Makransky & Glas (2014) の発想**

- **連続更新法 (continuous updating strategy)**

- マルチフェーズ法の拡張
- 項目が露出するたびに項目パラメタを随時再推定して更新する
→ 尺度が常にフレッシュな状態になる (ドリフトを残さない)

この方法に対する法的な問題点

- **ハイステークスな試験では法的な問題に発展しうる**
 - 試験実施後に項目パラメタの値を再推定して変更することは、受検者がもつ“同一尺度上で得点が表示される権利”を損なう
 - 裁判で勝つのは難しいだろうとの見解
(“This strategy … would be hard to defend in the court …,” Fink et al. 2018, p.329)
- **もちろん，ハイステークスでなければ…OK？**
 - 学力調査では，個々のスコアを推定することではなく，推算値を用いた集団統計量（平均・分散・各分位点など）の推定が目的
- **テスト理論だけでなく，さまざまな観点から試験・学力調査への波及効果を検討する必要がある**

本日の話題提供内容

- 項目反応理論とテストの等化【初めての方向け】
- 項目パラメタドリフトの概要と検出方法
 - 特異項目機能 (Differential Item Functioning, DIF) 分析との関係
- ドリフトの原因と方向性ーそのドリフトはどこから？
- ドリフトの影響範囲ーそのドリフトはどこへ？
- 分析事例の紹介
- ドリフト検出時の対応
- まとめ

受験者の権利

- 試験問題を非公開として尺度化ができた場合でも，その先に受験者の権利についていろいろな検討事項がある

私は，試験を2回受けたときに自分の得点が正しく比べられないのはイヤだな。
あまり時間変化に左右されずに，一定期間，同じ問題は同じ難易度・識別力として
取り扱ってほしい。



私は，ドリフトが確認され次第，難易度や識別力を新しいものに更新してほしいかな。
難しくなった問題をそのままにされたり，測定の精度には関係なくなった問題が出たりして，
場合によっては自分の得点が低く見積もられちゃうこともあるわけだね。

- どのような考え方が受容されるか，丁寧に調べる必要がある

不易流行

不易を知らざれば基立ちがたく
流行を知らざれば風新たならず 松尾芭蕉

- **能力測定・尺度構成の考え方の基本 = “不易”**
 - 受検者の能力値がいつでも比較可能であること、あまり時代の影響を受けない項目を出題することが基本
- **能力・項目の中身が時代に応じて変わる = “流行”**
 - まずは項目パラメタドリフトとして現れるのではないか
 - ある程度、断続的な変化が見られそうなら、能力分布も変化している可能性を検討する
 - いつまでも古い能力尺度のままでは、見たいものを見損ねる
- 「その本は一つなり」

不変の中に変化を取り入れていくバランスの良さが求められる

もちろん、実際にはいろいろな観点があって苦労するが…

<https://edo-g.com/wremarks/view/72>



ご清聴 ありがとうございました

試験実施機関等で丁寧な分析・検討をされている皆様に
敬意を表します

文献案内

- Bock, R. D., Muraki, E., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25(4), 275–285.
- Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60(3), 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf
- Glas, C. A. W. (2001). CUSUM statistics for large item banks: Computation of standard errors. *Law School Admission Council Computerized Testing Reports*, 98-11. https://research.utwente.nl/files/5130131/LSAC_CT-98-11.pdf
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20(4), 369–377.
- Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education*, 25(2), 97–117. <https://doi.org/10.1080/08957347.2012.660000>

文献案内

- Huang, C. Y., and Shyu, C. Y. (2003). The impact of item parameter drift on equating. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, Chicago.
- Kopp, J. P., & Jones, A. T. (2020). Impact of item parameter drift on Rasch scale stability in small samples over multiple administrations. *Applied Measurement in Education*, 33(1), 24–33. <https://doi.org/10.1080/08957347.2019.1674303>
- 熊谷龍一 (2012). 統合的DIF検出方法の提案—“EasyDIF”の開発—. *心理学研究*, 83, 35-43.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- Makransky, G., & Glas, C. A. W. (2014). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11(1), 1–20.
<http://www.jattjournal.com/index.php/atp/article/view/48350/39220>

文献案内

- Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement*, 34(5), 365–369.
- Mislavy, R. J. (1982). Five steps toward controlling item parameter drift. *Paper presented at the annual meeting of the American Educational Research Association*, New York.
- Park, Y. S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*, 7:255.
<https://doi.org/10.3389/fpsyg.2016.00255>
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26(1), 77–87.
- Witt, E. A., Stahl, J. A., Bergstrom, B. A., & Muckle, T. (2003). Impact of item drift with non-normal distributions. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago, IL.