全国的な学力調査のCBT化検討WG（第8回）
令和3年3月30日　10:00－12:00
Web会議

資料１

IRTの概要とCBT化への適用可能性

学校評価・経年変化分析調査・個別最適な学習・個の発達

東北大学大学院教育学研究科
柴山　直
教育情報アセスメント講座・教育評価測定論領域
（教育学・心理学・統計学・データサイエンス）

教育測定学とテスト理論

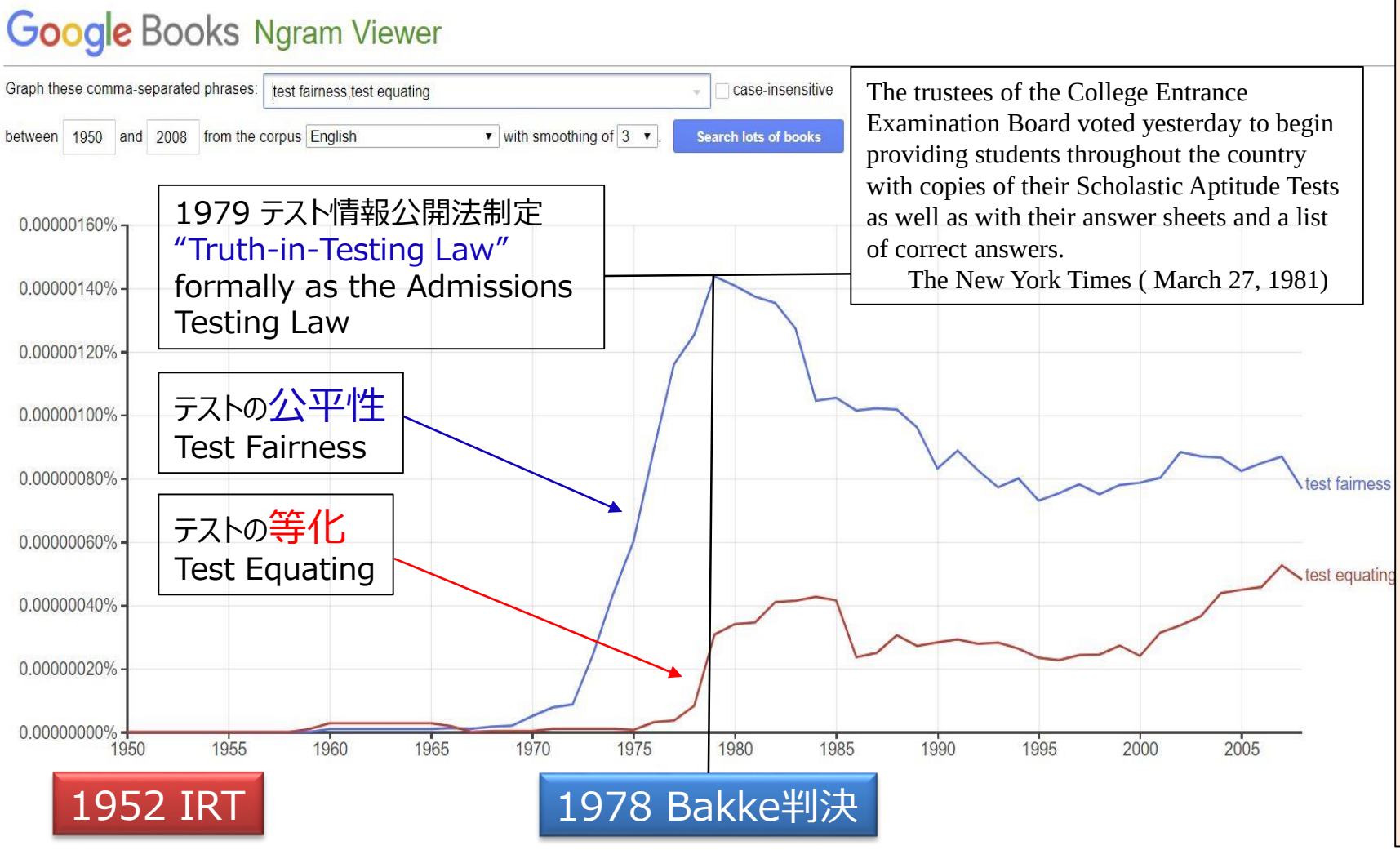
- 学力，知的能力，適性，パーソナリティ，認知および非認知的能力などの物理的には実体のない抽象的な存在の心理学的特性を，
- 統計学的モデルを通して数値化し，それらを
- 教育のエビデンスとすることを目指す研究分野

  教育測定学(Educational Measurement)
  心理測定学(Psychometrics)
  精神測定学(Psychometrics；精神医学分野における訳語)

- Evidence Based Accountability in Educationを支える基盤分野
- 役割表記：Psychometrics and analysis（ex.：PISA2018　ETS から 11 名）
- 専門職名：Psychometrician　　理論体系：テスト理論

テスト理論は測定技術でもある

例：Test Fairness (公平性)と Test Equating(等化) の出現頻度の変遷(1950-2008)



1979 テスト情報公開法制定
"Truth-in-Testing Law"
formally as the Admissions Testing Law

テストの公平性
Test Fairness

テストの等化
Test Equating

The trustees of the College Entrance Examination Board voted yesterday to begin providing students throughout the country with copies of their Scholastic Aptitude Tests as well as with their answer sheets and a list of correct answers.
The New York Times (March 27, 1981)

1952 IRT

1978 Bakke判決

◆テストの目的◆
によって技術の使い方が異なる

1）個人の処遇：個人スコア
（◆総括的評価◆のための測定）
・選抜(対訴訟：極めて厳格な測定)
・医療系大学間共用試験CBT
・全国学調_**本体調査**

2）集団の実態：集団スコア
（◆EBPM◆のための測定）
・全国学調_**経年変化分析調査**
・PISA・TIMSS・NAEP等

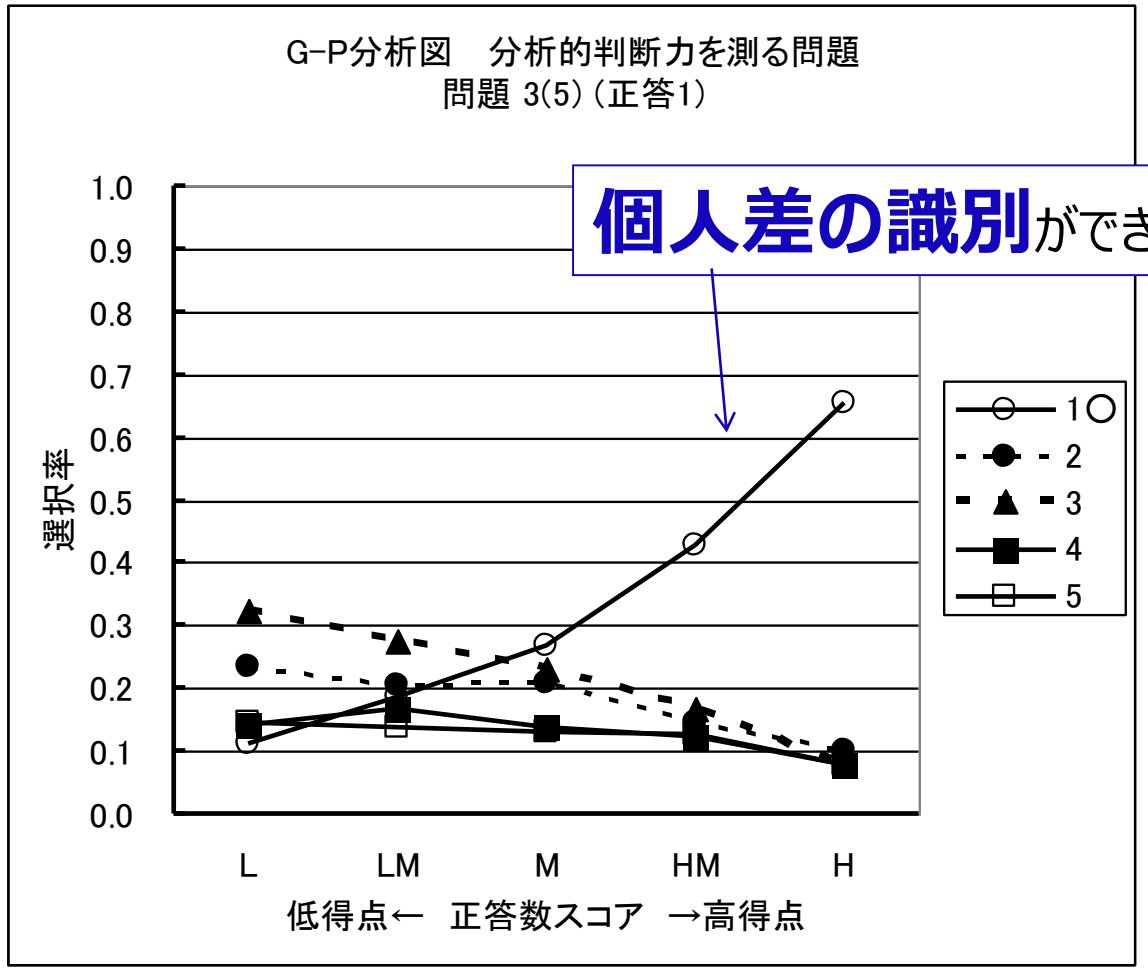
3）個人の進捗：個人スコア
（◆「学び」◆のための測定）
・GIGA個別最適な学習
・形成的アセスメント

4）個人の成長：個人スコア
（◆追跡◆のための測定）
・学力発達のサポート
・埼玉県学力調査

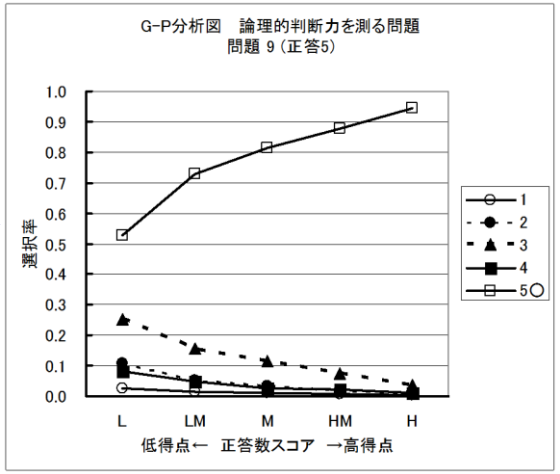
IRTモデルの
定式化の例　1/2

選択肢から見た良い項目

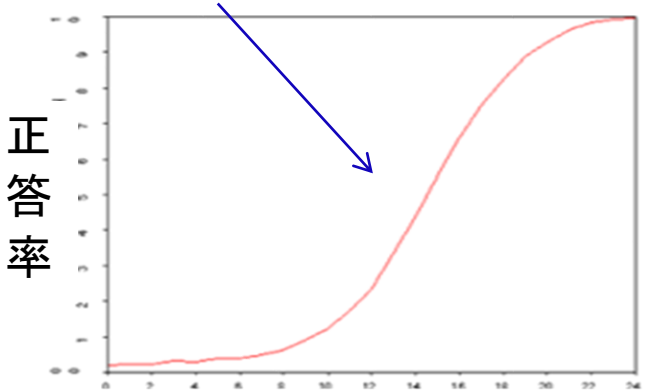
（法科大学院統一適性試験：日弁連法務研究財団より掲載許諾済 ）
https://www.jlf.or.jp/jlsat/touitsu_kakokekka/



個人差の識別ができている



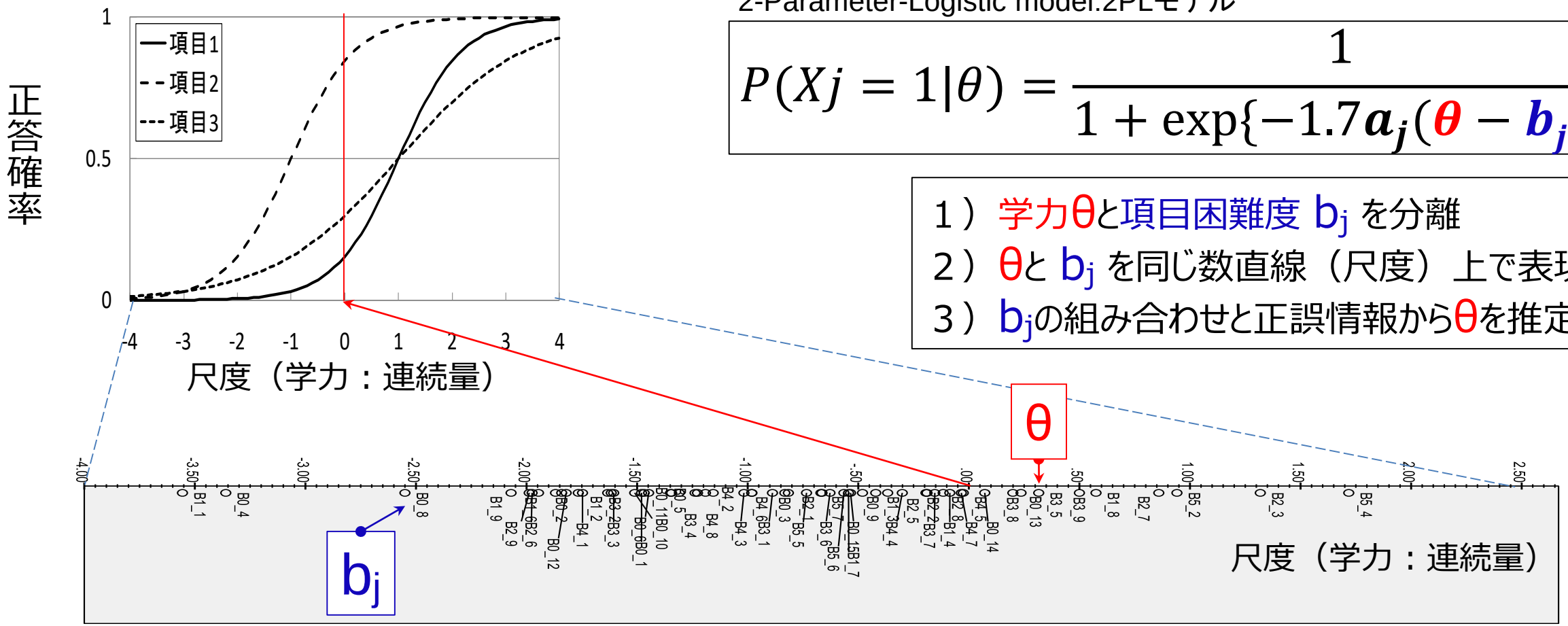
良い問題の項目特性曲線：DNC数学Ⅰ（柴山が復元）



IRTモデルの
定式化の例　2/2

IRTモデルは統計モデル

思考過程・認知過程を記述するプロセスモデルではない



2-Parameter-Logistic model:2PLモデル

$$P(Xj = 1|\theta) = \frac{1}{1 + \exp\{-1.7\boldsymbol{a_j}(\boldsymbol{\theta} - \boldsymbol{b_j})\}}$$

1）学力θと項目困難度 b_j を分離
2）θと b_j を同じ数直線（尺度）上で表現
3）b_jの組み合わせと正誤情報からθを推定

平成22年度文科省委託調査研究報告書 p.77
https://www.mext.go.jp/b_menu/shingi/chousa/shotou/085/shiryo/attach/1312362.htm

IRTモデルによるテストの得点分布の予測

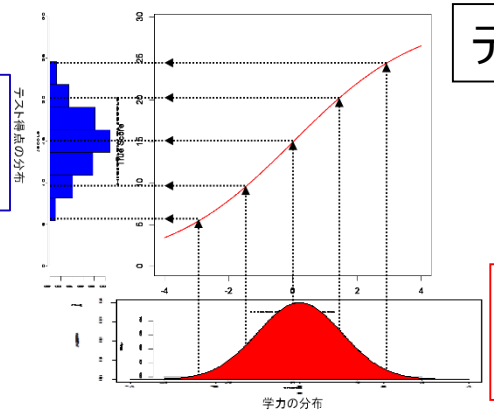
平成22年度文部科学省委託調査研究報告書p.5
https://www.mext.go.jp/b_menu/shingi/chousa/shotou/085/shiryo/attach/1312362.htm

素点分布
（得点分布）

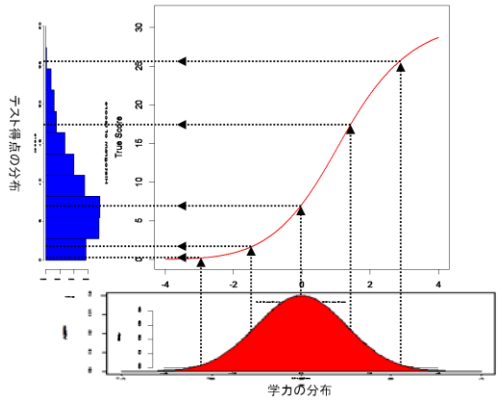
テスト特性曲線

学力分布
（θの分布）

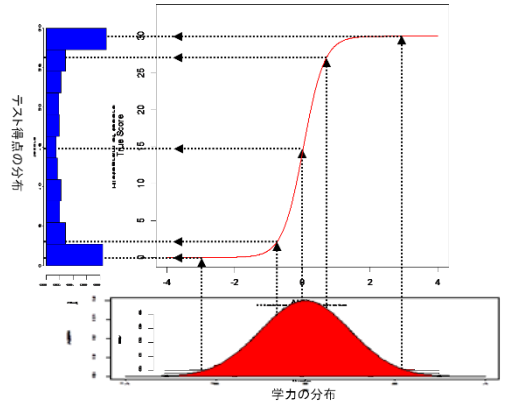
1) 一般的なテスト



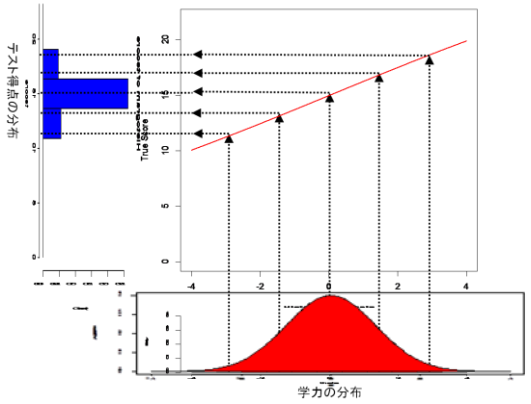
2) 高い学力層を選抜する場合



3) 資格の認定に使う場合



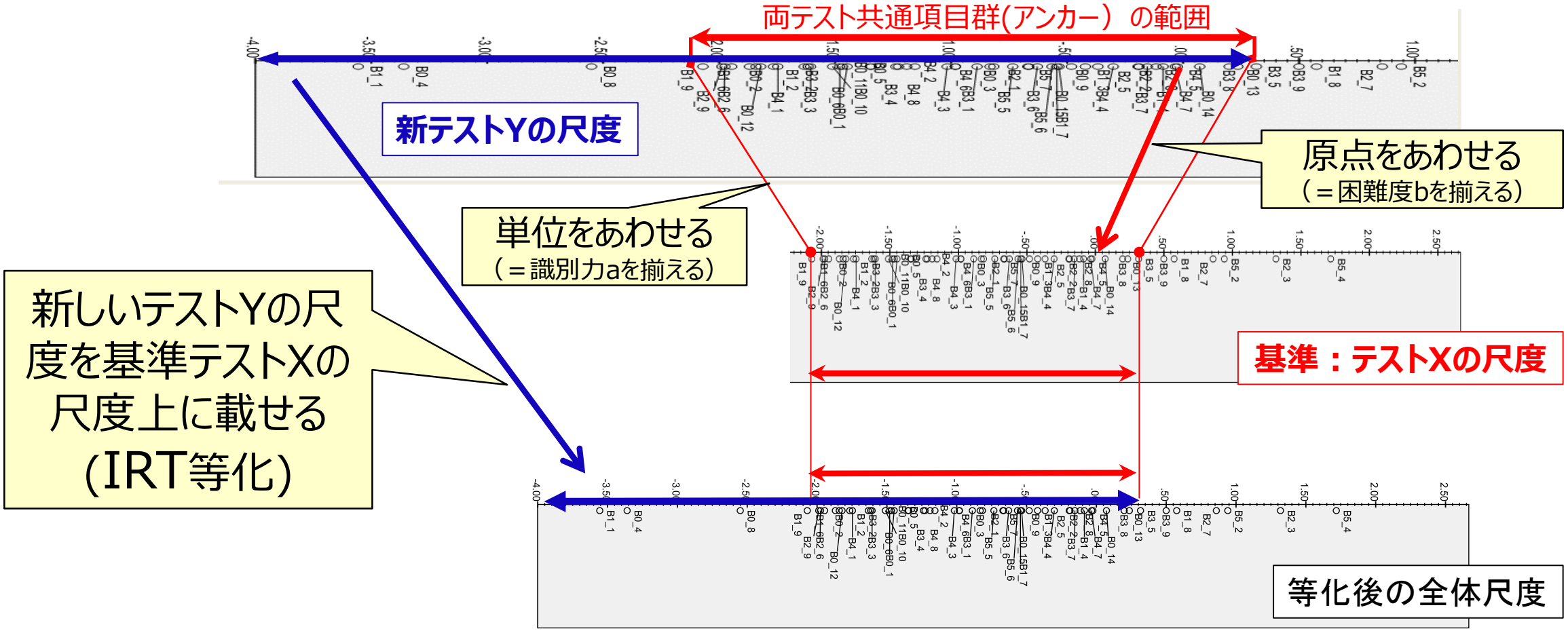
4) 個人差を小さく見せる場合



重要：
学力分布は同じでも，
テストの特性が変われば，
素点分布／得点分布の形は変化する

IRTによる尺度等化（IRT等化）の基本

線形変換により原点と単位を合わせる



新テストYの尺度を基準テストXの尺度へ等化し構成した全体尺度

IRT等化のためのデータ収集デザイン

（等化デザイン：受検者または項目/問題に共通する部分がある）

- 等価グループデザイン

グループ	テスト X	テスト Y
P_1	○	
P_2		○

- 単一グループデザイン

グループ	テスト X	テスト Y
P	○	○

- カウンターバランス デザイン

グループ	テスト X	テスト Y
P_1	1	2
P_2	2	1

- アンカーテストを伴う不等価グループデザイン（NEAT）

グループ	テスト X	テスト A	テスト Y
P	○	●	
Q		●	○

※P1，P2は同じ母集団Pからの異なる標本を、また、P,Qはそれぞれ互いに異なる母集団からの標本（児童・生徒）を表す。
※カウンターバランスデザインの表中の数字は実施順を示す。

IRTの
柔軟性の
具体例1/2

学校評価のIRTスケールに基づくCS分析法の開発

※CS分析=（Customer Satisfaction分析；顧客満足度）分析

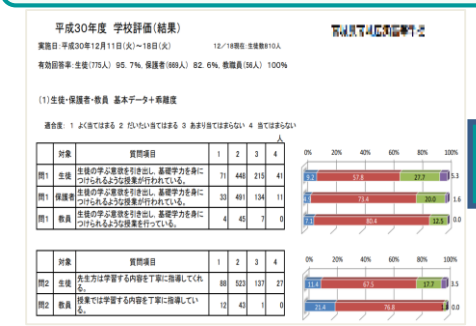
①

①顧客集団内の満足度を可視化する
CS分析 → どの項目から着手すれば効率的に
顧客の満足度を向上できるかが分かる

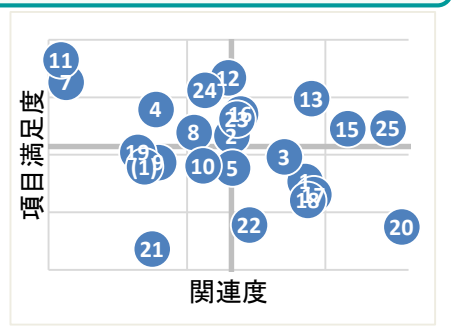
②GRM(段階反応モデル)に基づいた
CS分析を実証
※EasyEstimation http://irtanalysis.main.jp/

② ⇦

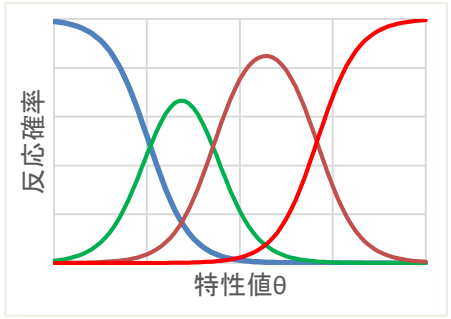
IRT導入
による
柔軟性
汎用性
の実現



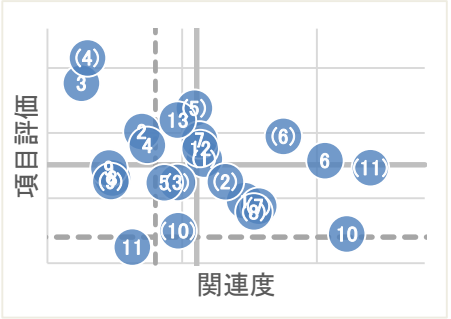
項目ごと単純集計のみ



CS分析グラフ



GRMのカテゴリ確率曲線



IRT−CS分析グラフ

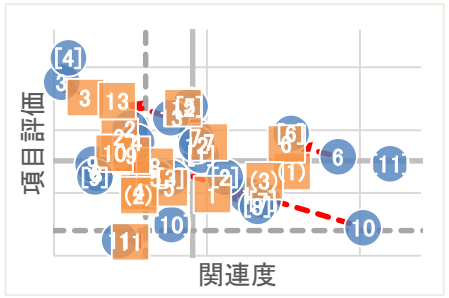
③

③IRT尺度に基づくCS分析によって，
改善目的での学校内年度間比較等が可能

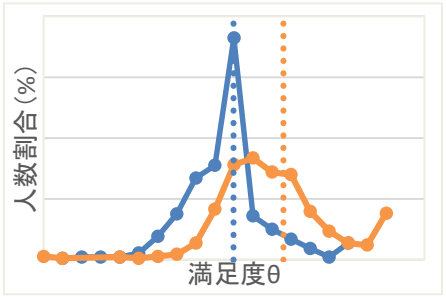
④様々な角度からの分析を自動的にできる
Excel ツールを開発

等化デザイン
の組み込み

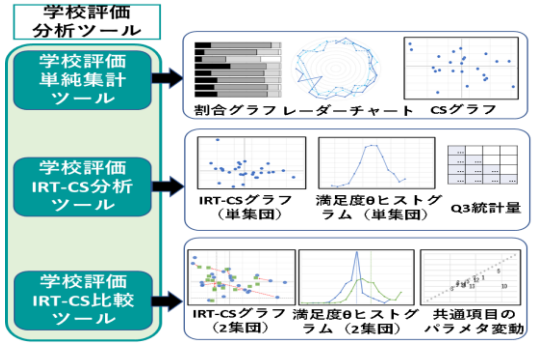
グループ	テスト X	テスト A	テスト Y
P	○	●	
Q		●	○



IRT-CS比較グラフ



満足度θのヒストグラム



④

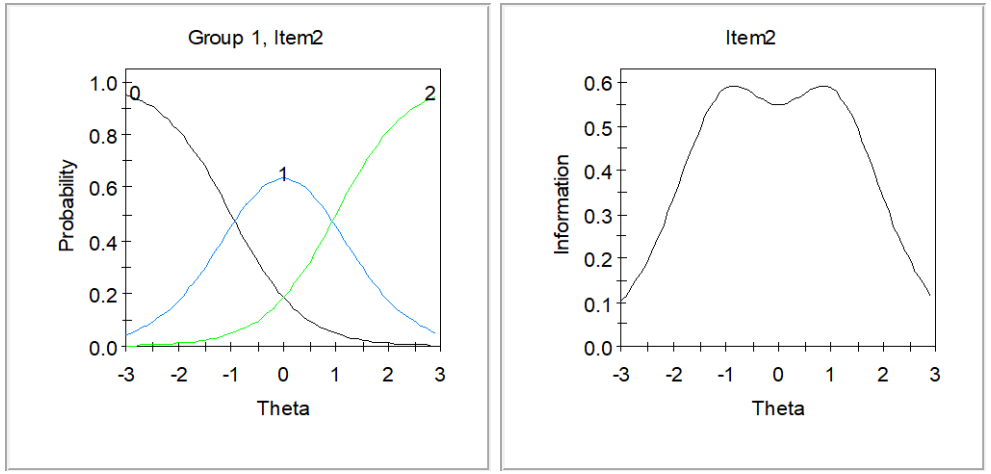
成果：
学校評価の
客観化・自動化
（DX）を実現

三品（2021）東北大学大学院教育学研究科修士学位請求論文（教育情報学）/東北大学機関リポジトリにて公開予定 https://tohoku.repo.nii.ac.jp/

IRTの
柔軟性の
具体例2/2

記述式問題へIRTモデルの適用は可能

- **記述式問題：** この文章では最後の一文だけ「～です」が使われていますが，それによってどのような表現効果があるか書きなさい。
- 採点基準 → **手採点** → 0，1，2
- **段階反応モデル**（Samejima's **GRM**)の適用



カテゴリー確率曲線（左） と 項目情報量（右）

- 複数のIRTモデルの同時適用も可能
- **自然言語処理（NLP）**にもとづく自動採点による結果（数値化）に，**高い信頼性**が担保できれば，GRM/IRTのCBT適用は可能
- 採点作業の格段の**効率化**
  - 現時点では大規模アセスメントにおける手採点の代替になるかは慎重な判断が必要

平成23年度文部科学省委託研究：2.2 多値項目反応モデルの導入
http://www.mext.go.jp/b_menu/shingi/chousa/shotou/085/shiryo/attach/1323273.htm

平成24年度文部科学省委託研究調査報告書Pp.10-19
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/1344827.htm

経年変化分析調査のための分冊デザイン

（Item-Matrix Sampling： 標本調査法 ＋ 実験計画法 ：重複テスト分冊法）

グループ	テスト X	テスト Y
P_1	○	
P_2		○

グループ	テスト X	テスト Y
P_1	1	2
P_2	2	1

グループ	テスト X	テスト Y
P	○	○

グループ	テスト X	テスト A	テスト Y
P	○	●	
Q		●	○

	分冊1	分冊2	分冊3	分冊4	分冊5	分冊6	分冊7
位置1	3	4	5	6	7	1	2
位置2	5	6	7	1	2	3	4
位置3	6	7	1	2	3	4	5
位置4	7	1	2	3	4	5	6

・どの項目セットも等しく４回使用されている → **使用頻度**
・どの項目セットも互いに等しく２回ずつ会合する → **組合せ**
・どの項目セットも等しく別の位置に配置される → **出現順**

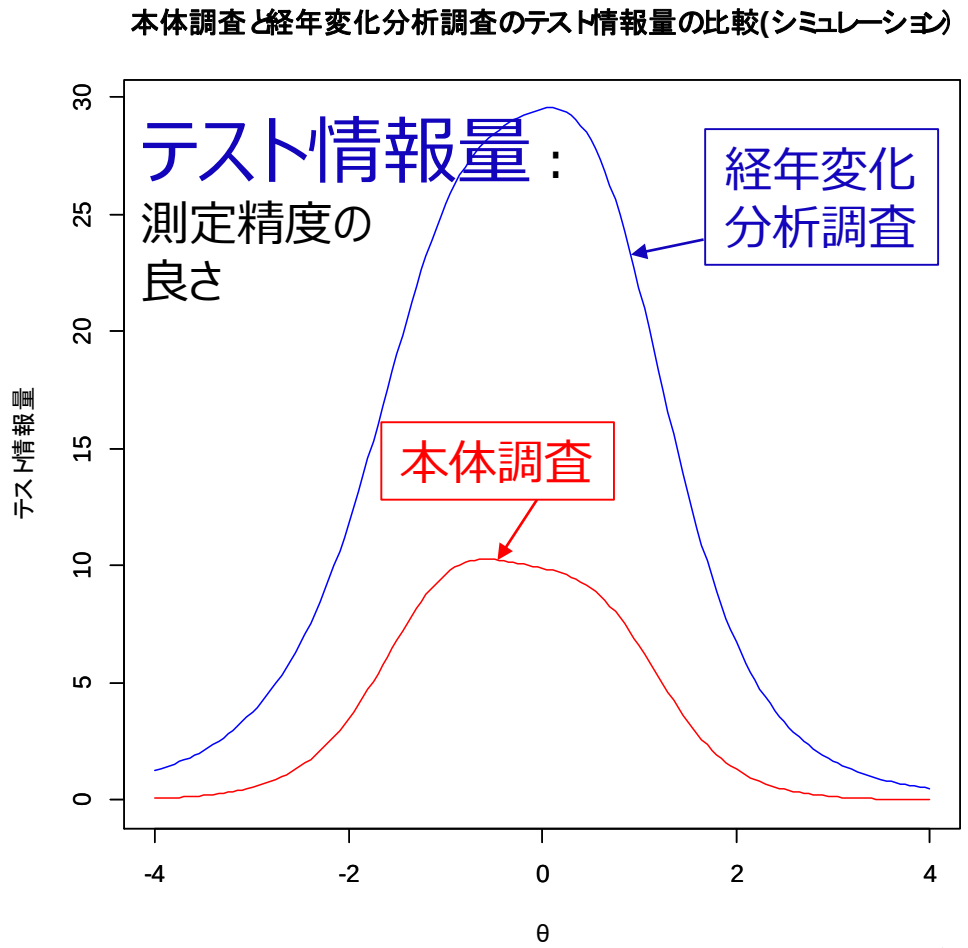
の効果を相殺

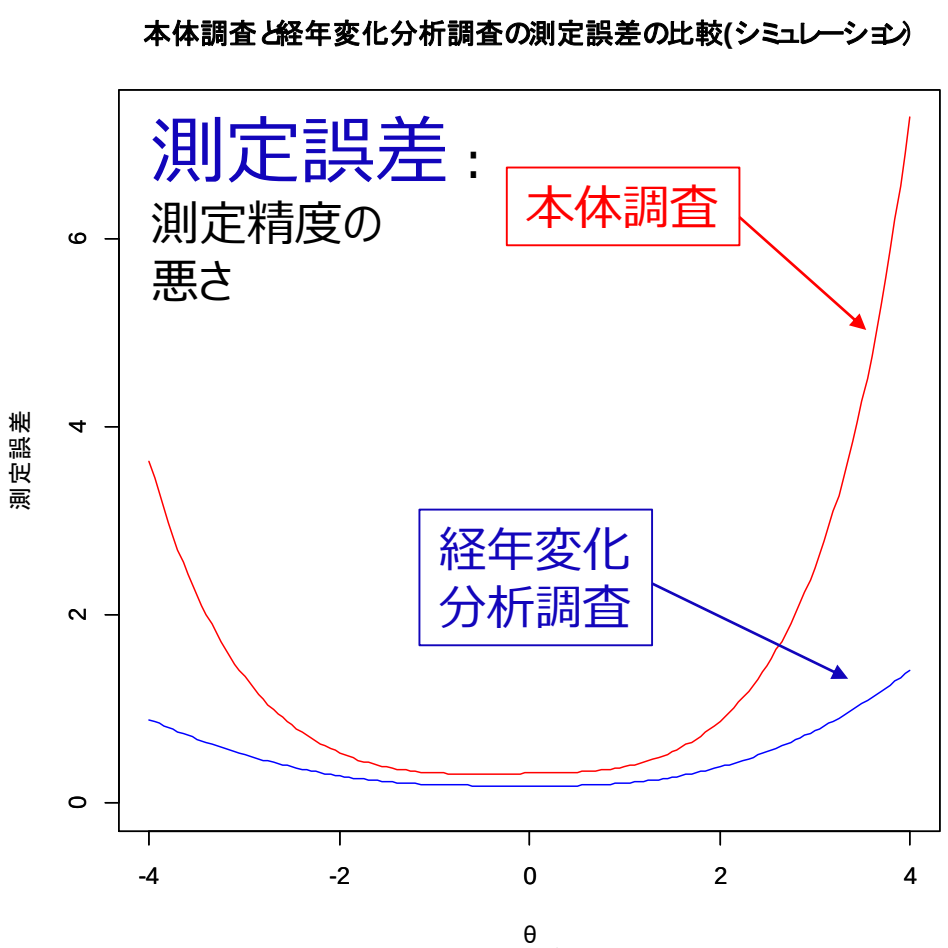
※表内の数字は項目セット（＝PISAではcluster）
※実験計画はBIBD＝Balanced Incomplete Block Design（釣合い型不完備ブロックデザイン：Yuden方格）

注意：

H28年度 経年変化分析調査では13分冊
実施報告書（文部科学省・国立教育政策研究所）

本体調査と経年変化分析調査の測定精度





本体調査の問題は
予備調査なし
毎年**公開**
- 15問で精度を確保
- 作問能力の高さ

経年変化分析調査の
問題は**非公開**
- 曝露効果の防止
- 予備調査必須
- コストの削減
- 挙動異常の問題は
  差し替え可能
- サンプル問題公開

注意：あくまでも相対的な比較をしているだけで，本体調査の測定精度が「悪い」といっているわけではない

◆EBPM◆のための測定　3 /3
Test Linking(仕様の異なるテスト間の「**対応づけ**」:前提条件が**等化**より緩い)

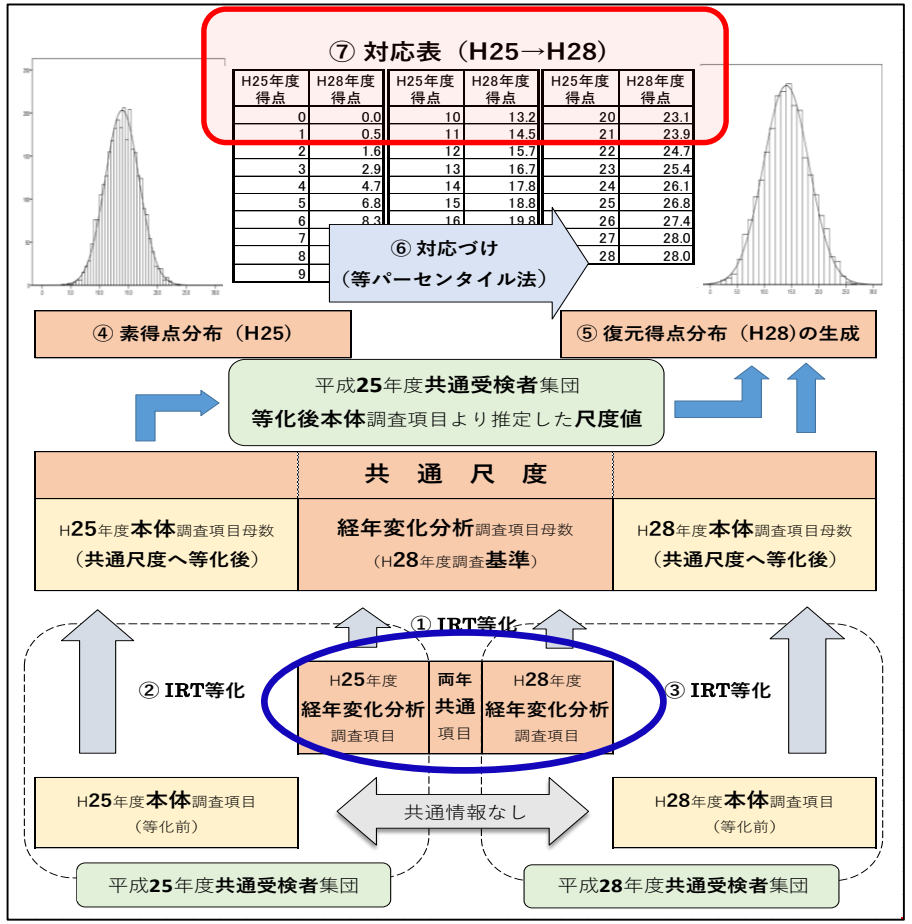
経年変化分析調査との対応づけによる本体調査の年度間比較の試み
（H25分布をH28分布に変換する）

派生成果：**領域ごとの年度間比較**のために**推算値**（PVs）を利用する方法の提案（Pp.38-46）

重要：経年変化分析調査とPISA（筆記型）の仕様の違い

- **構成概念（学力の定義）が異なる**
- **分冊内は同一教科の問題のみ**
  - -子ども達の混乱を避けるため
  - -問題数を確保し測定精度を確保
- **下位分布などの構造を組み込んでいない**
  - 潜在回帰/能力回帰モデルのθへの組み込みなし
  - subpopulationsの属性定義が不安定で困難
  - 等化作業の煩雑化/等化誤差の混入を回避
  - 保護者調査等で収集する属性情報を使う
- **測定モデルはIRT-2PLモデル**（PISAも2015からRasch→2PL）

注意：経年⇔PISA
等化　→　×
対応づけ　→　×
相関関係　→　○



平成29年度文科省委託調査研究報告書
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/1406895.htm

地方独自調査と本体調査との対応づけ
のための調査デザインも可能

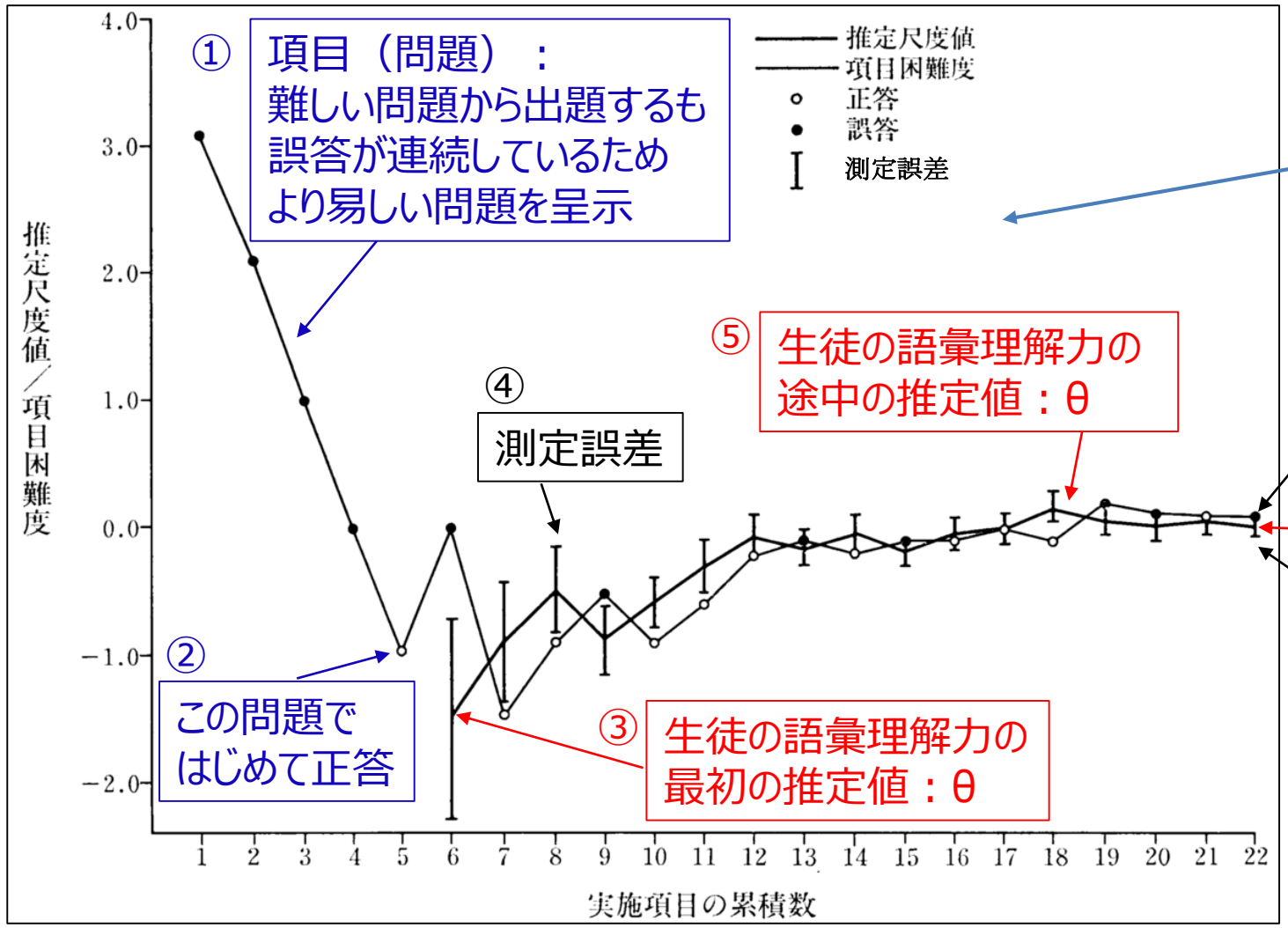
先行例：法科大学院適性試験（JLSAT）

日弁連法務研究財団 ←（※対応表）→ DNC

※但し，選抜利用のためほとんど等化に近い厳格な条件での対応づけ

◆「学び」◆のための測定　←　項目管理の技術

コンピュータ適応型テスト(CAT)の原理



① 項目（問題）：
難しい問題から出題するも
誤答が連続しているため
より易しい問題を呈示

② この問題で
はじめて正答

③ 生徒の語彙理解力の
最初の推定値：θ

④ 測定誤差

⑤ 生徒の語彙理解力の
途中の推定値：θ

項目プール（アイテムバンク）

・項目母数既知の**241**項目
・難易度：**小1～大学生**
・同一尺度上にIRT等化済み

最後に呈示された項目の困難度

⑥ 生徒の語彙理解力の最終的な推定値：θ

生徒のθに関する最終的な測定誤差

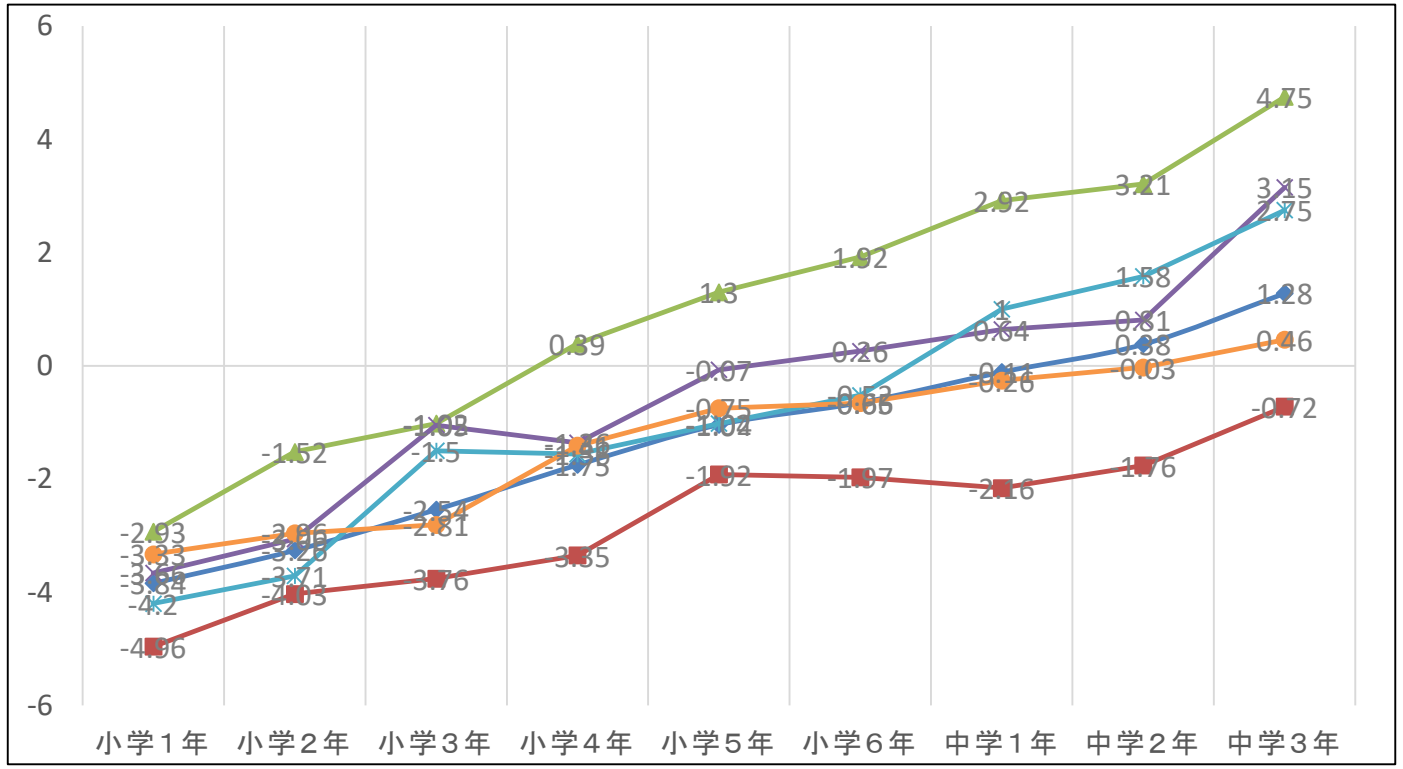
・ CATでは大きなアイテムバンクは必須
・ それ以外のアイテムバンクの大きさは
目的・予算・セキュリティ・更新回数等
に応じてケースバイケース

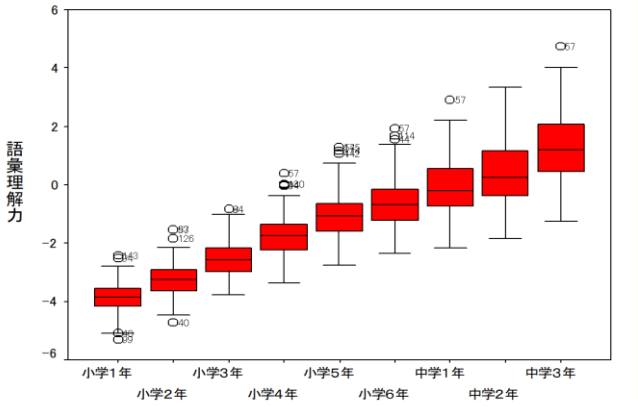
柴山・野口・芝・鎌原（1987）「最適化テスト方式による語彙理解力の測定」 https://doi.org/10.5926/jjep1953.35.4_363

◆追跡◆のための測定：Vertical Scaling（垂直尺度化）

児童・生徒の語彙理解力発達の追跡研究

（東京杉並区の小中学生の9年間）





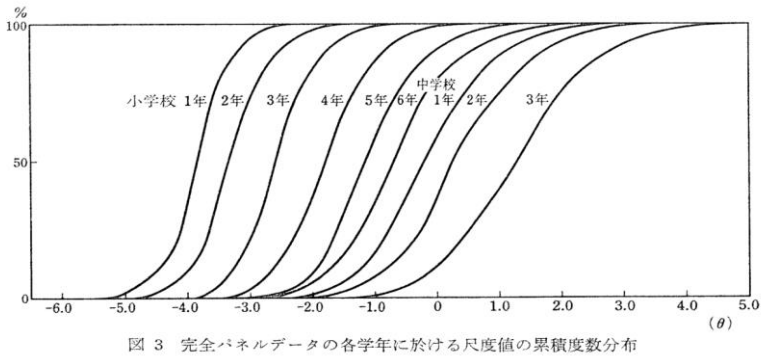


図 3　完全パネルデータの各学年に於ける尺度値の累積度数分布

芝・野口・柴山（1986）「語彙理解力の発達に関する追跡的研究」 http://doi.org/10.15083/00029860

参考1：柴山（2016－2020）「発達段階をトレースできる到達度評価のためのＩＲＴ垂直尺度構成の試み」
https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-16H03731/

参考2：澁谷（2019）「異なる難易度のテスト項目のIRT**垂直尺度化**」 http://hdl.handle.net/10097/00125343

最後に

目的ごとにIRTの使い方は異なる

1）個人の**処遇**：個人スコア

◆総括的評価◆のための測定

・全国学調_**本体調査**

※あらたな視点・発想で作り直す必要あり

2）**集団の実態**：集団スコア

◆EBPM◆のための測定

・全国学調_**経年変化分析調査**

3）個人の**進捗**：個人スコア

◆「学び」◆のための測定

・GIGAスクールにおける個別最適な学習

4）個人の**成長**：個人スコア

◆追跡◆のための測定

・埼玉県学力調査

経年変化分析調査CBT化の課題

A) PBT（筆記型）とCBT測定の**同等性・継続性の検証**

B) 良い問題の開発には**測定すべき「学力」の明確化**が重要

A）伝統的な「教科学力」 ⇔ 「情報活用力」

B）「操作リテラシー」 ⇔ 「情報活用力」

C) 心理学的構成概念**（ことばの定義）の整理**が必要

A）「操作リテラシー」

- IT機器の文具・鉛筆・紙冊子並の**利便性**が前提

B）「情報活用力」（マジックワード的に使用されている）

- 「教科学力」を **IT で augment する力**等

大規模アセスメントのCBT化

IRTの柔軟さ × simpleで確実なIT

TOHOKU UNIVERSITY