テスト理論入門

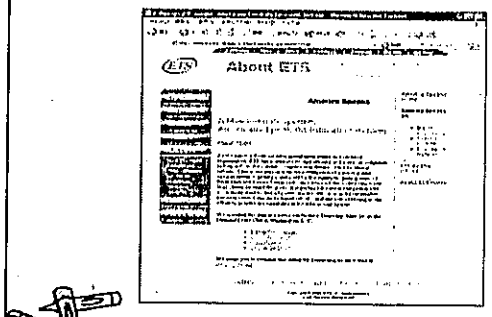
植野真臣

長岡技術科学大学

1. ETS

Educational Testing Service (ETS)

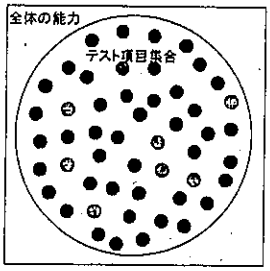
- 1947年設立
- SATテスト、GREテスト、TOEFL, TOEIC 等



2. テストの考え方の基本

測定対象となる膨大なテスト項目への被験者の反応を推定するために、限られた数のテスト項目によって、全体の項目への反応を推定する。

測定される能力はテスト項目集合によって決定される。



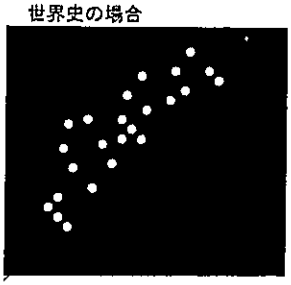
出題されるテスト項目

全体よりランダムに抽出されなければならない。

3. 項目数が信頼性を決める

世界史の場合

浪人時代の偏差値

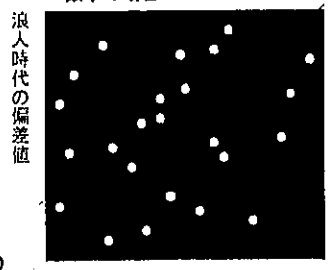


現役時代の偏差値

サンデー毎日1977. 3. 27

数学では無相関

数学の場合



4. テスト構成の課題

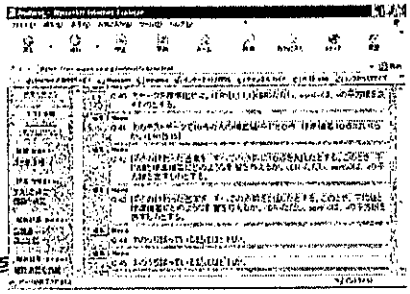
- 測定したい能力、特性分野から如何にバランスよく項目を選択するか。
- 信頼性を保持し、如何に項目数を減らせるか。

5. アイテムバンクの設計

- 統計データを伴う項目のデータベース
- 項目ID

評価観　1.「情報活用の実践力」、2.「情報の科学的な理解」、3.「情報社会に参画する態度」
目標　1.「知識・理解」、2.「思考・判断」、3.「技能・表現」、4.「興味・関心・態度」
教科番号A
分野番号1：情報を活用するための工夫と情報機器
単元番号1：問題解決の工夫
目標番号1：問題解決を効果的に行うためには，目的に応じた解決手順の工夫とコンピュータや情報通信ネットワークなどの適切な活用が必要であることを理解する。

アイテムバンクの例



6. IRT(項目反応理論)

- 過去のテストデータからテスト項目の特性値を数学的に推定。

Advantages

- 項目の良し悪しの評価とフィードバック
- 異なるテスト項目を受けた被験者を同じ尺度上で評価できる(等化)
- 自動テスト構成
- 適応型テスト(CAT)

モデル

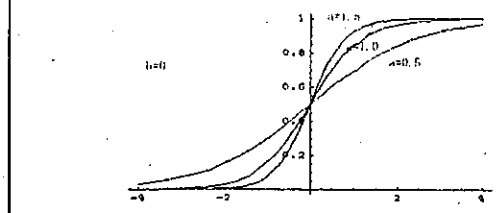
$$P(x_i=1|\theta)=\frac{1}{1+\exp(-1.7a_i(\theta-b_j))}$$

パラメータa：項目の識別力

パラメータb：項目の難易度

Θ:受験者の能力

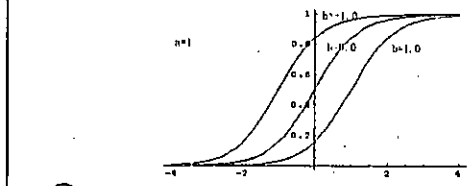
パラメータaを変化させると？



パラメータaの解釈

- 識別力パラメータ
- 正答確率が0．5となるところの曲線の傾きを示している。
- 被験者の能力θをどれくらい識別できるかを示している

パラメータbを変化させると？



パラメータbの解釈

- 難易度パラメータ
- 正答確率が0．5となるところの能力を示している
- パラメータbが高い項目ほど正答するのに必要な能力が高くなる

項目応答理論の例

問題

1. 方程式 x + 8 = -2を解きなさい。
2. 方程式 −4 x = -6を解きなさい。
3. 方程式4x+7= -5を解きなさい。
4. 方程式5x-6=-4x+3を解きなさい。

被験者

中学1年生 428名

パラメータ推定結果

	正答率	a	b
1	0.7287	0.7975	-0.09176
2	0.6726	1.3471	-0.52678
3	0.6510	0.8739	-0.56454
4	0.6438	0.95784	-0.50647

項目分析

- 識別力パラメータについては、2>4>3>1である。項目1のように低い項目は、もう一度、内容を検討してみる必要がある。
- 難易度パラメータについては、1>4>2>3である。正答率の順序と異なる。

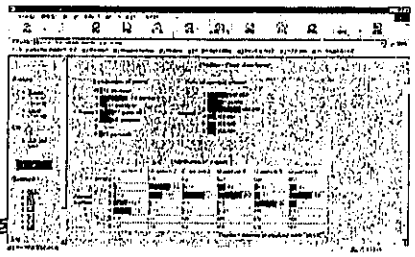
テスト構成のための項目情報量

$$I_i = \frac{[\frac{\partial}{\partial\theta} P'(x_i = 1 \mid \theta)]^2}{P(x_i = 1 \mid \theta)[1 - P(x_i = 1 \mid \theta)]}$$

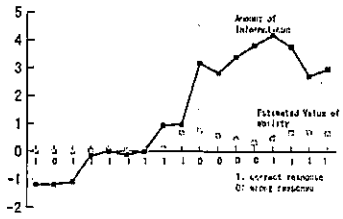
テスト情報量による テスト構成

- 対話的テスト構成支援
- 適応型テスト

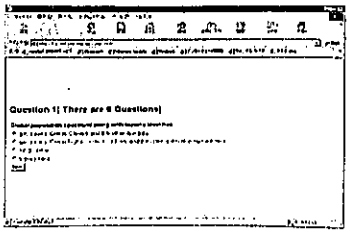
テスト構成支援システム



適応型テストの動作例



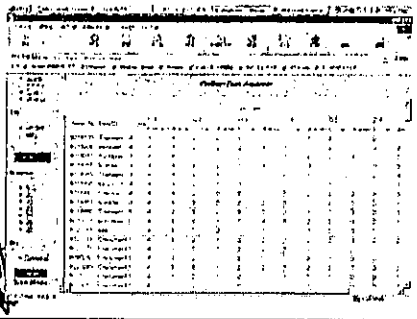
Computer based Testing



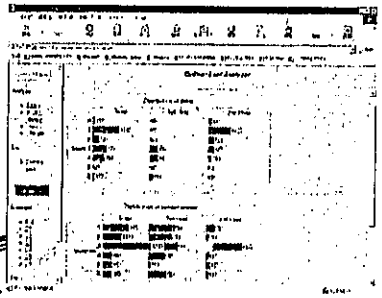
Advantages

- ペーパーテストでは測定できない能力を測れる
- ペーパーテストでは測定できないデータが測定できる
- 自動採点と即時フィードバック

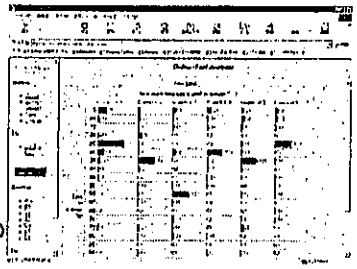
テストデータ行列



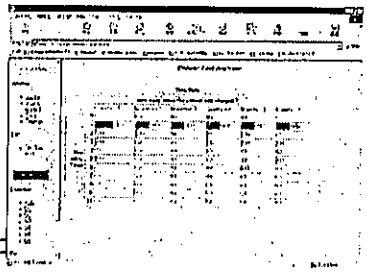
テスト得点分布



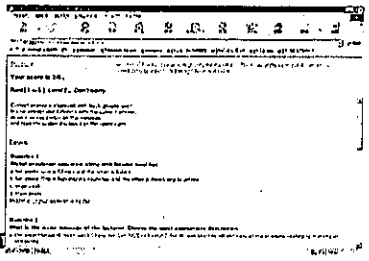
所要時間分布



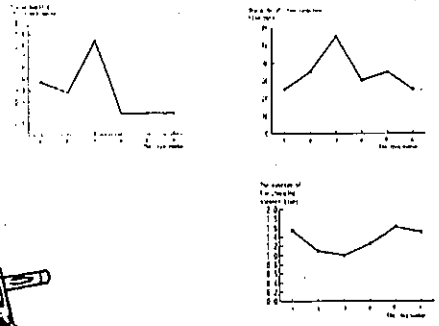
回答書き直し回数分布



受験者への即時フィードバック



多様な項目分析



入試問題の意味

- 学習者の学力を正当に効率よく評価するためのツール
大量の問題出題と同程度の精度を持つ少数項目テストの開発

⇔

- 新たなる学力観のインセンティビティ
- ↓
- 作成者間での入試問題の意味の共有度の少なさ

ある大学入試問題における選択問題の例

問題1

問題2

問題3　　問題4

分析の問題

- 選択問題3と4を選択した人では、どのような違いがあったのか？
- 選択問題3と4を選択した人を同一尺度上で評価したい

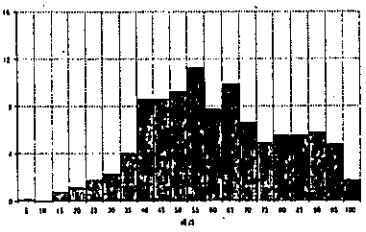
↓

項目応答理論による分析

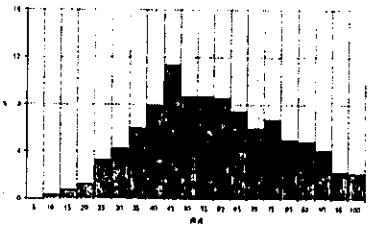
選択問題3と4

- ほぼ例年、平均値は有意差なし
- 作問者の努力は、この点数差をなくすために行われる

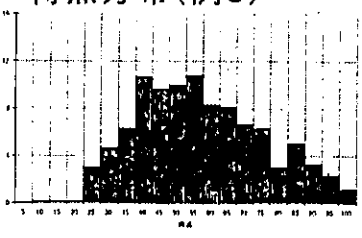
得点分布(例1)



得点分布(例2)



得点分布(例3)

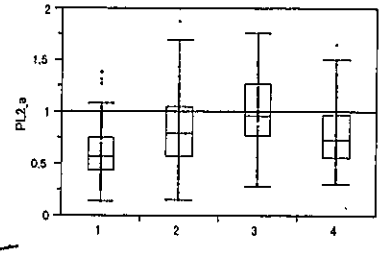


IRTによる分析

- 選択問題3と4の扱い
- 選択問題3を受けた受験生の答案では、問題4への反応を欠測値として、選択問題4を受けた受験生の答案では、問題3への反応を欠測値として扱った。

→ IRTでしかできない分析

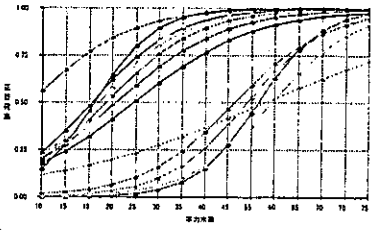
問題別識別力パラメータ分布



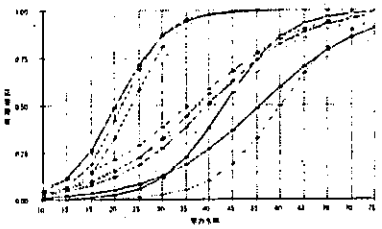
何がわかるか？

- これまで圧倒的に問題3が識別力が高く、問題1は非常に低く、問題2，4もかなり低いことが分かる。このことから、テスト全体の得点は、問題3の結果に非常に左右される傾向があると言える。

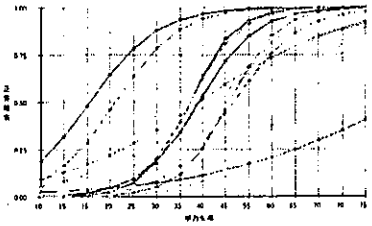
問題1の例



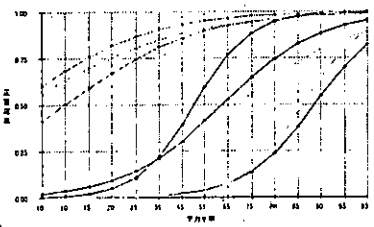
問題2の例



問題3の例



問題4の例



問題3と問題4の差は？

- しっかり勉強してきたのであれば、問題3を選択すれば、自己の努力は報われる。
- 勉強をあまりしてきておらず、実力が十分でない場合は、問題4を選択すれば、よい点が取れる確率が高い。

何が差異なのか？

- 問題は二つの選択問題の平均点の差ではない（なぜなら項目応答理論を用いれば等価という技術を用い、科学的に異なるテストのテスト結果を一次元上で評価できる）
- 選択問題における識別力の差は、問題である。問題4の問題作成方略を検討すべきである。

まとめ
テスト作りで間違いやすい点

- テストは、平均点が50点になるような正規分布になるようなテストがよいとは限らない。
- 平均点の調整だけが問題ではなく，識別力の高い問題をどのようにして作問するかということをもっと考えるべきである。