eテスティングの基礎

植野真臣
電気通信大学

１．基礎用語と概念の準備

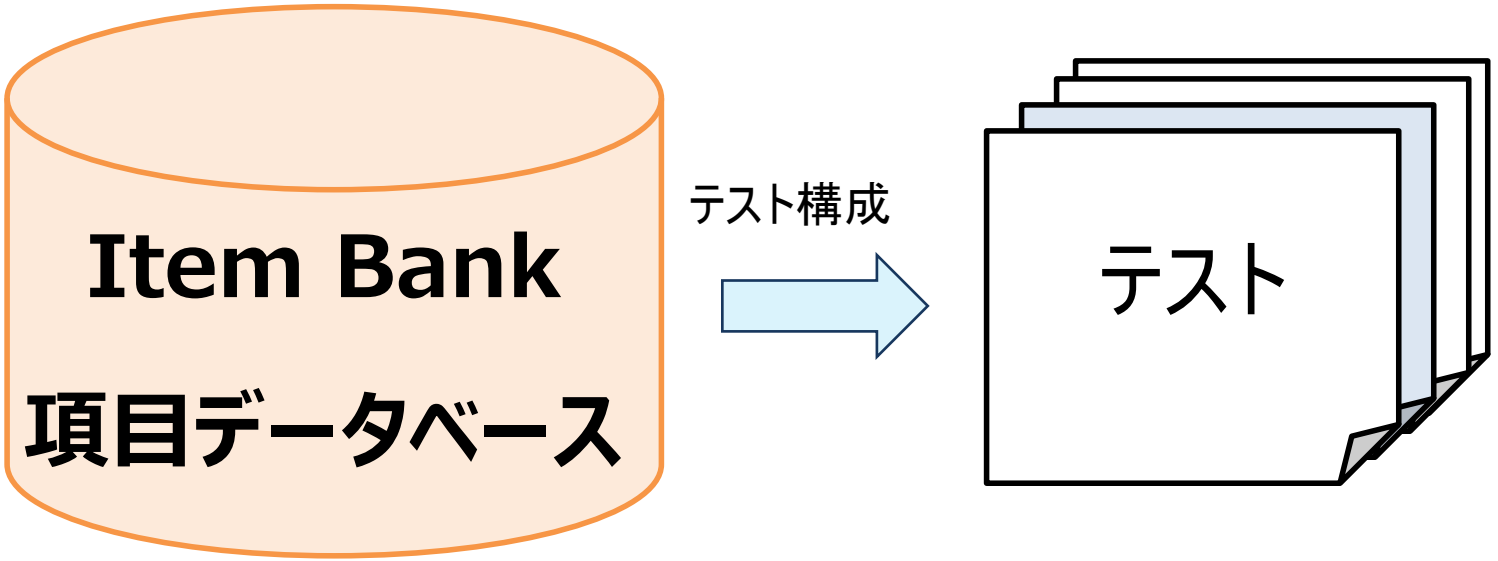
Computer based Testing (CBT)

- コンピュータによって行われるテストの総称

アイテムバンク

テスト問題項目がその属性データとともに格納されているデータベース。

テスト運営にとってアイテムバンクを構築・管理することが最もコストが高い。



テストの信頼性

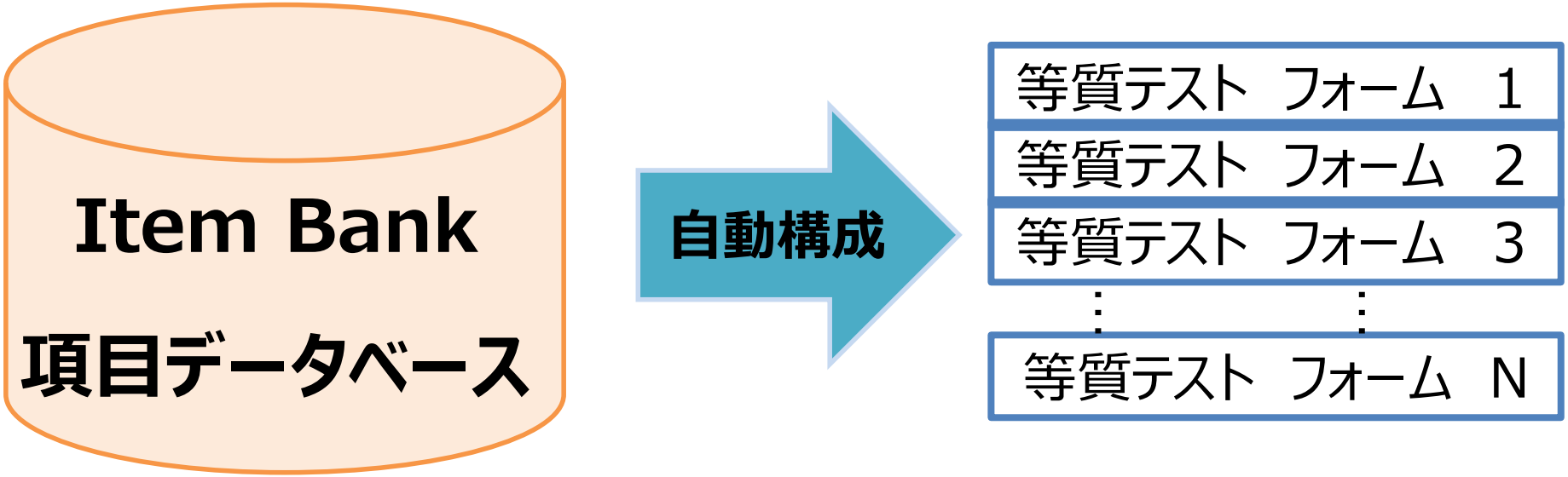
同一の能力の被験者が異なる項目より成るテストを複数回受検しても、同一の得点が得られる。

eテスティング

- 同一能力の受検者が異なるテストを何度受検しても同一の得点スコアを返すCBT
- 信頼性の高いテストを実現する技術
- 心理学×統計×コンピュータサイエンスの複合分野

eテスティング：等質テスト生成

できるだけ多くの等質テストを生成



我が国のテストでは２００００問程度の問題項目が蓄積されていることが多い.
毎年項目は追加されるが，作問した問題項目がすべて登録されるわけでなく非常に貴重！！

２．信頼性の高い公平なテストを実施するために

異なる項目より構成されたテストの評価の問題

A組のＡ君とB組のＢ君は　異なるテスト問題によって構成されるテストを受験した。

A君の得点は75点　Ｂ君の得点は80点

問題

このとき、Ｂ君のほうがＡ君より　学力が高いといえるか？

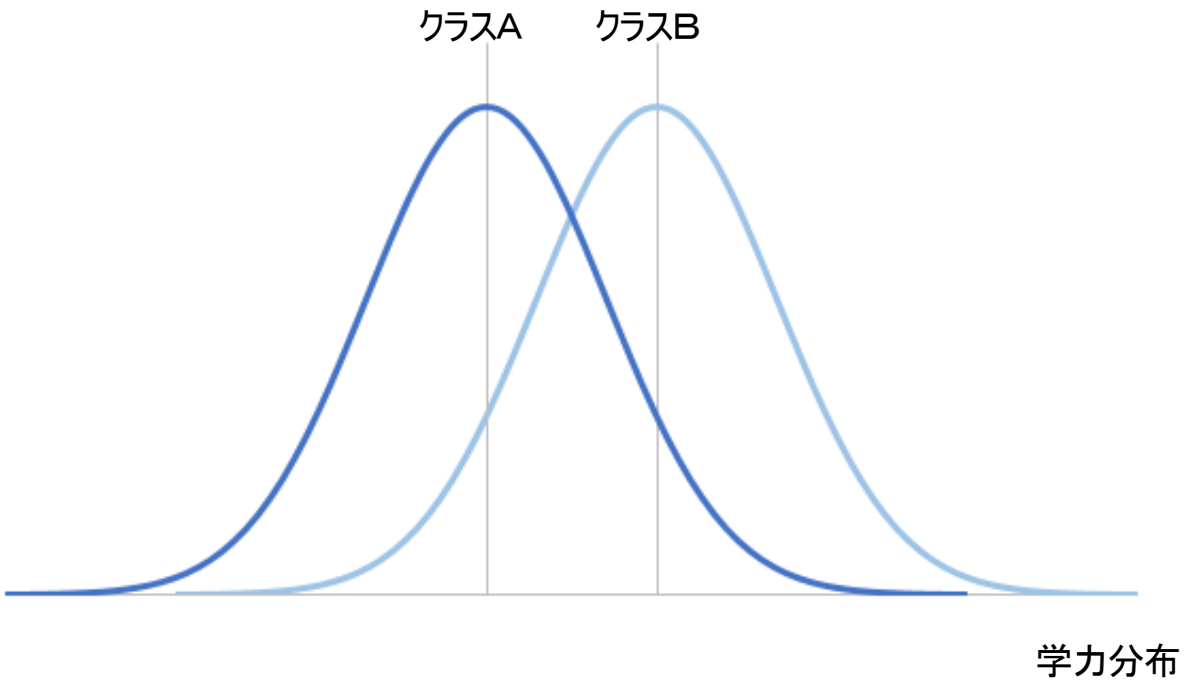
正解

テストの難易度が違うかもしれないので何も言えません！！

じゃあ、偏差値を用いればよい

- 偏差値＝$\dfrac{\text{得点}-\text{平均値}}{\text{標準偏差}}$

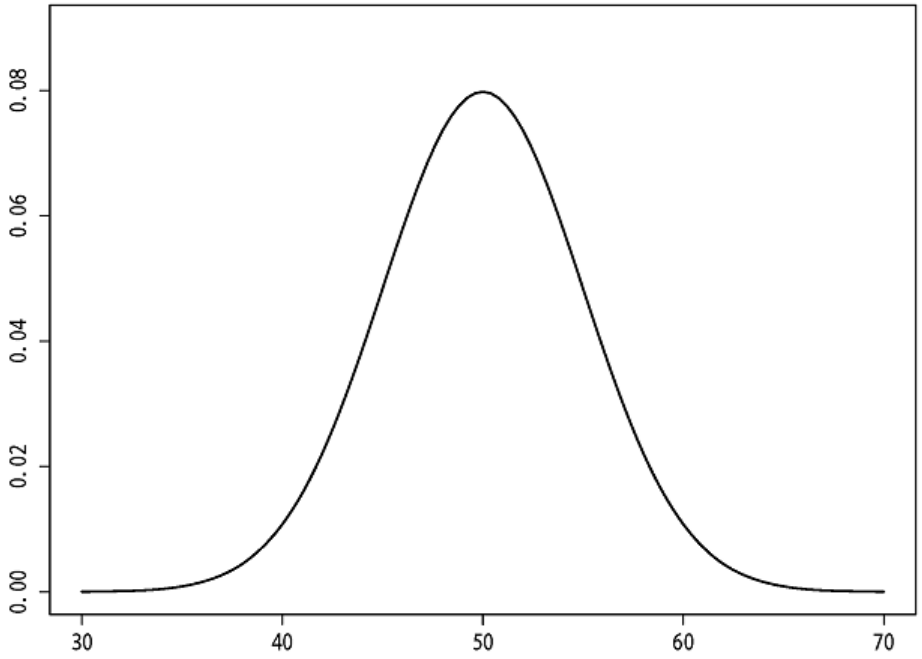
ＡとＢの所属するクラスの学力分布に依存してしまう：ＢがＡより学力の高いクラスなら何も言えない



受検者母集団の得点分布を一定にするテストを作成すればよい！！

- 違う項目で作成されるテストが同一の分布を持つように作成しなければならない。

平均点を５０点と決めよう！！



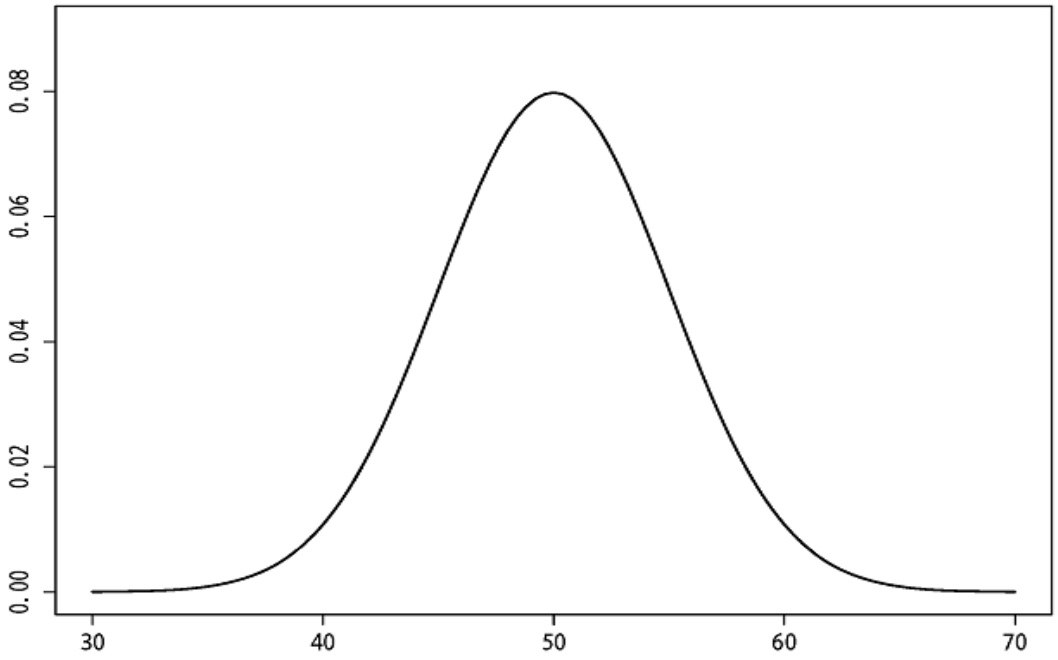
テストの作り方

- 平均点50点のテストを作る手法を考えてみましょう！！

問題

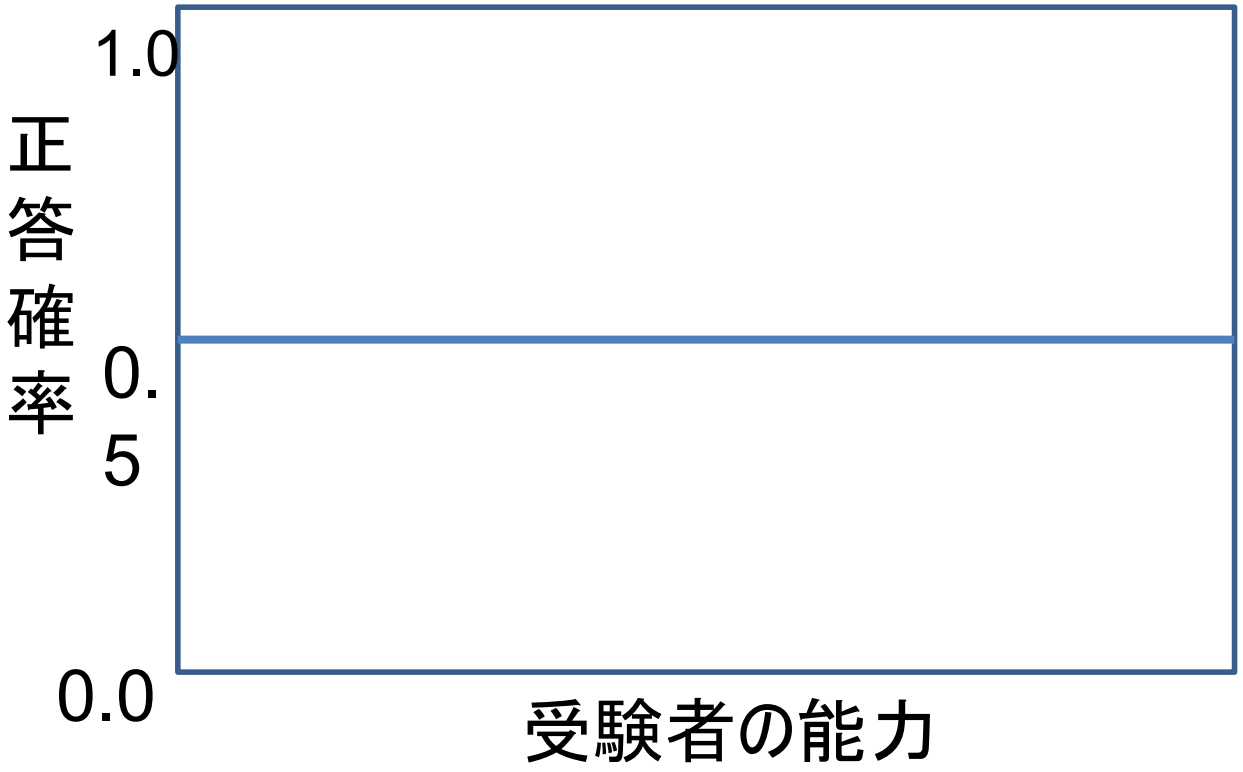
- 「今からコインを投げる。　表がでるか裏が出るか？」を１００回　問うテスト

平均点は５０点の分布になる

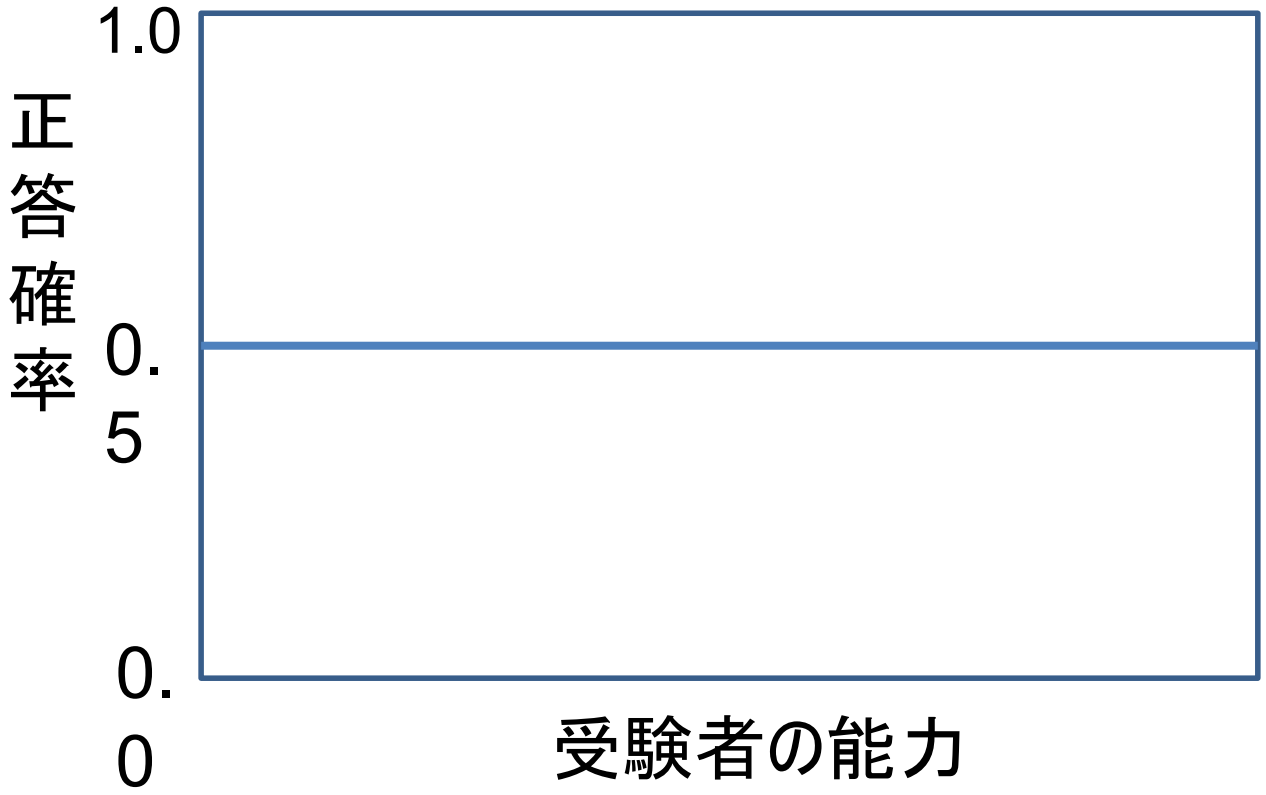


こんな問題おかしい！！

何がおかしい？



受験者の能力を何も測ってません！！



結論

- 得点分布を等質にできても　受検者の得点は偶然の産物で、テストをやるごとに点数は変わってしまう！！
- 信頼性の高いテストを保証しない！！

解決法

得点分布を等質にするのは難しい。

項目反応理論は異なる問題項目で構成されたテストの得点を同一尺度で評価できるらしい！！　使ってみよう！！

しかし

- 項目反応理論を機械的に用いたからといって信頼性のあるテストはできないのです！！
- テストには誤差があり、この誤差を平等に可能な限り小さくなるように構成しないといけないからです。

3．項目反応理論

- 項目反応理論
- Item Response Theory (IRT)

項目反応理論
２パラメータロジスティックモデル

$$P(x_j = 1|\theta_i) = \frac{1}{1 + \exp(-1.7a_j\theta_i + b_j)}$$

項目反応理論

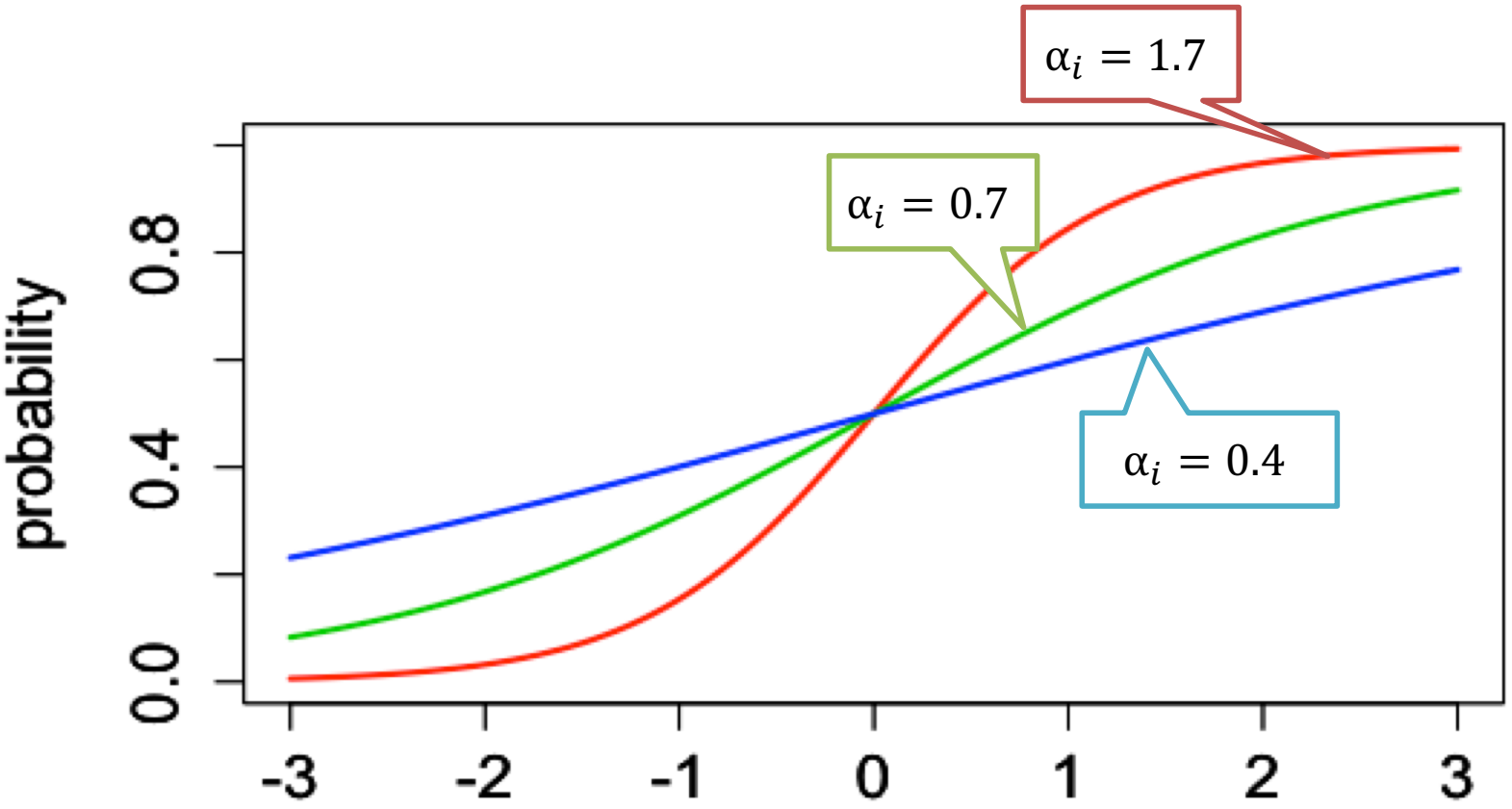
$$P(x_j = 1|\theta_i) = \frac{1}{1 + \exp(-1.7a_j\theta_i + b_j)}$$

項目iの識別力

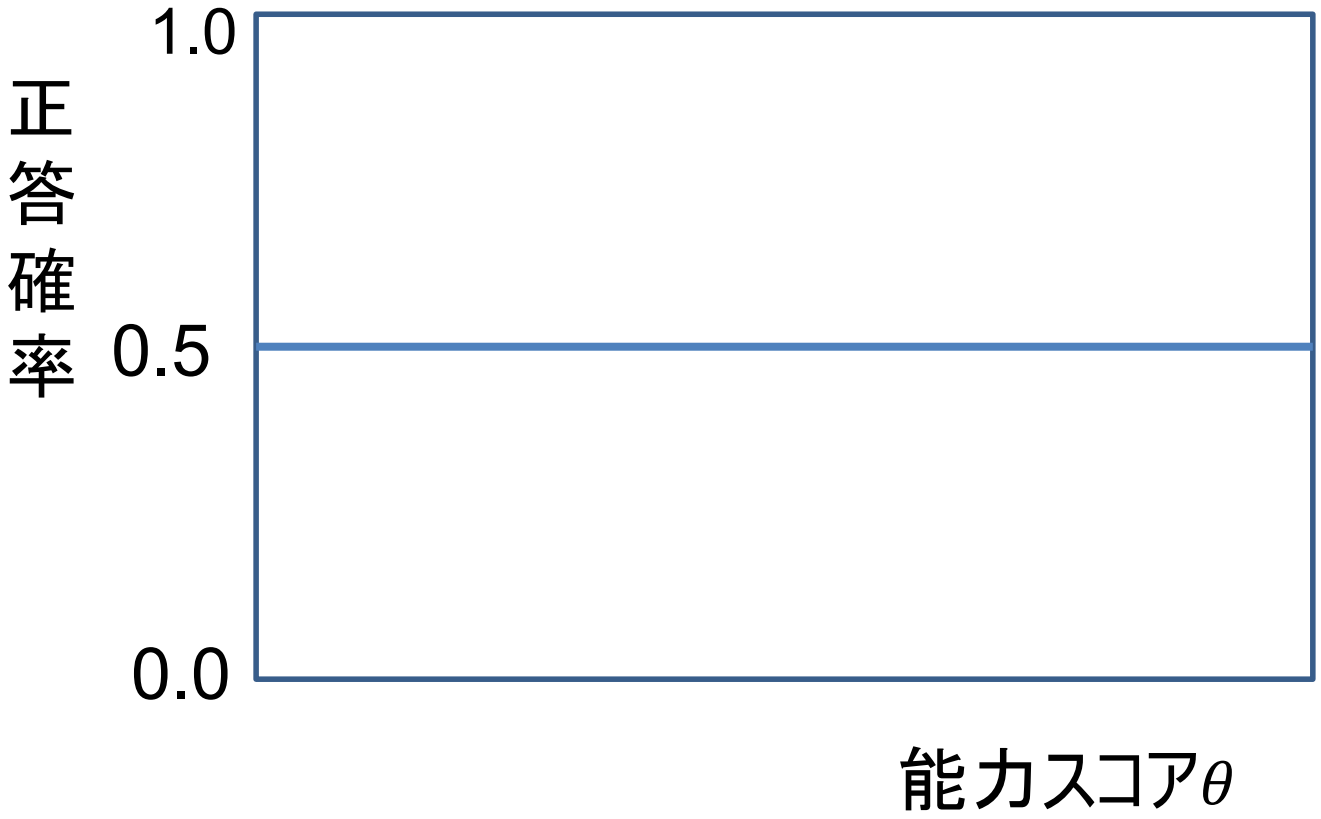
受験者の能力

項目iの難易度

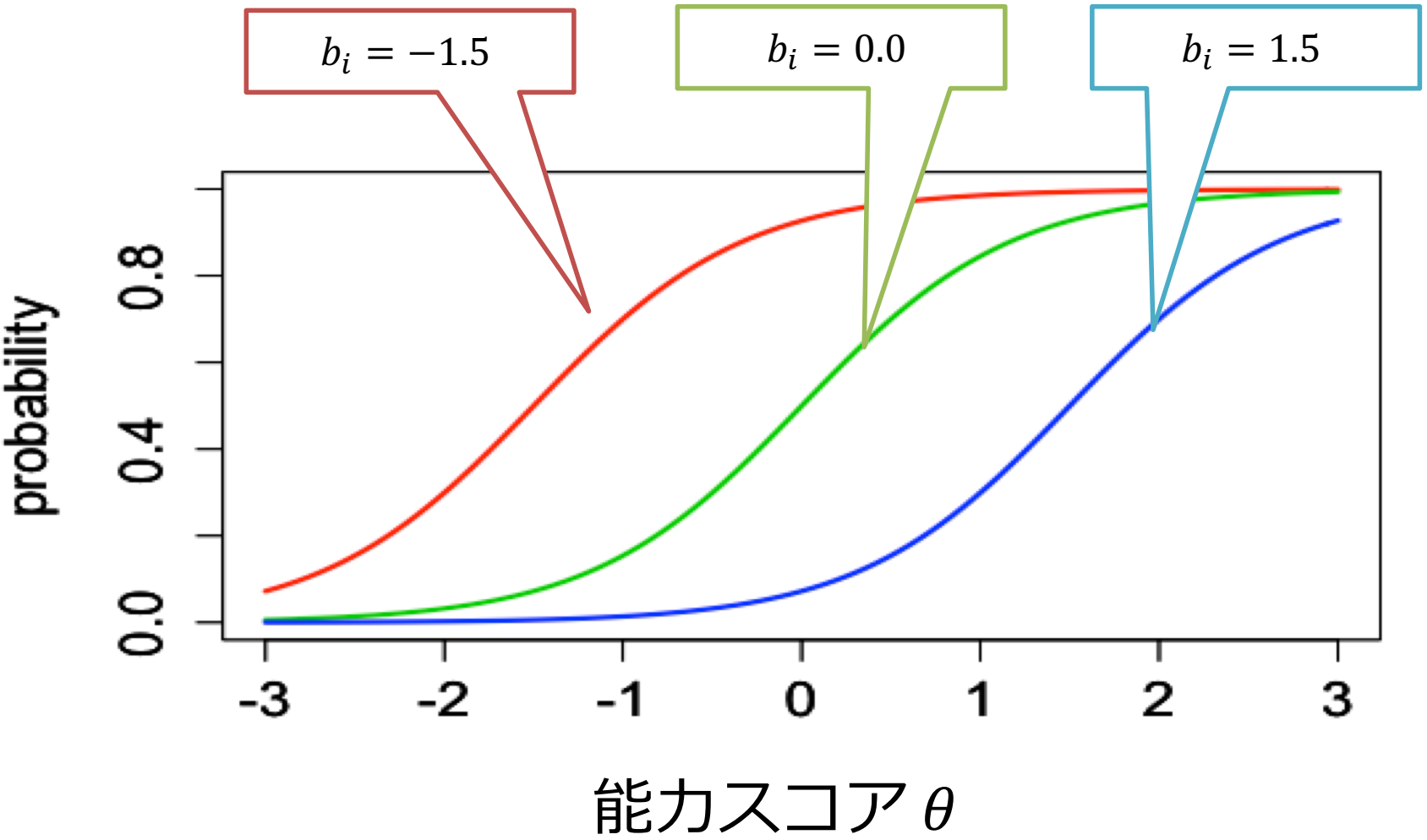
識別力パラメータa_j


$\alpha_i = 1.7$
$\alpha_i = 0.7$
$\alpha_i = 0.4$
probability
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
能力スコア θ

識別力パラメータ$a_j = 0$



難易度パラメータb_j



4. パラメータの推定とアイテムバンク

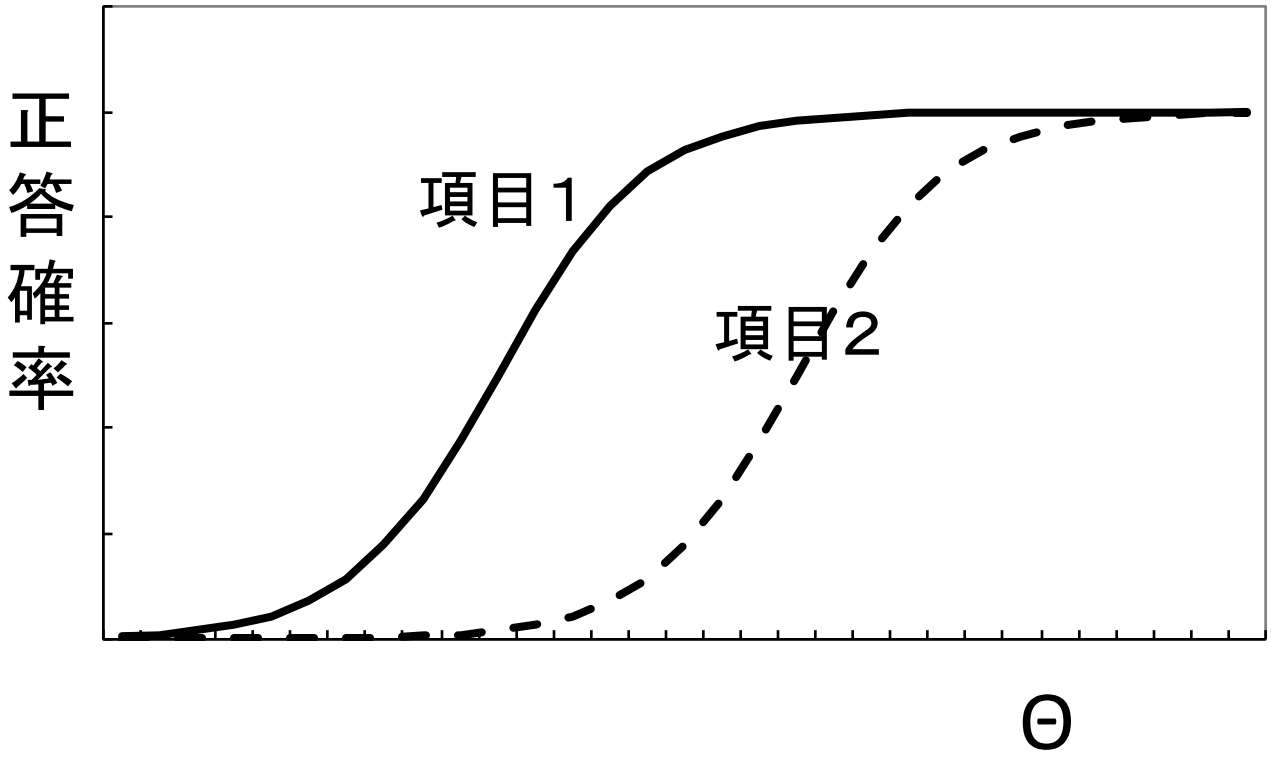
- あらかじめ受験者集団にテストを受験させ、そのテスト反応データから、パラメータ推定値を求める。
- ベイズ推定の普及により安定した予測精度の高いパラメータ推定値を得ることができる。
- 項目データベースである　アイテムバンクに　項目内容や正答のほかに、項目反応理論における推定値を格納しておく。
- アイテムバンクの項目は非公開！！

理想的には

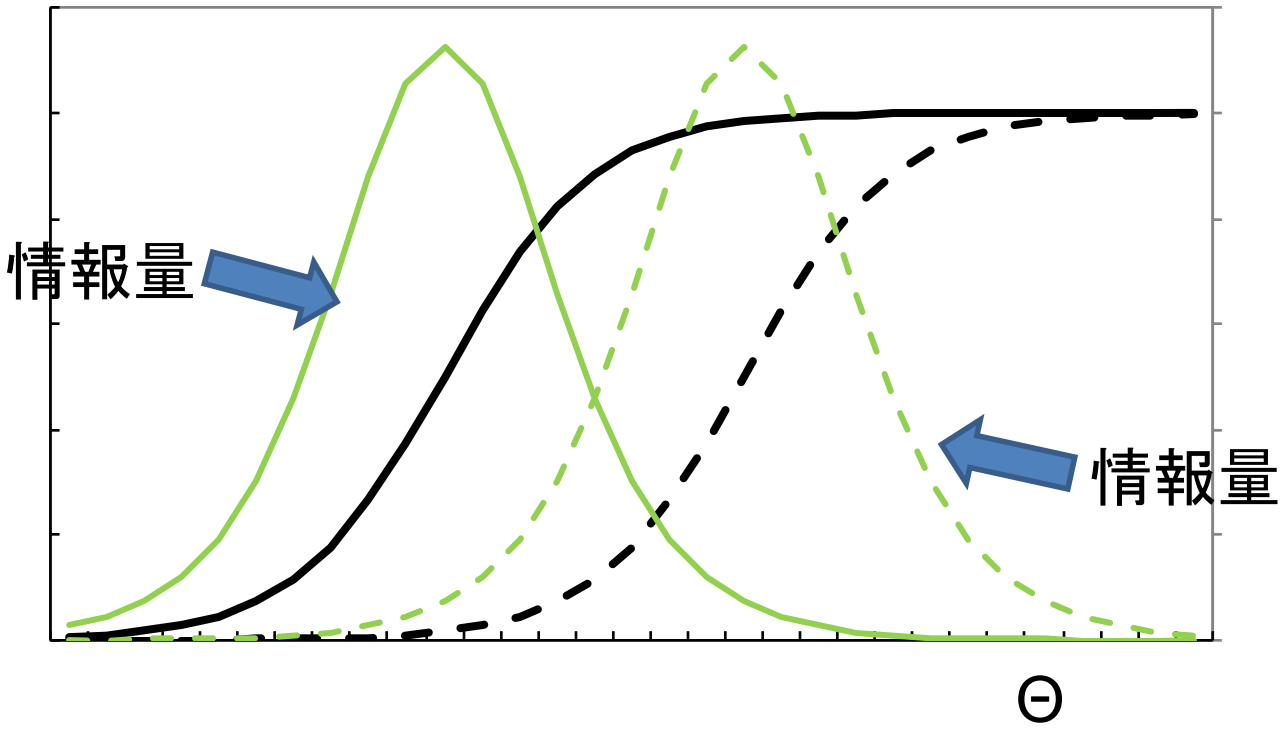
難易度が受検者の能力スコアに近く、識別力の高い項目を出題すればよい。

項目情報量の高い項目を出題

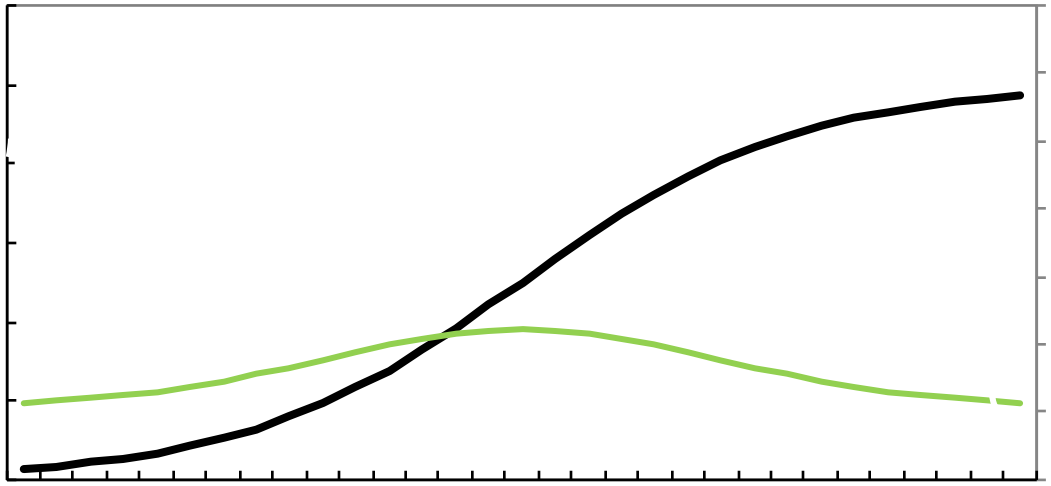
項目反応曲線と情報量関数



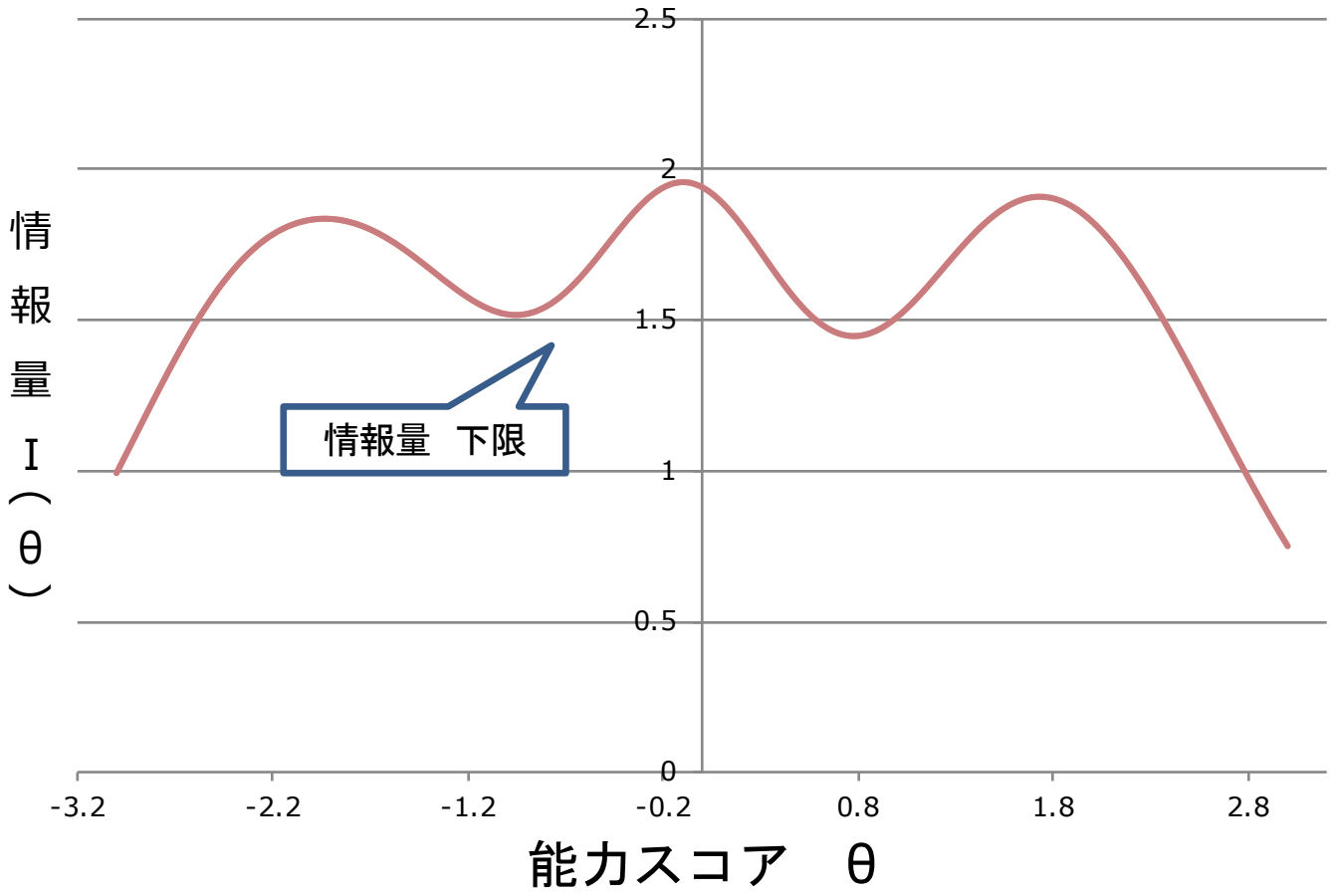
項目反応曲線と情報量関数



識別力の低い項目の情報量関数



テスト情報量＝テストを構成する項目情報量の和



情報量とは誤差の逆数

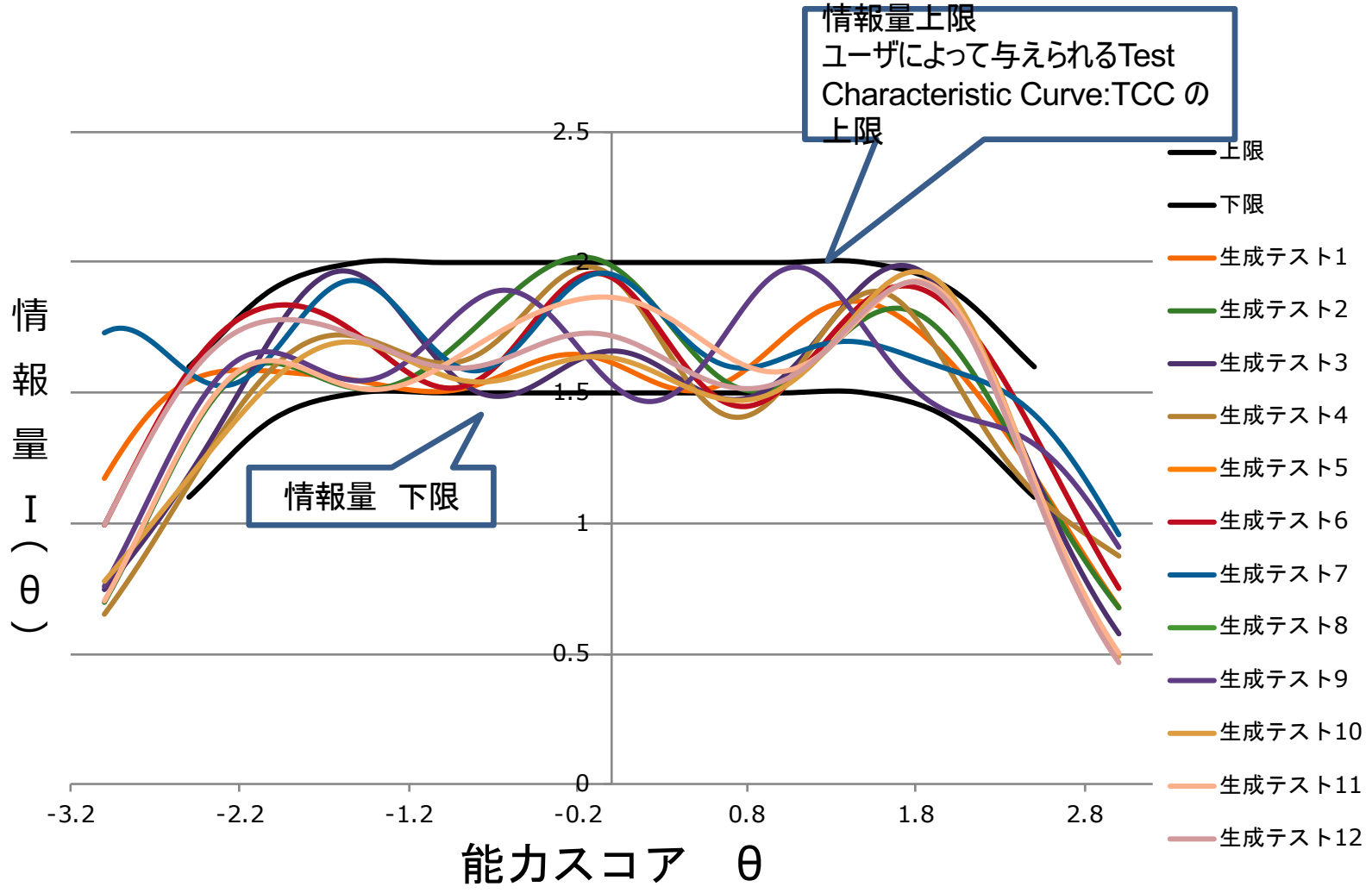
数学的な性質として情報量の逆数が能力スコアの測定誤差に収束することが知られている。

情報量の大きなテストを作成するとスコアの誤差の小さなテストができる。

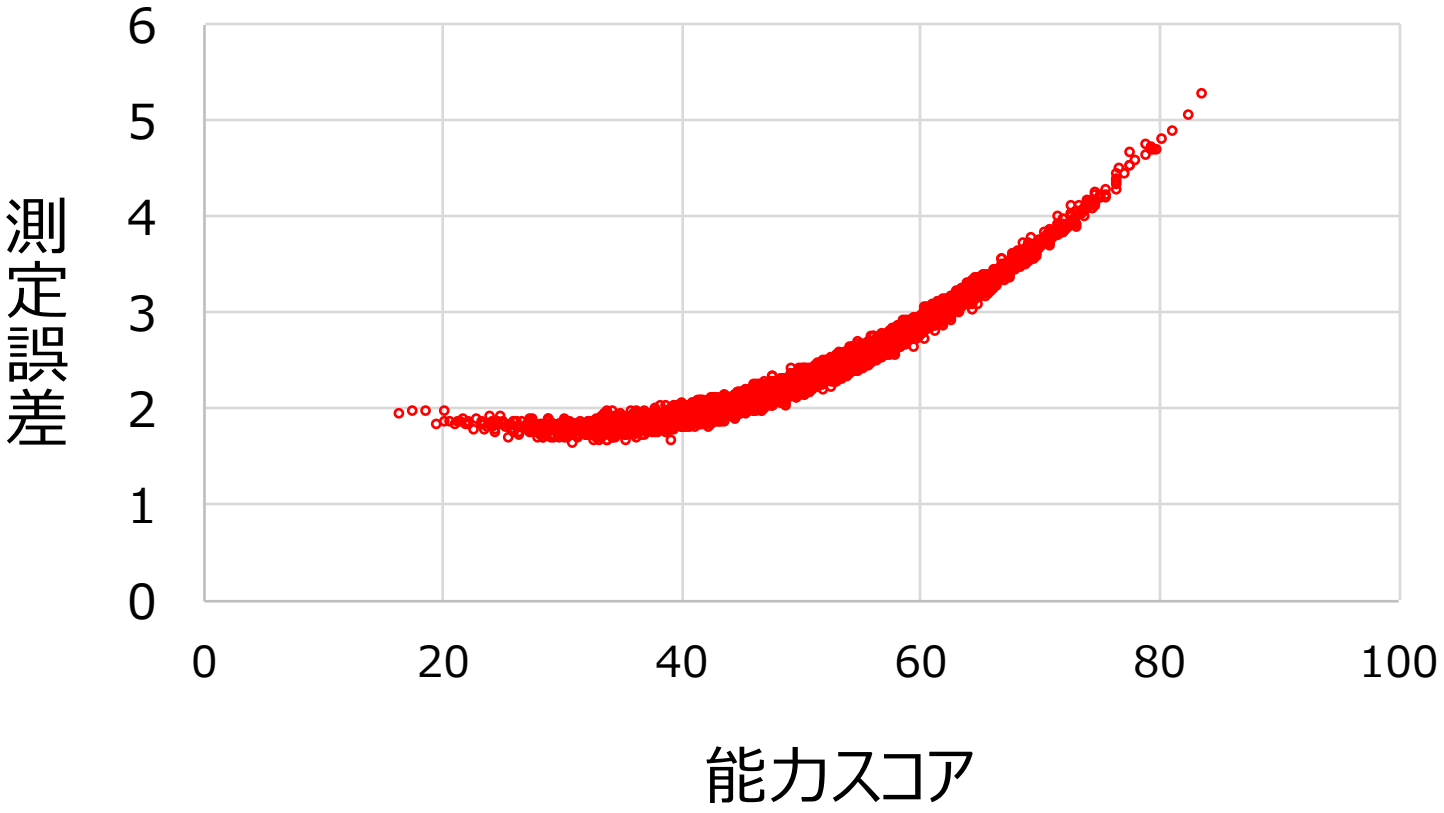
すべてのテストの情報量を大きく等質にできれば、同一受検者が異なるテストを受けても同一得点スコアを返すことができる。

信頼性の高いテストの実現！！

上限と下限を決めて等質化



ある公的試験の実際の等質テストのスコアの測定誤差



5. eテスティング技術
等質テストの項目組み合わせを最適化

どのように等質テストの項目組み合わせを求めるのか？

①Big Shadow Test

W. J. van der Linden, : Liner Models for Optimal Test Design, Springer (2005)

線形計画法を用いた計算量の小さい実用的手法。欠点として誤差が大きい。

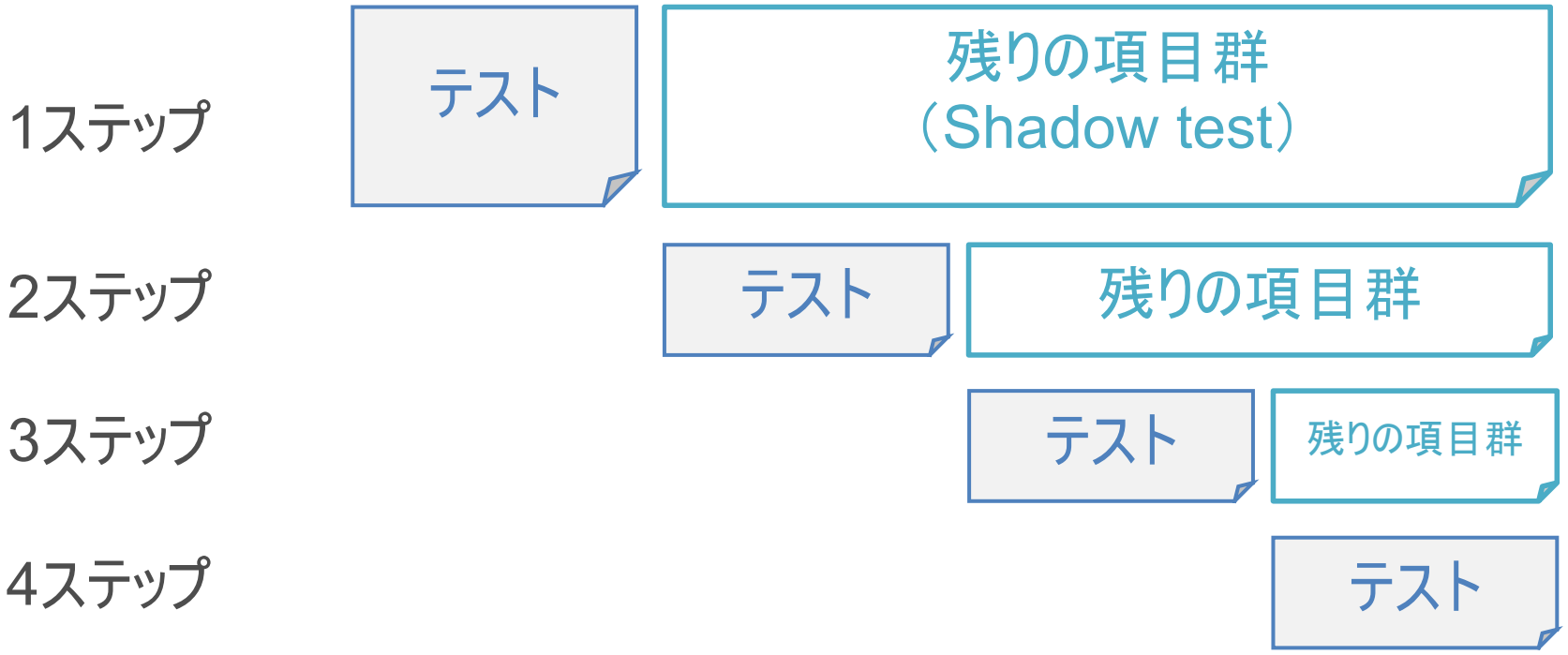
②最大クリーク手法

Takatoshi Ishii, Pokpong Songmuang, Maomi Ueno, “Maximum Clique Algorithm and its approximation for Uniform Test Form Assembly”, IEEE Transactions on Learning Technologies, Vol.7(1), pp.83-95, 2014.

最大クリーク問題として解く。計算量は大きいが、厳密解が求められる。

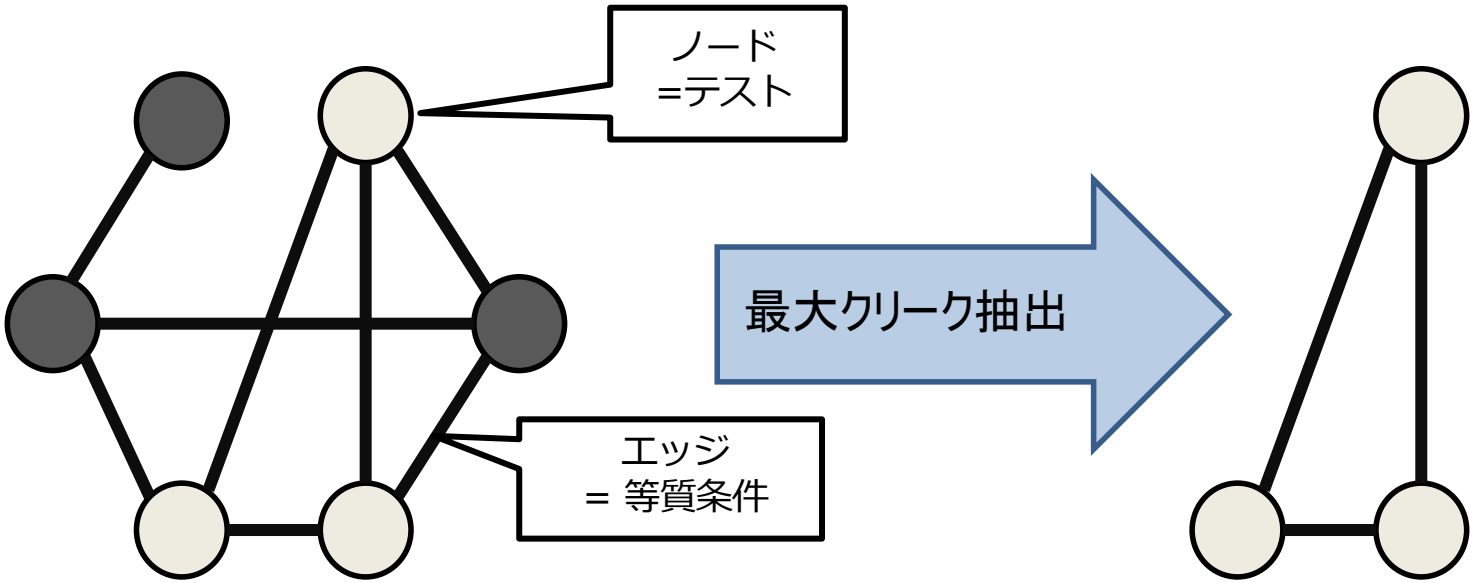
①Big shadow test(W. J. van der Linden2005)

1. アイテムバンクから逐次的にテストを構成
2. 「構成中のテスト」と「アイテムバンクの残りの項目群の平均」のテスト情報量の差異を線形計画法により最小化



②最大クリーク法（現在 最も正確で最大のテスト構成数を可能にする）

2018年度電子情報通信学会論文賞受賞



1. Takatoshi Ishii, Maomi Ueno, “Algorithm for Uniform Test Assembly Using a Maximum Clique Problem and Integer Programming”, International Conference on Artificial Intelligence in Education (AIED), LNAI 10331, pp. 102–112. 2017
2. Takatoshi Ishii, Pokpong Songmuang, Maomi Ueno, “Maximum Clique Algorithm and its approximation for Uniform Test Form Assembly”, IEEE Transactions on Learning Technologies, Vol.7(1), pp.83-95, 2014.

アイテムバンク500項目でのテスト構成数

重複項目数	Big Shadow法	最大クリーク法
0	15	10
1	19	22
2	19	65
3	19	223
4	19	936
5	19	4324
6	19	19817
7	19	61740
8	19	93678
9	19	99469
10	19	99979

アイテムバンク1000項目でのテスト構成数

重複項目数	Big Shadow法	最大クリーク法
0	25	17
1	39	61
2	39	282
3	39	1585
4	39	9793
5	39	46162
6	39	90127
7	39	99396
8	39	99979
9	39	99998
10	39	100000

アイテムバンク2000項目でのテスト構成数

重複項目数	Big Shadow法	最大クリーク法
0	61	32
1	79	186
2	79	1436
3	79	12456
4	79	62424
5	79	96859
6	79	99891
7	79	99991
8	79	100000
9	79	100000
10	79	100000

等質テストで何ができるか？

- 毎回の異なるテストが等質であることが保障され、いつでも何度でもテストを受けることができる。
- テスト実施後にその精度が評価でき、テストの評価をその都度行なえる。
- １００点満点の素点を受検者に返せ、それをもとに合否やランキングを決めることができる。

Q&A　eテスティングでのスコアにはIRT能力推定値を使いますか？

- 英語や医療系テストのようにIRTスコアが世界的に標準的な科目ではIRTの能力推定値をスコアに用いることが多い。IRTはそもそも真の能力を測定するためにサイコメトリック分野で開発された数理モデル。
- eテスティングではすべてが等質テストであるので素点が用いることができる。

Q&A　IRTは推定値に誤差が大きい？

IRTは推定値に誤差が大きいので使えないという人がいます。しかし、信頼性の概念は　同一能力の受検者が異なるテストを受験しても同等のスコアを返すことです。ですので一般のテスト得点は必ず誤差があります。その誤差を小さくするためにIRTやアイテムバンクを用いるのです。また、IRTは現在では十分に確立された技術でもあります。正しいモデルと推定法を用いれば真の同時確率分布を漸近的に推定でき、予測精度を非常に高く推定できるのです。ただ、誤用してしまうと誤差が大きくなってしまう危険性があります。正しい知識を持ちましょう。

Q&A eテスティング技術はテスト技術なので教育に役立たない？

No！！一般にIRTにおける能力推定値は一つの値を返すだけで学習者へのフィードバックとしては弱く見えるかもしれません。しかし、eテスティングは、受検者の各問題項目への学習者の反応の予測精度の高いモデルを採用していることを思い出してください。IRTを含むこれらのモデルは学習者の未知の項目への理解度の予測精度が高く、学習者がどこで分からなくなっているかなどを精度高く予測できるモデルなのです。これらは学習者に適応して教育するアダプティブラーニングのエンジンとして採用されています。

Q&A IRTモデルはどのモデルを使えばよいですか？

- 一般的には2パラメータロジスティックモデルでよい。同時確率分布が真の分布に収束することが証明できる。また、DeepLearningの研究では、離散分布の近似で微分可能なロジスティック関数が他の離散分布より精度が高いことが近年明らかになっている。学習者の未知の項目への反応予測の精度が高くどこで行き詰っているかなどのアダプティブラーニングにも有効である。
- 多肢選択式テストの場合は3パラメータロジスティックモデルを適用してもいい。モデルの説明力と予測精度は2パラメータと同等。本当に当て推量で正答する受検者がいる場合、それが識別力を下げているのを補正することができ、解釈性を向上させる。
- ラッシュモデルや１パラメータロジスティックモデルは基本　テスト素点と同じ情報で見た目をIRTスコアに見せているだけで予測精度が低く、結果として信頼性も低くなる。
- eテスティングでは等質テストに予測精度の高い手法が必要でIRTにこだわらずDeep Learningなど様々な手法が提案されている。

Q&A　IRTのパラメータ推定はどうすればよいですか？

- MLE(Maximum Likelihood)
  尤度を最大にするパラメータ
- MAP(Maximum a posteriori)
  ベイズ事後分布を最大にするパラメータ
- ◎ 推奨　EAP(Expected a posteriori)

ベイズ事後分布におけるパラメータの期待値（予測を最大化する推定値）

（MCMC法(Markov chain Monte Carlo、変分ベイズ法、ガウス近似法)

Q&A アイテムバンク構築のためにリンケージ（等化）はどのようにすればよいですか？

線形変換の代表的な算出手法

- **Mean/Sigma法**：共通項目の困難度の平均値と標準偏差に基づいてリンケージ係数を求める方法
- **Mean/Mean法**：共通項目の平均困難度と平均識別力に基づいてリンケージ係数を求める方法
- **項目特性曲線法**：共通項目の項目特性曲線に基づいてリンケージ係数を求める方法
  （代表的な方法：Haebara法，Stocking&Load法）

Q&A アイテムバンク構築のためにリンケージ（等化）はどのようにすればよいですか？

線形変換の代表的な算出手法

- **Mean/Sigma法**：共通項目の困難度の平均値と標準偏差に基づいてリンケージ係数を求める方法
- **Mean/Mean法**：共通項目の平均困難度と平均識別力に基づいてリンケージ係数を求める方法
- **項目特性曲線法**：共通項目の項目特性曲線に基づいてリンケージ係数を求める方法
  （代表的な方法：Haebara法，Stocking&Load法）

Q&A アイテムバンク構築のためにリンケージ（等化）はどのようにすればよいですか？

数学的には　Θ全体について線形変換は成り立たない。大きな誤差を避けられない。個々のΘの値を条件とした上で変換しなければならない。

実際には、テスト構成時に全体のテスト項目数の５％から１０％に匹敵する新規項目を付け加えてテスト出題し、採点には用いず推定にのみ用いる。既知の項目パラメータのみを用いて受検者の能力値パラメータを推定する。新たに推定された能力値パラメータを所与として、テストデータより未知の新項目のパラメータを推定する。これらのプロセスを　所望の推定精度まで異なる受検者に実施し、パラメータ推定精度が十分になるとその項目はアイテムバンクに格納する。採点に用いない項目が出題されていることは受検者には知らせない。

おしまい