テストのデータサイエンス

植野真臣
電気通信大学　情報理工学研究科

今後のスケジュール（予定）

4月7日　授業の概要とガイダンス
4月14日　ベイズの定理
4月21日　ベイズはどのように誕生したか？
4月28日　　ベイズはコンピュータ、人工知能の父である！！
5月12日　　アランチューリングとベイズ
5月19日　　ビリーフとベイズ
5月26日　　尤度と最尤推定(1)
6月2日　　尤度と最尤推定(2)
6月9日　　ベイズ推定と事前分布(1)
6月16日　ベイズ推定と事前分布（２）
6月 23日　　データサイエンス：ルービン因果推論
6月30日　　テストのデータサイエンス
7月7日　　階層ベイズとデータサイエンス
7月14日　ベイジアンネットワークと因果推論
7月28日　　国際会議で休講
8月 4日　　テストと総括

テストのデータサイエンス

異なる問題から構成されるが、同一受検者が受検すると同一得点を返す複数の異なるテストの構成を行いたい。

メリット

1．毎回、等質なので不公平がない。

2．いつでもどこでも何度でも受けられる試験が実現できる。

↓

これは違うテストを受検した受検生の得点を予測して公平・公正に受検者の実力を比較する（信頼性の高いテスト）というデータサイエンスの問題である。異なる背景（異なるテストを受検）を持つ受検者をバイアスを調整して公正に比較するというデータサイエンスの問題。

1．テストの信頼性とは？

同一の能力の受検者に　同一テスト（異なる項目）を行った場合、同一の測定値が得られるか？

異なる項目より構成されたテストの評価の問題

Ａ君とＢ君はそれぞれＡ組、Ｂ組に所属している。同一科目でも教員が異なり、異なるテスト問題によって構成されるテストを受験している。

A君の数学の得点は75点　Ｂ君の数学の得点は80点

問題

このとき、Ｂ君のほうがＡ君より　数学の学力が高いといえるか？

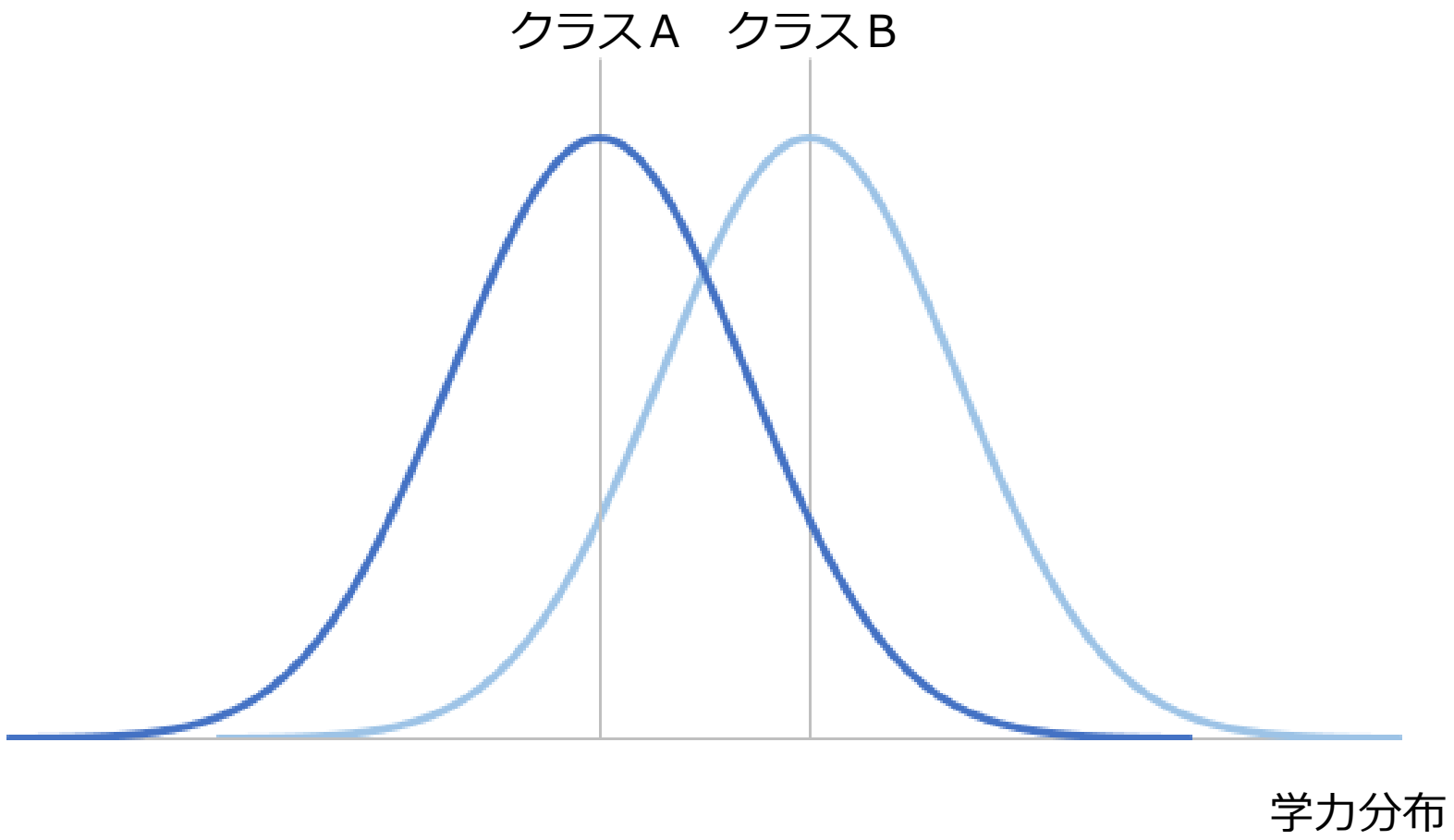
正解

Ａ君、Ｂ君の受けた二つの異なるテストは　それぞれの難易度が違うかもしれないので何も言えません！！

偏差値を用いて平均と標準偏差をそろえればよい？

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

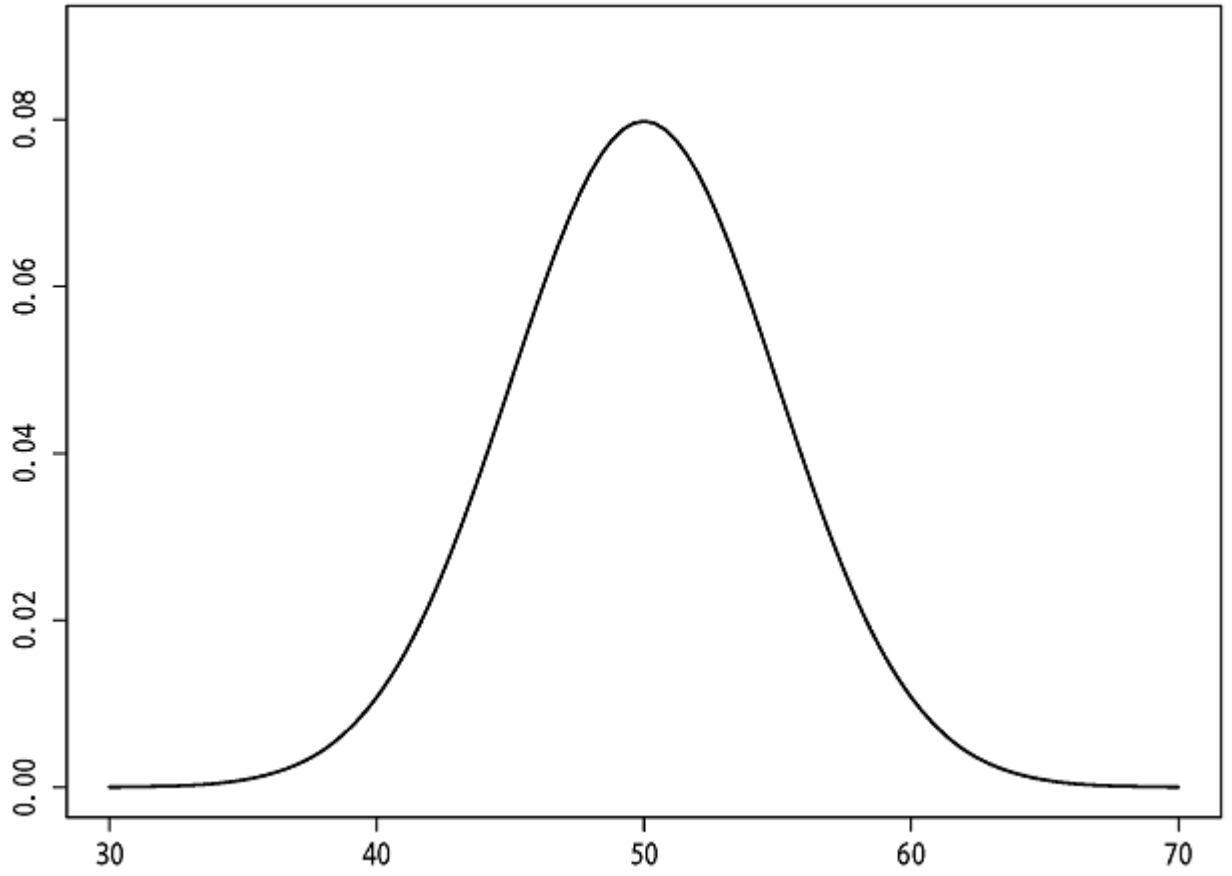
問題：ＡとＢの所属するクラスの学力分布に依存してしまう：クラスＡとクラスＢの学力分布が違う場合、偏差値は意味をなさない。



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。

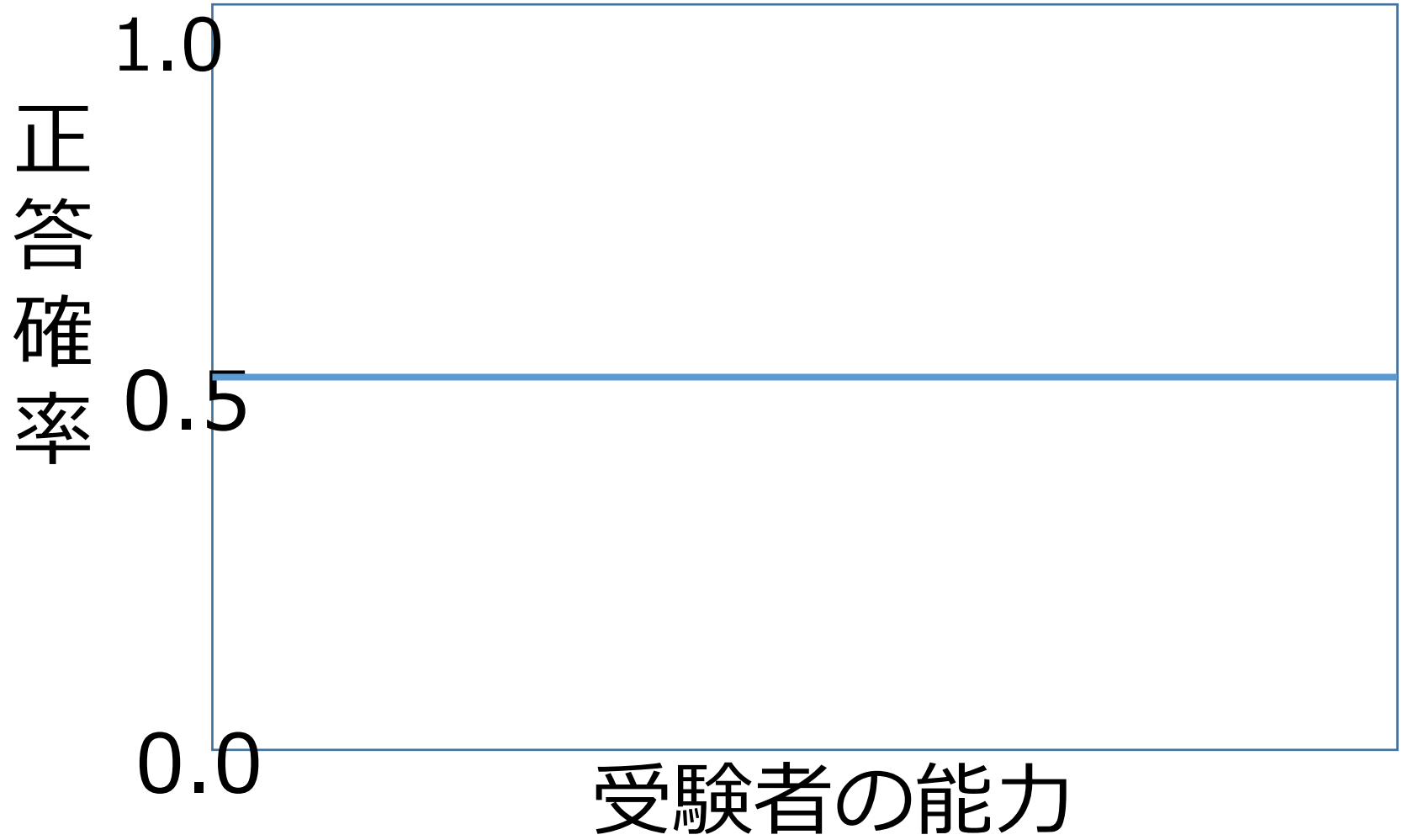
平均点を５０点の同一の得点分布にするように問題を作ればいい！！



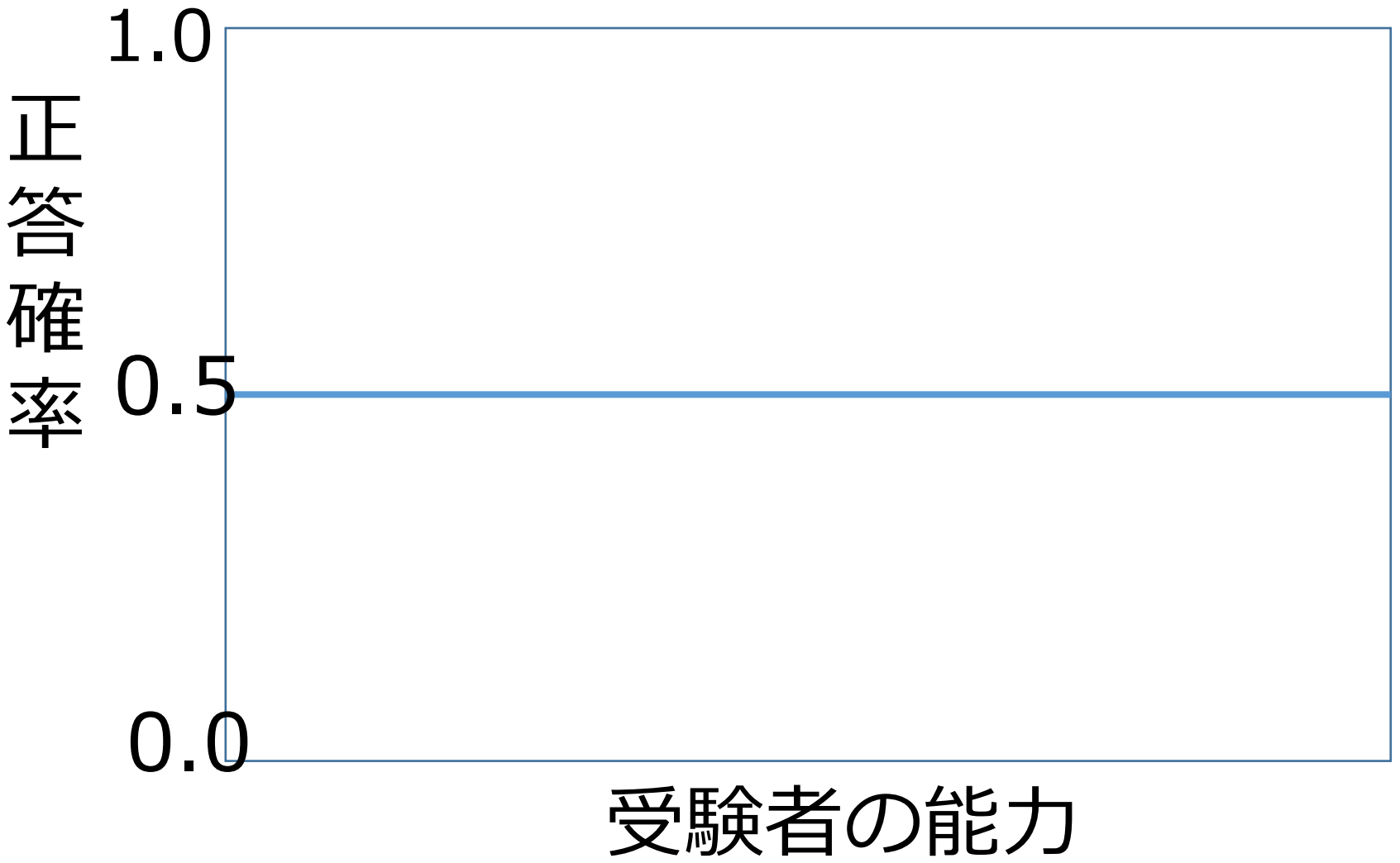
次の問題の正答確率分布

- 「今からコインを投げる。　表がでるか裏が出るか？」

何が問題か？



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(IRT)

- 項目反応理論
- Item Response Theory (IRT)

項目反応理論(IRT)
２パラメータロジスティックモデル

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

項目反応理論(IRT)

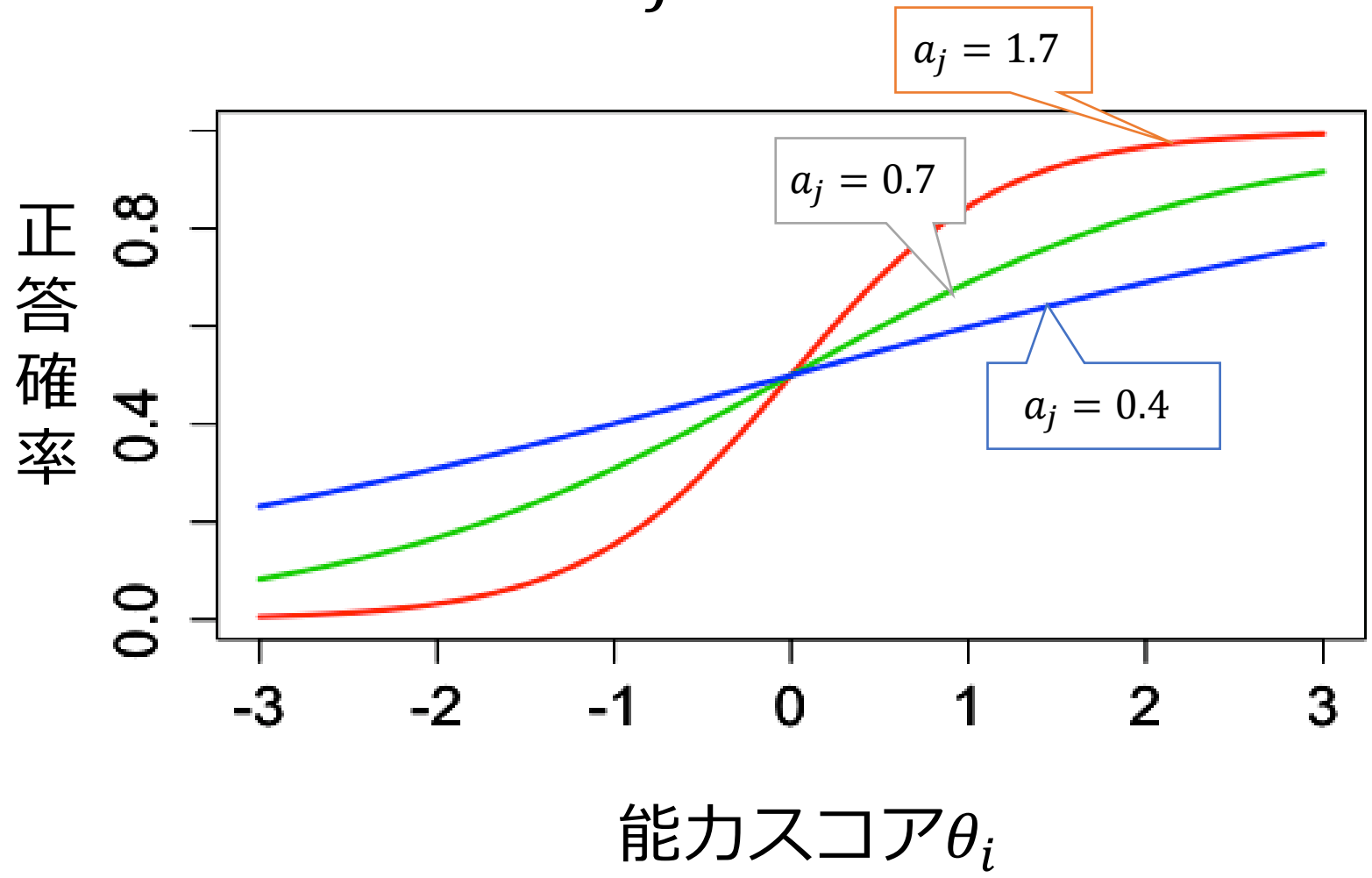
$$P(x_j = 1|\theta_i) = \frac{1}{1 + \exp(-1.7a_j\theta_i + b_j)}$$

項目jの識別力

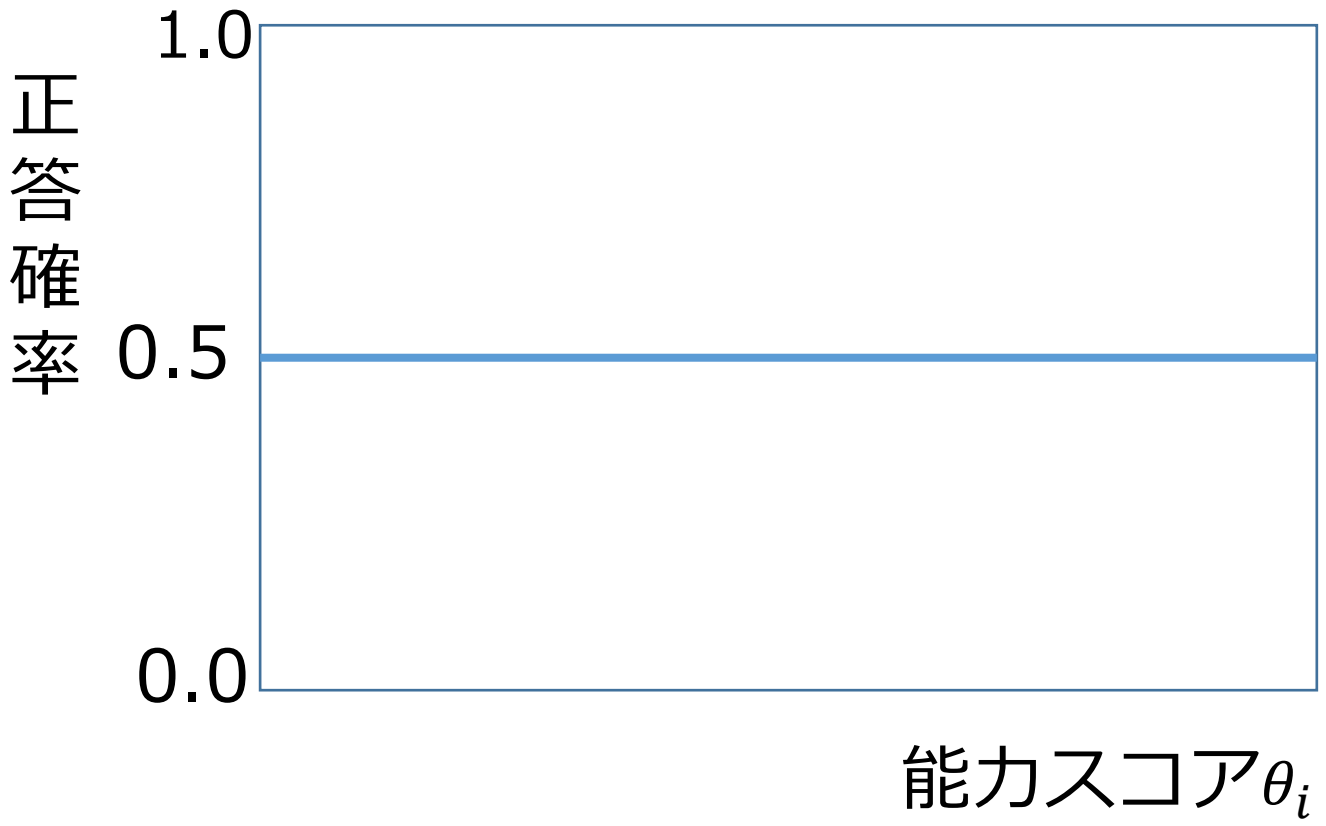
受験者iの能力スコア

項目jの難易度

識別力パラメータa_j



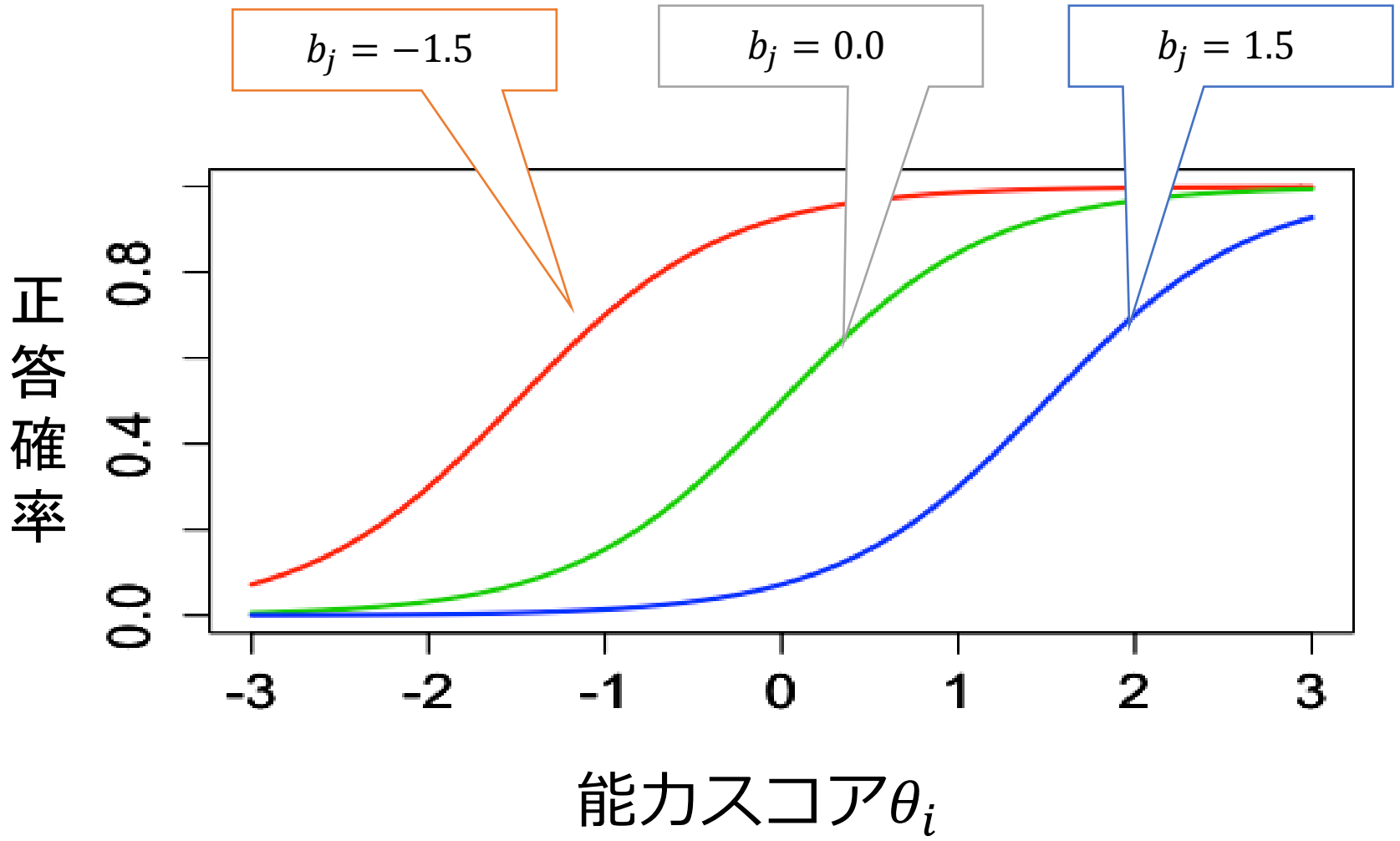
識別力パラメータ$a_j = 0$



つぎの問題とテストとしては同じ

- 「今からコインを投げる。　表がでるか裏が出るか？」

難易度パラメータb_j



パラメータの推定とアイテムバンク

- あらかじめ　異なる受験者集団にテストを受験させ、そのテスト回答データから、パラメータ推定値を求める。
- 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。
- テストはアイテムバンクより問題項目を複数抽出することにより構成する。
- アイテムバンクの項目は非公開！！

４．受検者のスコアΘ推定法の概要

受検者iの3問からなるテストへの正誤データX={正答, 正答,誤答}であったとする．能力値θ_iを持つ受検者iのこの正誤データの回答をする確率は以下の尤度$L(\theta_i|X)$に比例する．

$$L(\theta_i|X) = P(x_1 = 1|\theta_i)\ P(x_2 = 1|\theta_i)\ (1 - P(x_3 = 1|\theta_i))$$

事後確率は　　$p(\theta_i|X) = p(\theta_i)\ L(\theta_i|X)$

ここで$p(\theta_i)$ は能力値θ_iの事前分布$p(\theta_i) \sim N(0, 1^2)$ (標準正規分布)．そこで正誤データXを所与とした事後確率の期待値(EPA)

$$\mathrm{E}(\theta_i|X) = \int L(\theta_i|X)\, p(\theta_i) d\theta_i$$

を能力値θ_iの推定値とする．

対数尤度最大化のためのスコア関数

$$l = \ln L(\theta|x)$$

以下のθについてlを偏微分した関数=0となるθを求める.

$$\frac{\partial}{\partial\theta} l = \frac{\partial}{\partial\theta} \ln L(\theta|x) = 0$$

$\frac{\partial}{\partial\theta} l$をスコア関数と呼ぶ.

５．項目情報量：スコア関数の分散

スコア関数の分散

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta}l\right)=\mathrm{E}\left(\frac{1}{L(\theta|x)}\frac{\partial L(\theta|x)}{\partial\theta}\right)^2=-\mathrm{E}\left(\frac{\partial^2\ln L(\theta|x)}{\partial\theta^2}\right)$$

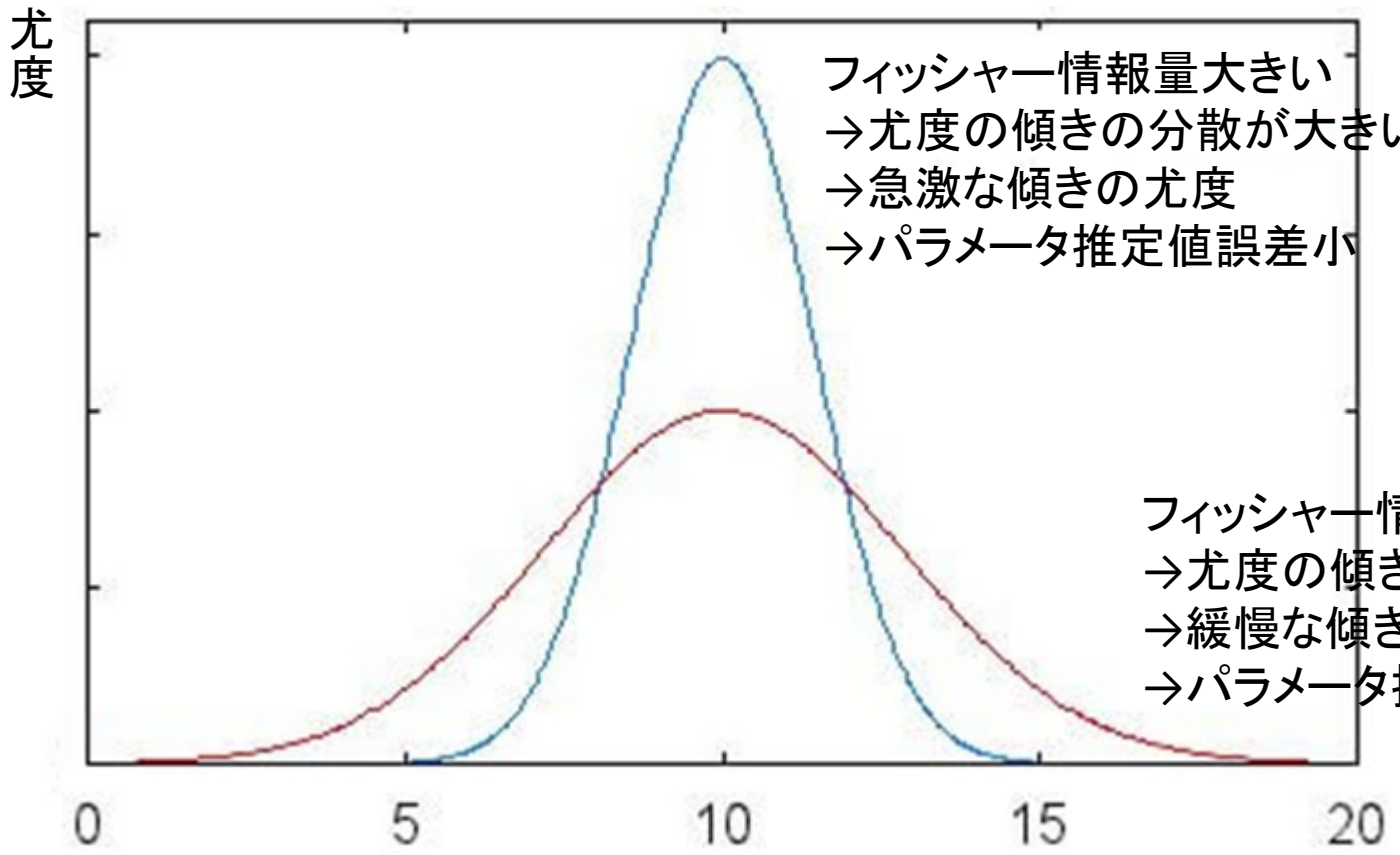
$$=a_j^2\ \mathrm{p}(x_j=1|\theta_i)[1-\ \mathrm{p}(x_j=1|\theta_i)]$$

統計学では　フィッシャー情報量と呼び,

テスト理論では項目情報量と呼ぶ.

$$I_j=a_j^2\ \mathrm{p}(x_j=1|\theta_i)[1-\ \mathrm{p}(x_j=1|\theta_i)]$$

フィッシャー情報量は推定値の信頼性を反映



フィッシャー情報量大きい
→尤度の傾きの分散が大きい
→急激な傾きの尤度
→パラメータ推定値誤差小

フィッシャー情報量小さい
→尤度の傾きの分散が小さい
→緩慢な傾きの尤度
→パラメータ推定値誤差大

情報量とは推定誤差の逆数の近似

数学的に情報量の逆数が漸近的に推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構成すると誤差の小さなテストができる。

最尤推定値の漸近正規性

定義

真の値θ^*の推定値 $\hat{\theta}$が**漸近正規推定量**（asymptotically normal estimator）であるとは，$\sqrt{n}(\hat{\theta} - \theta^*)$ の分布が正規分布に分布収束することをいう．すなわち，任意の$\theta^* \in \Theta^*$ と任意の実数に対して

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{as} N(0, \sigma^2(\theta^*))$$

$\sigma^2(\theta^*)$ を漸近分散（asymptotic variance）という．

最尤推定値の漸近正規性

定理

確率密度関数が正則条件（regular condition）の下で，微分可能のとき，

最尤推定量は漸近分散$I(\theta^*)^{-1}$をもつ漸近正規推定量である.

$$I(\theta^*) = E_\theta\left[\left(\frac{\partial}{\partial\theta} lnL(\theta|\mathbf{x})\right)^2\right]$$

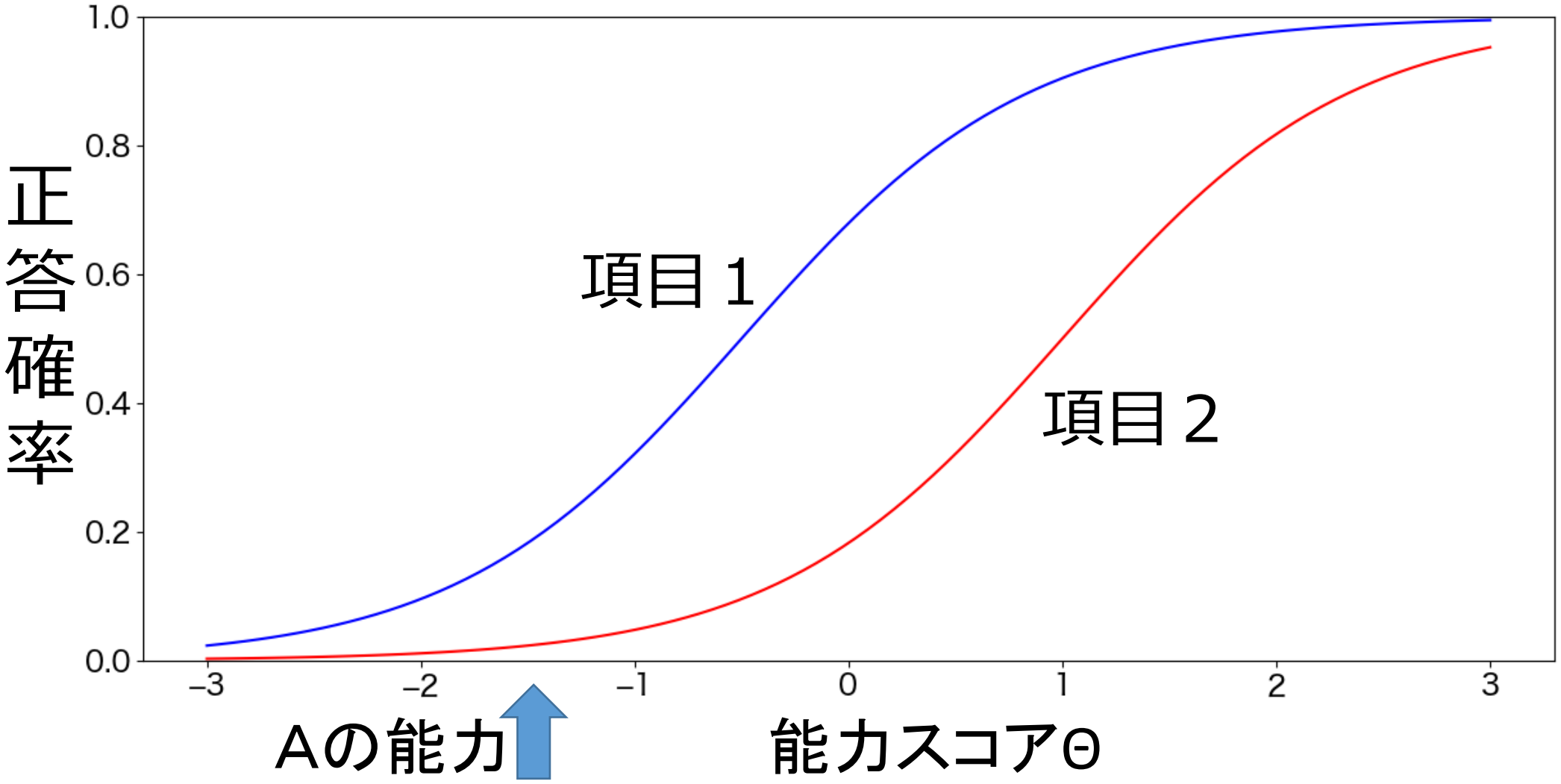
フィッシャー（Fisher）の情報量

証明

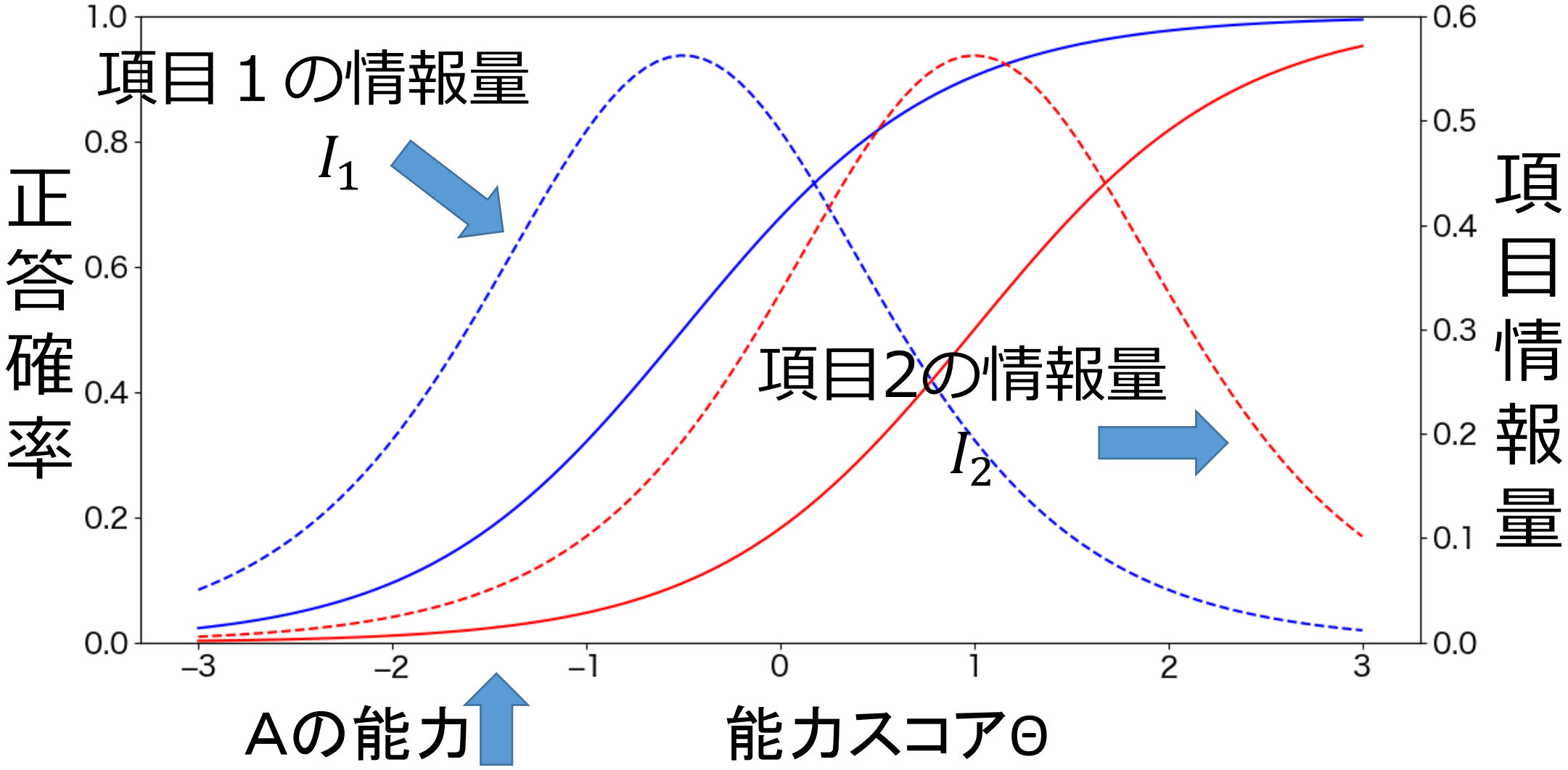
以下を参照

https://qiita.com/PePrs/items/8d758e38df7a68004304

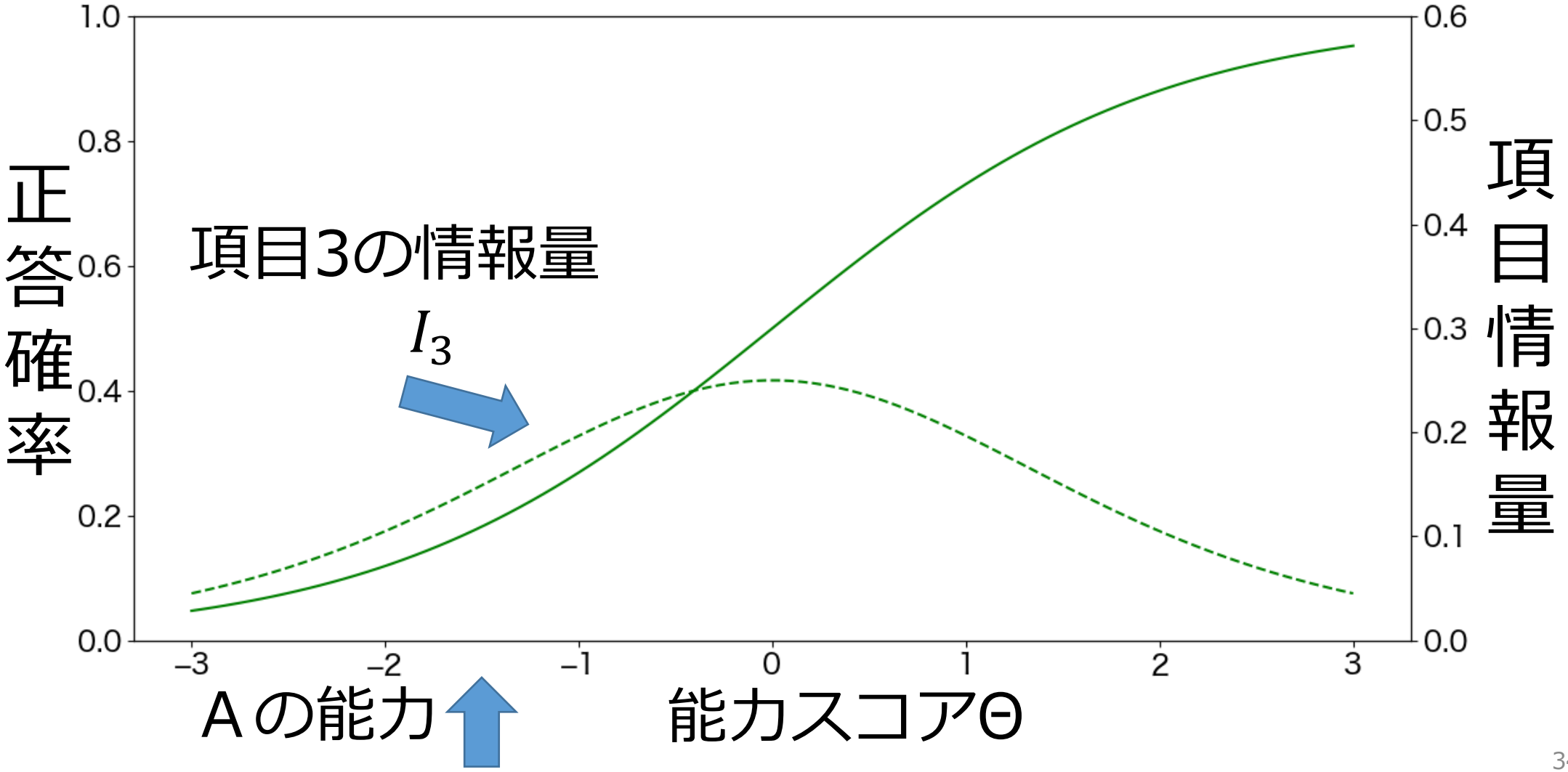
項目反応曲線と情報量関数



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

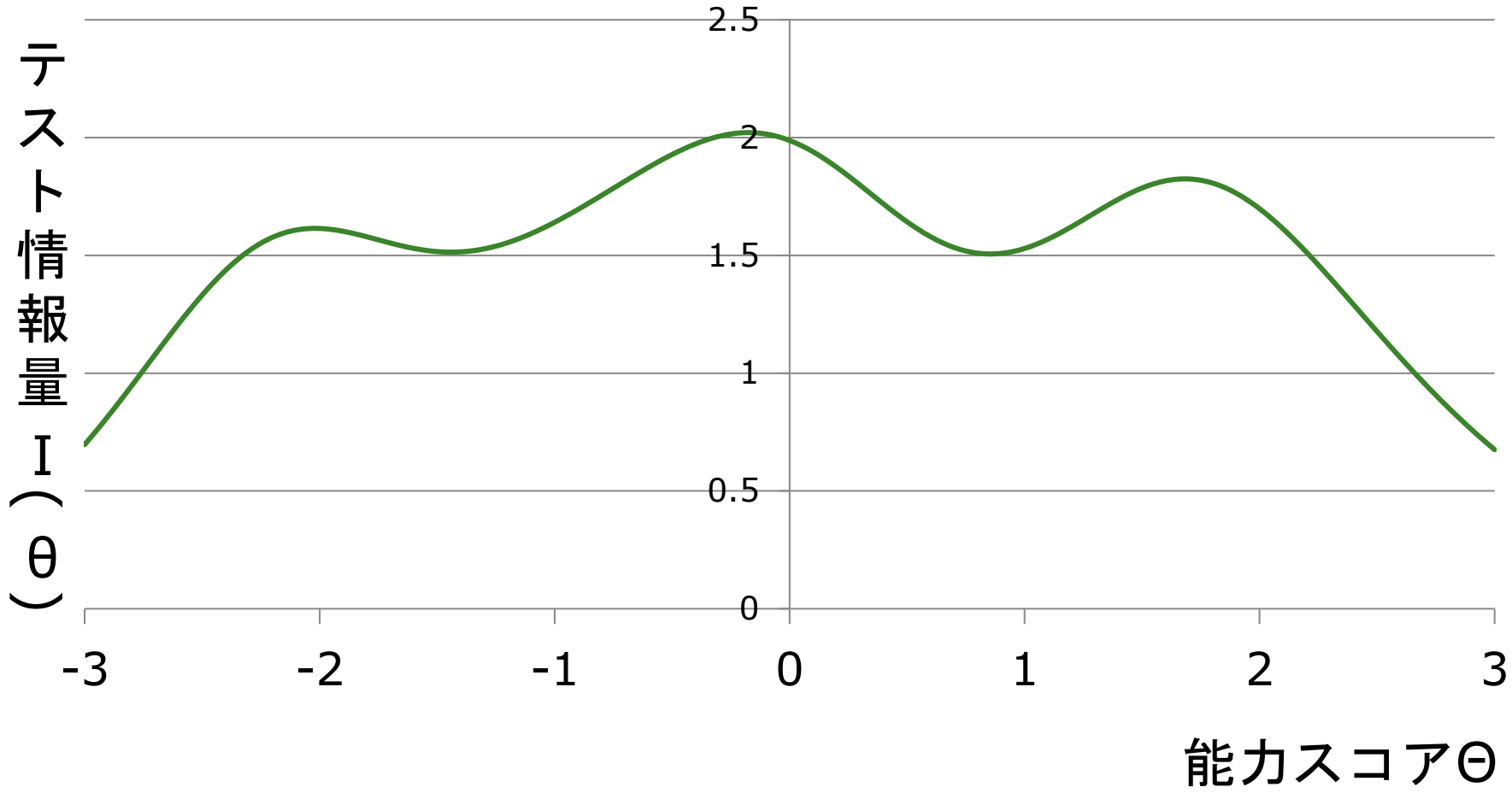


6. テスト情報量

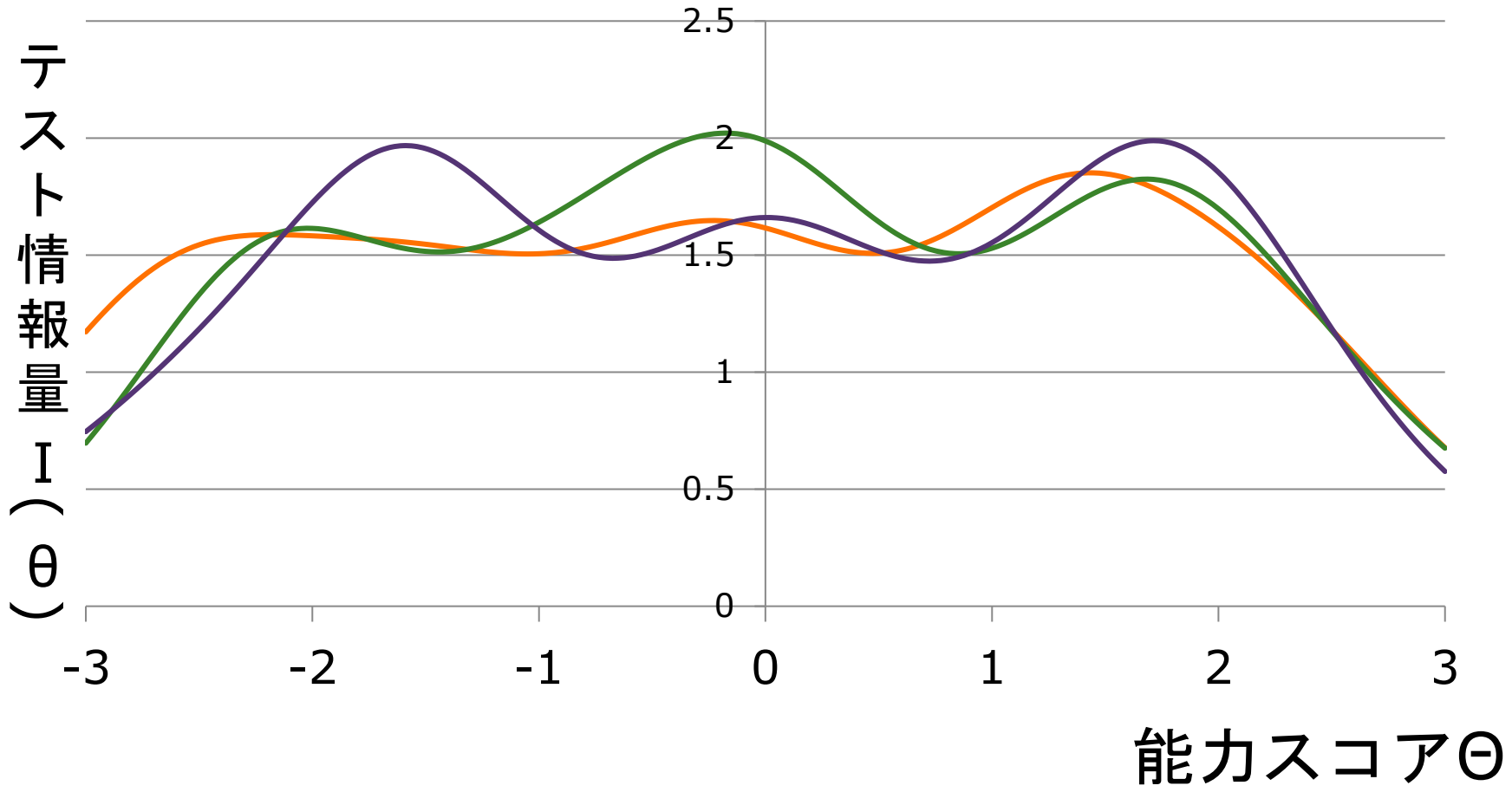
テスト情報量＝テストを構成する項目情報量の和

$$I_{Test} = \sum_{j \in Test} a_j^2 \ \mathrm{p}(x_j = 1|\theta_i)[1 - \ \mathrm{p}(x_j = 1|\theta_i)]$$

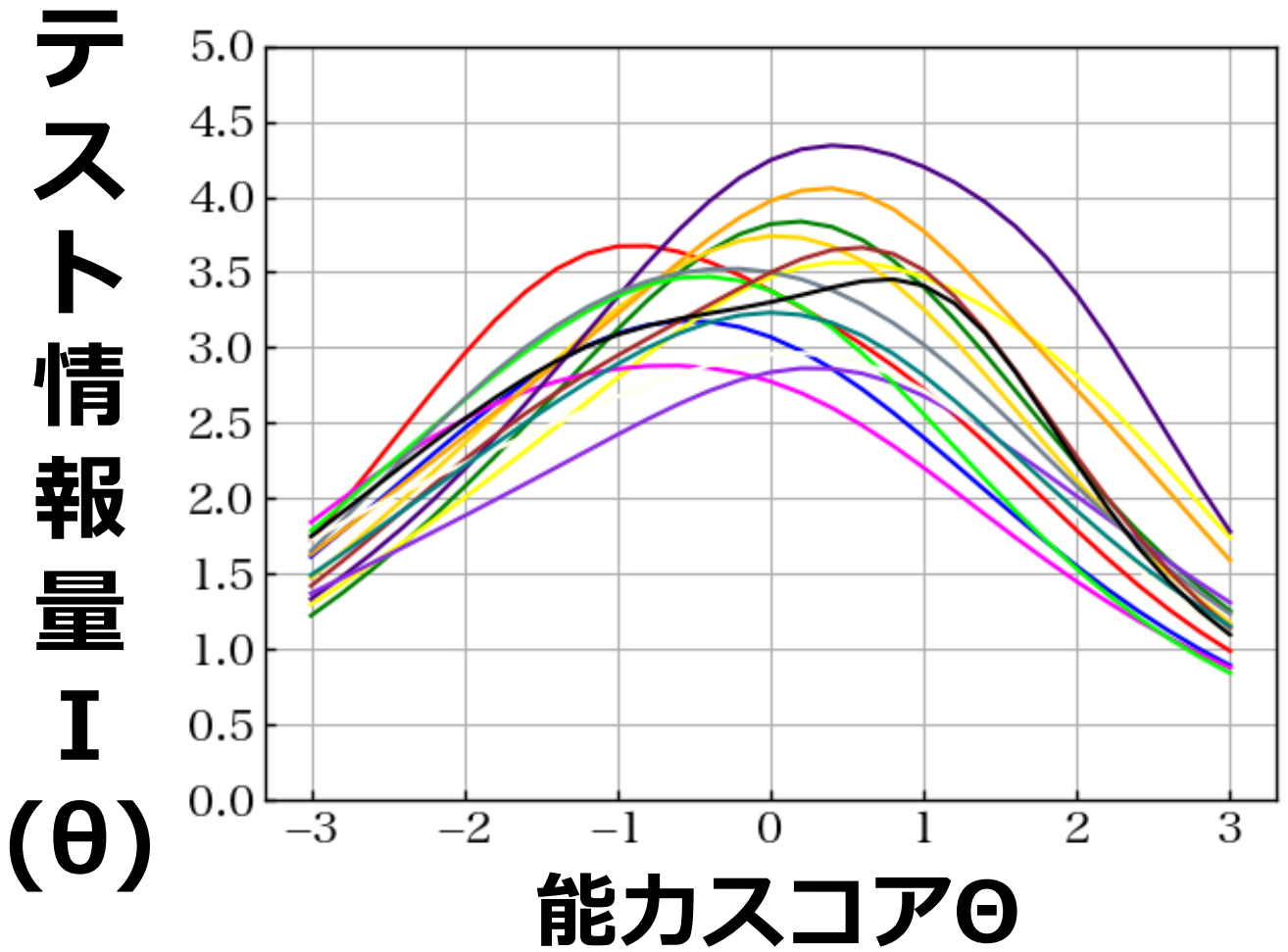
例　項目1,2,3で構成されるテストの情報量



３つの異なるテストの情報量曲線



実際の異なる10個のテストの情報量関数（950個のアイテムバンクより異なる25項目からなるテストをランダムに抽出）



７．テスト情報量とは推定誤差の逆数

数学的に情報量の逆数が漸近的に推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構成すると誤差の小さなテストができる。

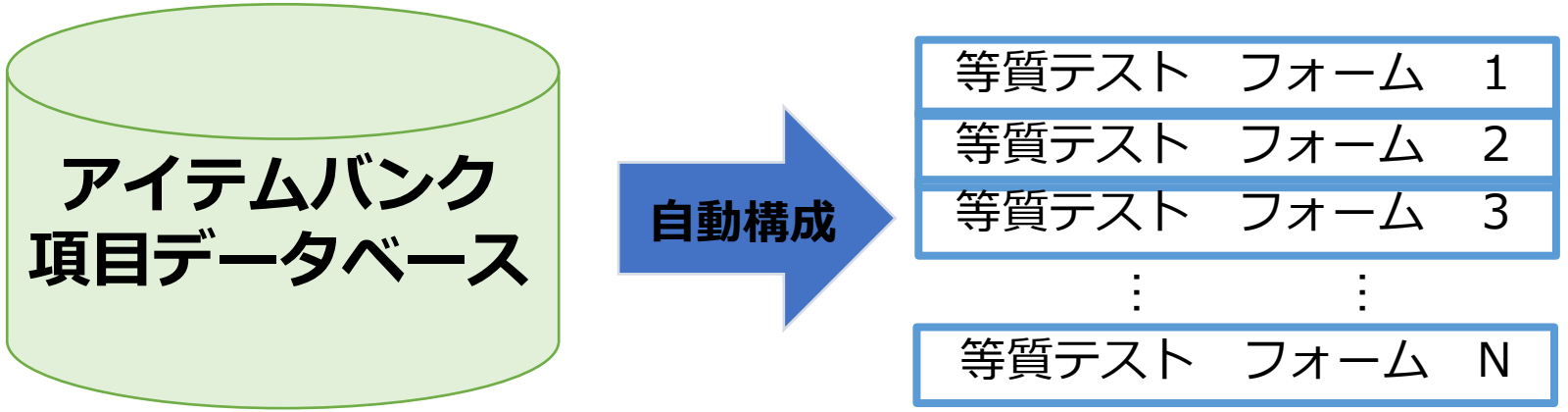
８．良いテストとは

良いテスト項目とは識別力が高く、難易度が合否判定ライン（クリティカルポイント）の受検者の能力スコアに近いように作成する。

→

項目情報量の高い項目をなるべく多く持つテストがテスト情報量が高く、スコアの誤差を小さくできる

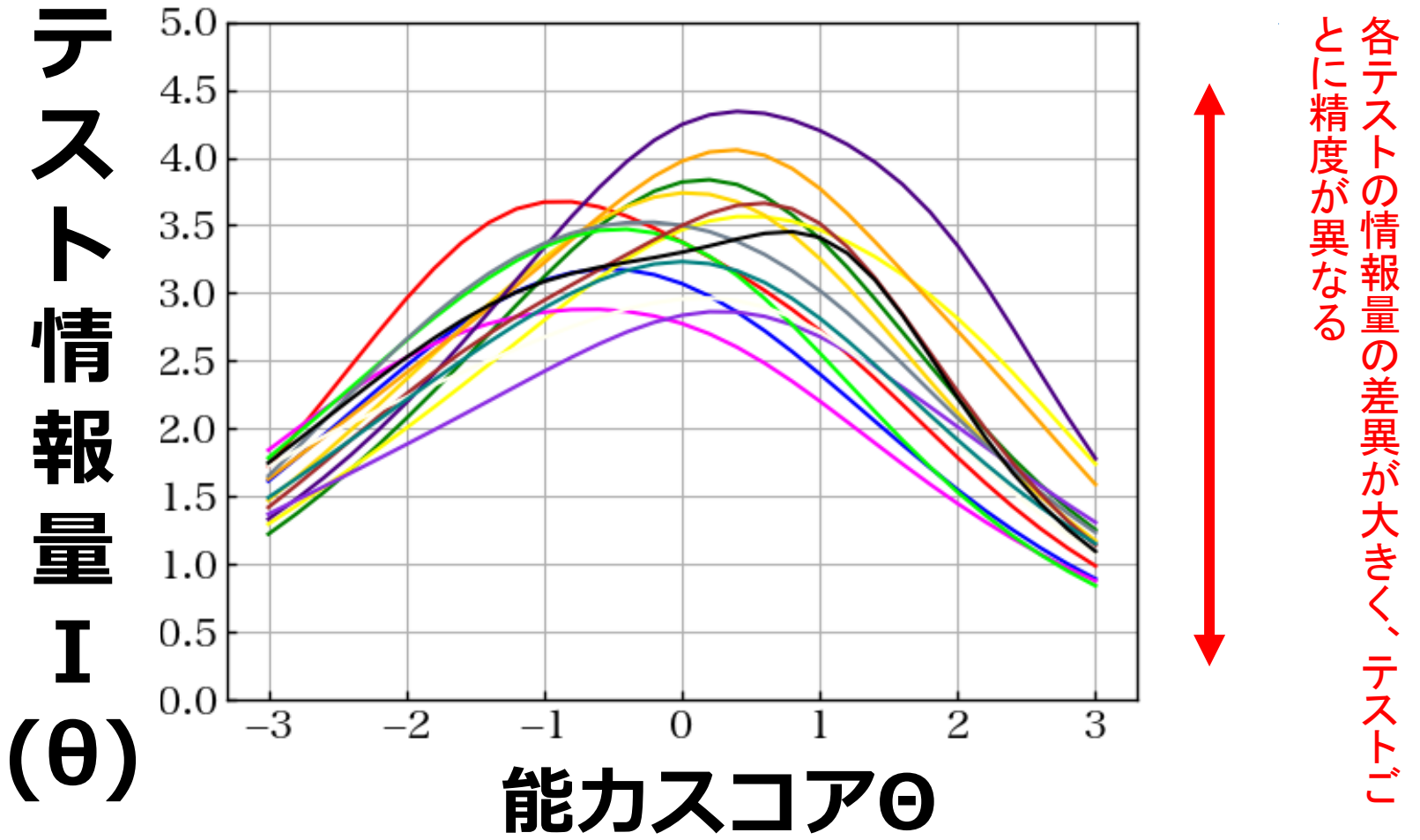
９．アイテムバンク方式（問題バンク方式）によるテスト構成



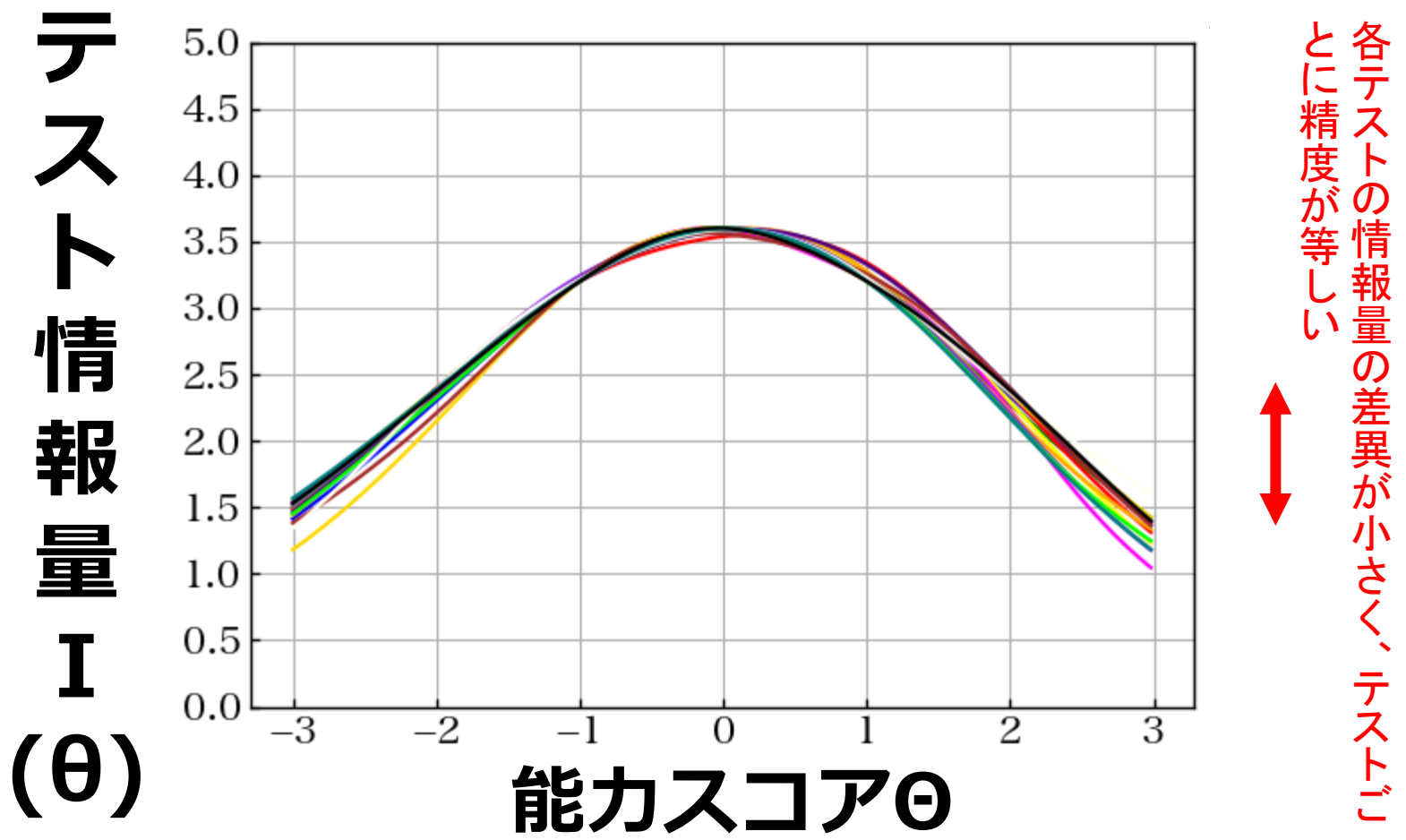
アイテムバンク（問題バンク）より，所望の条件を満たして等質になるように問題項目の組み合わせを自動的に抽出してテストを構成

Maomi Ueno: Advanced technologies for e-testing, IEEE Transactions on Learning Technologies, IEEE computer Society, Vol. 4, No. 3, 209-221（2003）

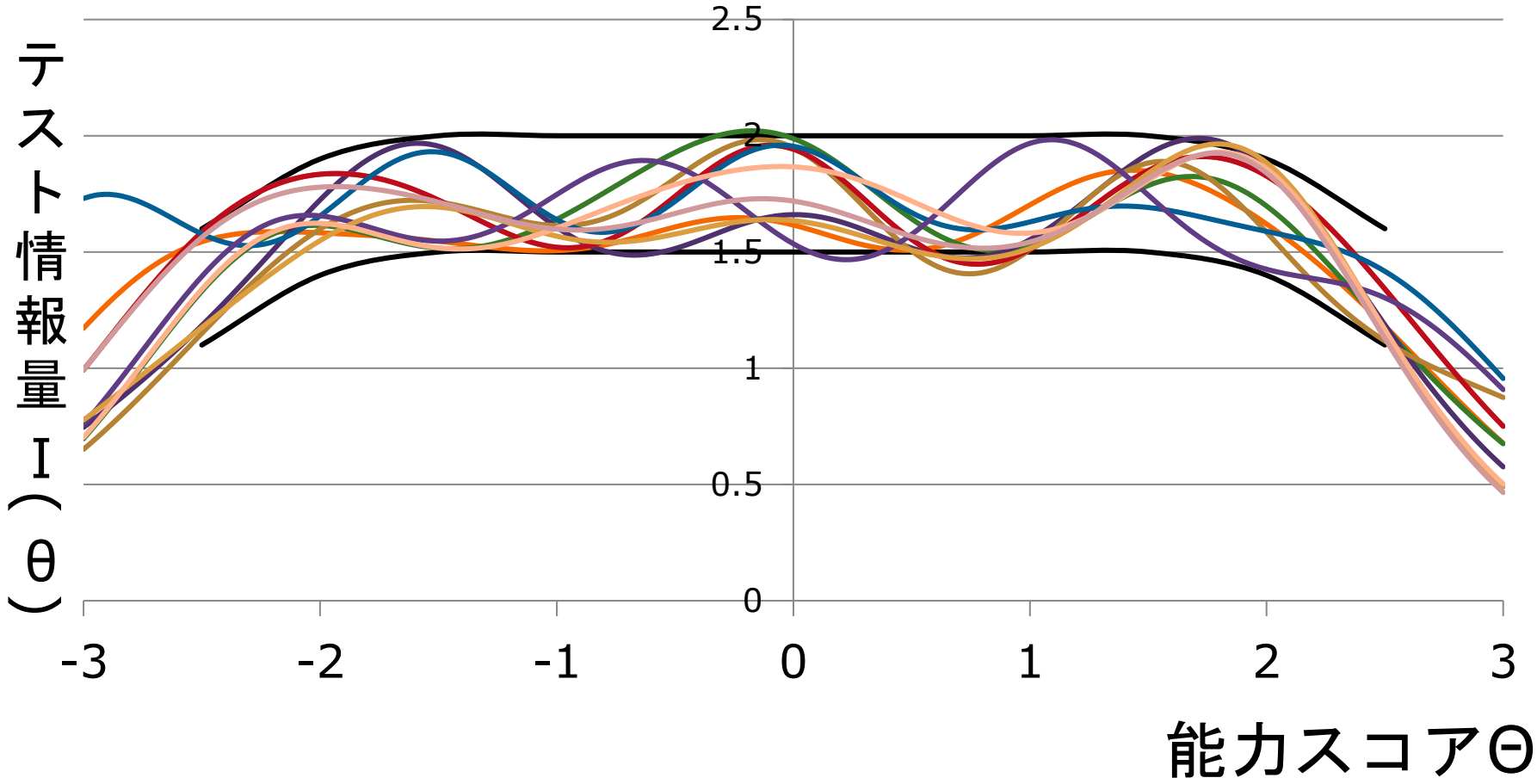
実際の異なる15個のテストの情報量関数（950個のアイテムバンクより異なる25項目の問題の組み合わせ（テスト）をランダムに抽出）



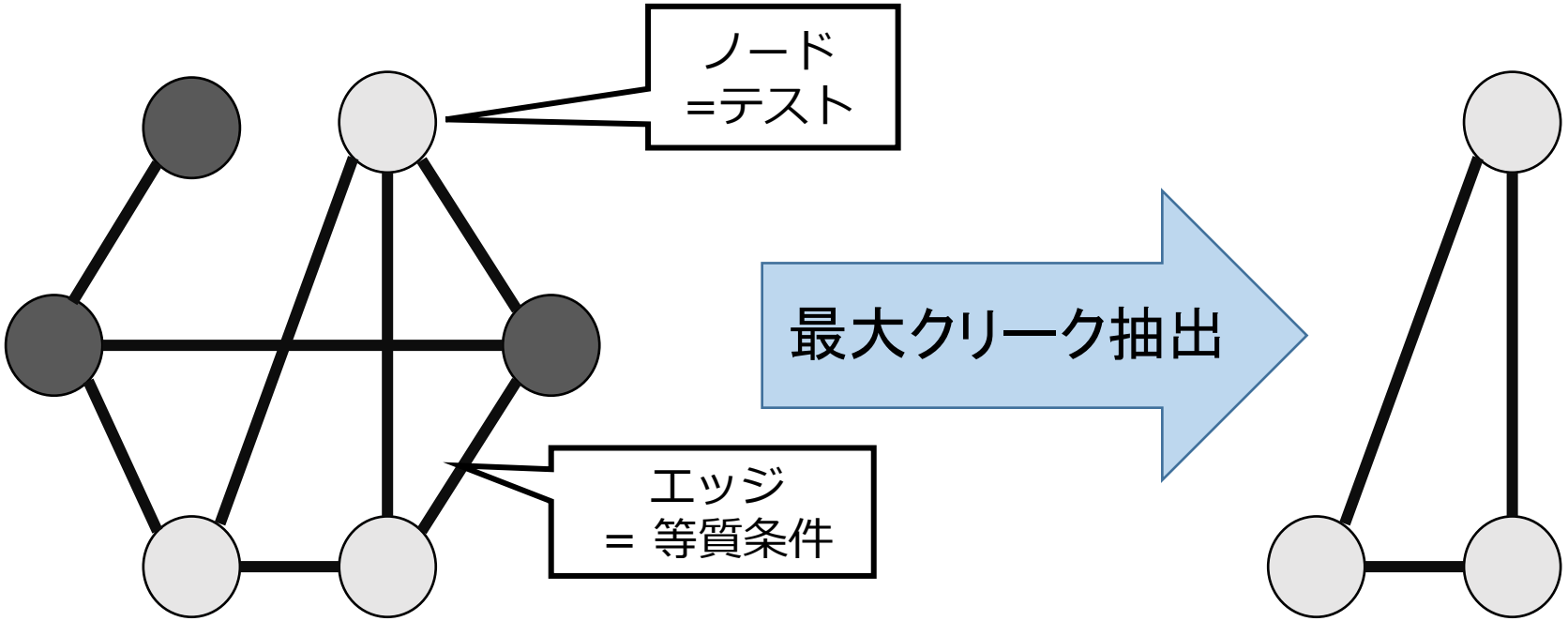
実際の異なる15個のテストの情報量関数
（ 950個のアイテムバンクより異なる25項目の問題の組み合わせ（テスト）を情報量が等質になるように抽出）



異なるテストが等質の情報量になるようにテストを構成（アイテムバンクより問題の組み合わせを抽出）する⇔　人工知能が自動的に構成（アイテムバンクより抽出）

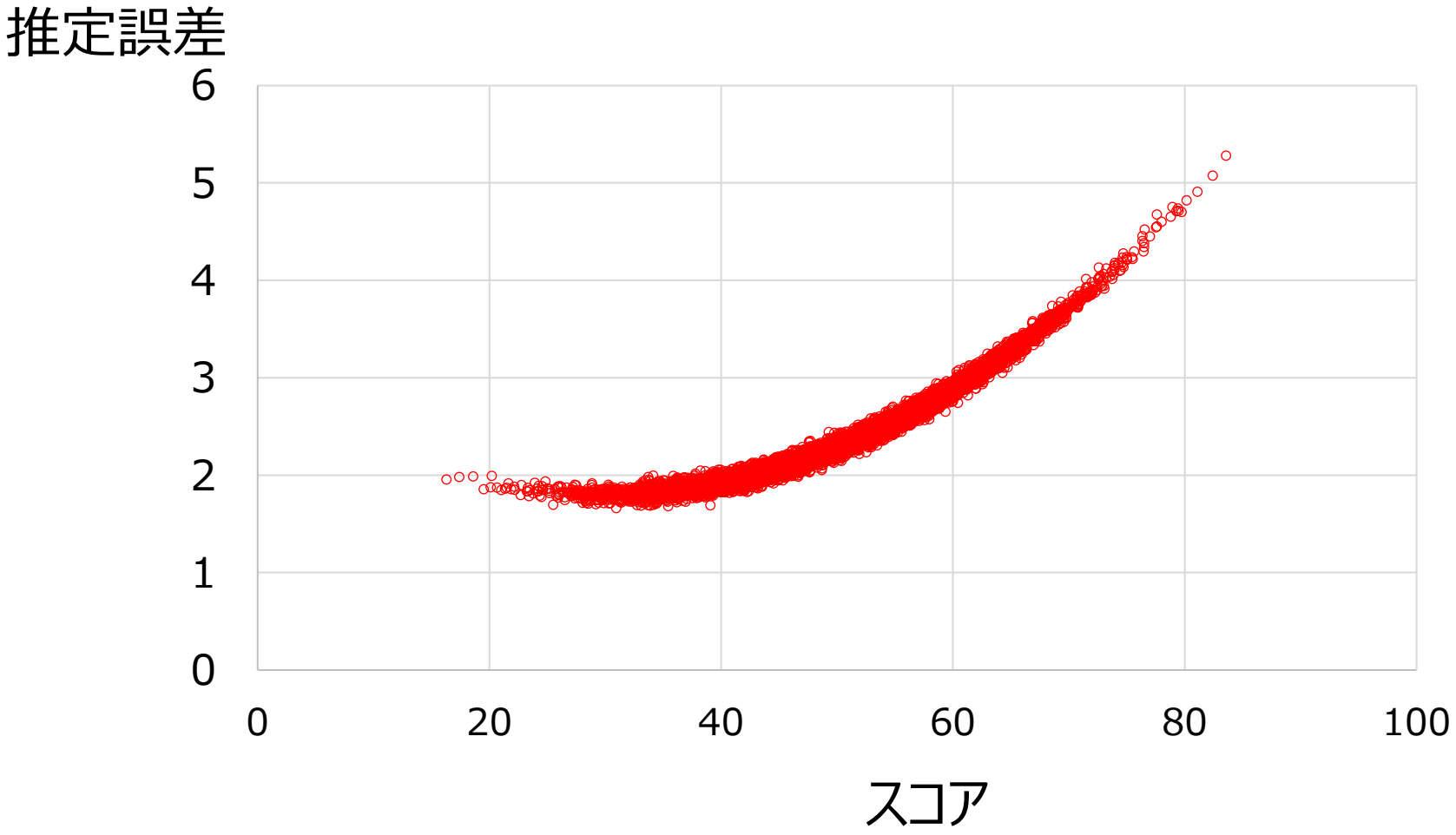


最大クリーク問題に帰着（10万以上のテスト生成）2018年度電子情報通信学会論文賞受賞



1. 石井隆稔・赤倉貴子・植野真臣“複数等質テスト構成における整数計画問題を用いた最大クリーク探索の近似法”, 電子情報通信学会論文誌D Vol.J100-D(1), pp.47-59, 2017.
2. 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
3. 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.

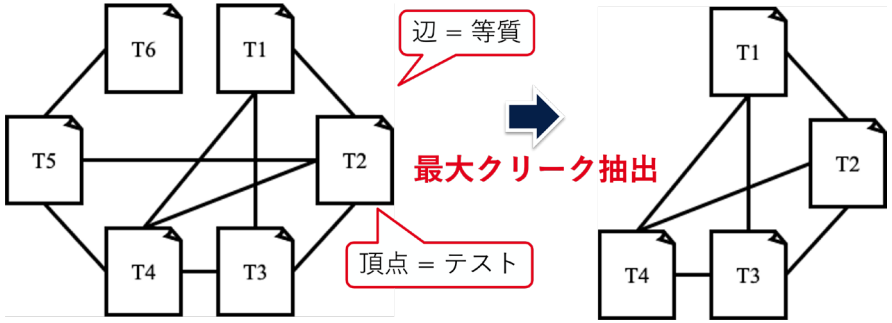
あるハイステークスな公的試験における等質テストのスコアの測定誤差



M.Ueno,K.Fuchimoto, and E,Tsutsumi : E-testing from artificial intelligence approach. Behaviormetrika, Vol. 48, No. 2, pp. 409–424, (2021)

テスト生成数をより向上させるアルゴリズムの開発

1. Fuchimoto, Kazuma, Shin-ichi Minato, and Maomi Ueno. "Automated Parallel Test Forms Assembly using Zero-suppressed Binary Decision Diagrams." IEEE Access (2023).
2. Kazuma Fuchimoto, Takatoshi Ishii, and Maomi Ueno: Hybrid Maximum Clique Algorithm Using Parallel Integer Programming for Uniform Test Assembly,IEEE Transactions on Learning Technologies, vol. 15, no. 2, pp. 252-264, 1 (2022)
3. 渕本　壱真，湊　真一，植野　真臣．Zero-suppressed Binary Decision Diagrams を用いた自動テスト構成，人工知能学会論文誌，37.5 A-M23_1，(2022).
4. 渕本 壱真，植野 真臣．等質テスト構成における整数計画法を用いた最大クリーク探索の並列化．電子情報通信学会論文誌D，Vol.J103-D，No.12，pp. 881-893，(2020).

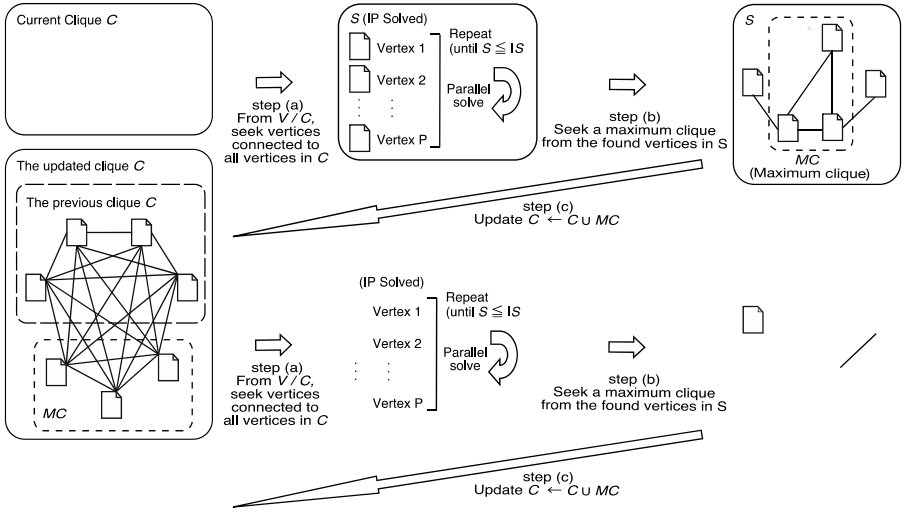


第一段階

時間計算量が小さいが

空間計算量が大きい最大クリーク法により

メモリの限界までテストを生成



第二段階

時間計算量が大きいが

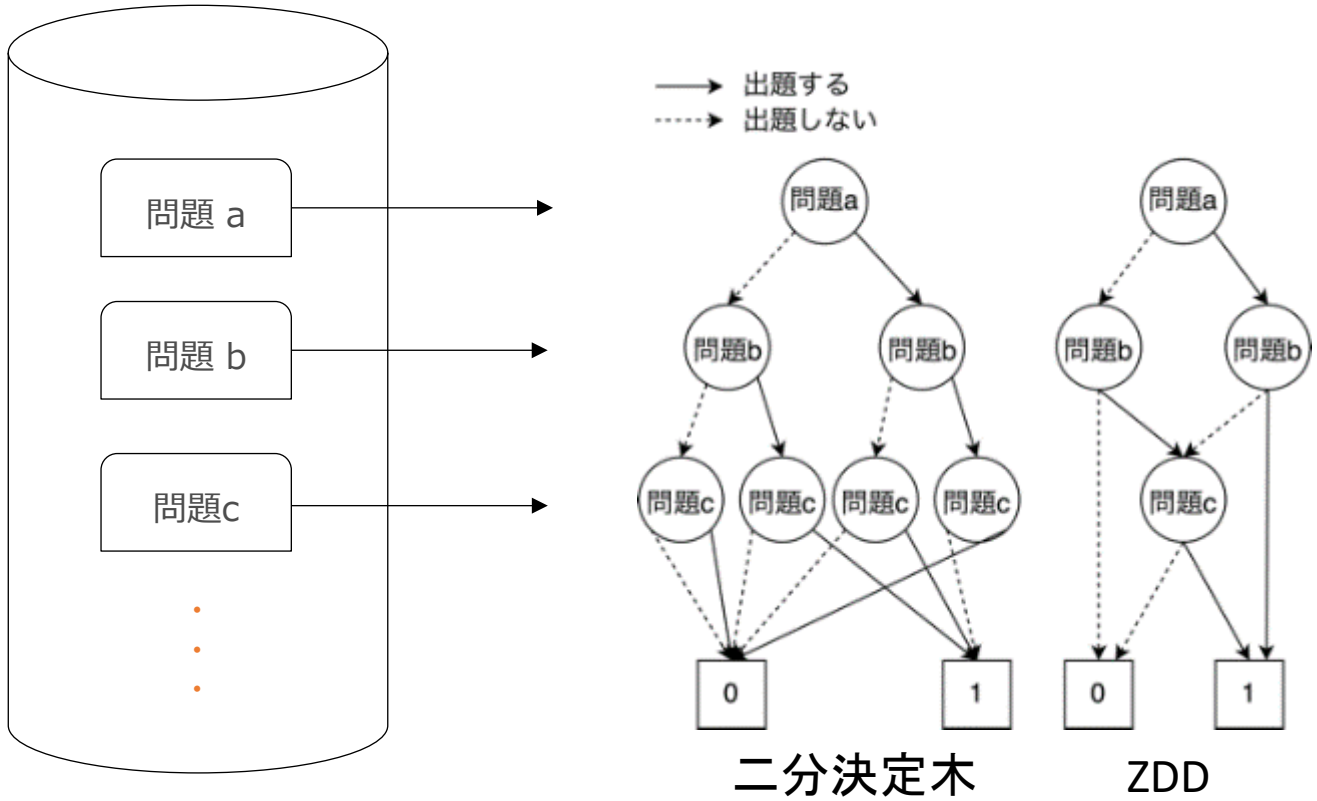
空間計算量が小さい整数計画法を用いた

並列探索アルゴリズムにより

テストを逐次的に生成

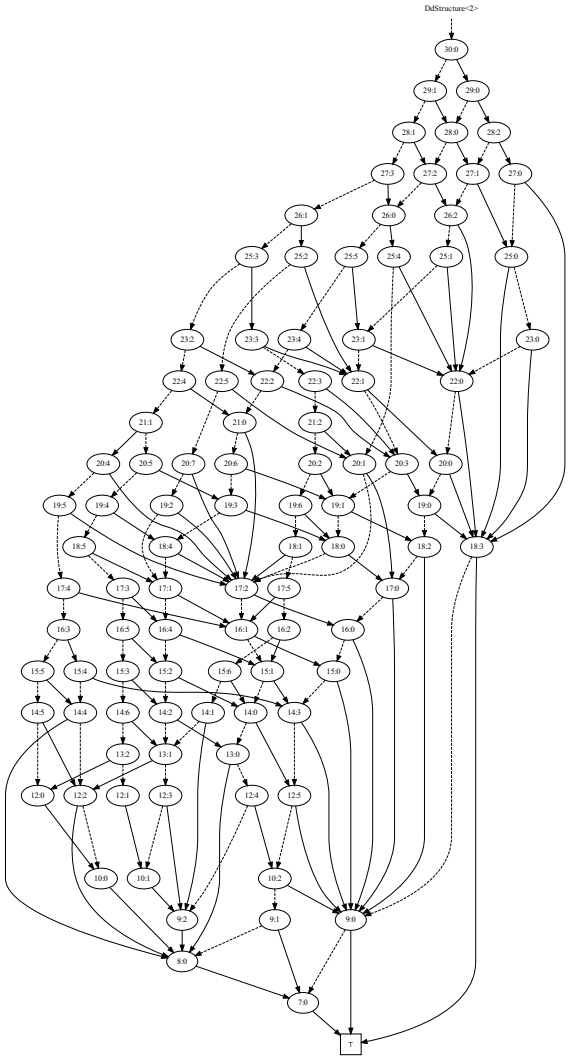
→ 1ヶ月で約45万のテスト生成に成功（従来の約4.5倍）

Fuchimoto, K., Ishii, T., and Ueno, M.: Hybrid Maximum Clique Algorithm Using Parallel Integer Programming for Uniform Test Assembly, IEEE Transactions on Learning Technologies (2022)



各テストの測定誤差が等質となるように
問題の組合せをZDD[Minato 93]を用いて最適化

Minato, Shin-ichi. Zero-suppressed BDDs for set manipulation in combinatorial problems. Proceedings of the 30th International Design Automation Conference. 1993.



ZDDを用いた並行テスト構成例

- 24時間で約150万のテスト生成に成功

Fuchimoto, Kazuma, Shin-ichi Minato, and Maomi Ueno. "Automated Parallel Test Forms Assembly using Zero-suppressed Binary Decision Diagrams." IEEE Access (2023).

項目露出を一様とする等質テスト自動構成アルゴリズム

1.渕本 壱真，植野 真臣．項目露出ペナルティを用いた整数計画法による自動並行テスト構成，統計数理 (2024)

2.植野 晶，渕本 壱真，植野 真臣．項目露出を考慮した整数計画法による等質テスト構成，電子情報通信学会論文誌D，(2022).

- 最大クリークやZDDを用いた手法ではテスト間に
  問題項目の重複を一定数許すことでテスト数を増加
- 各問題項目の出題頻度（露出数）に偏りが生じる

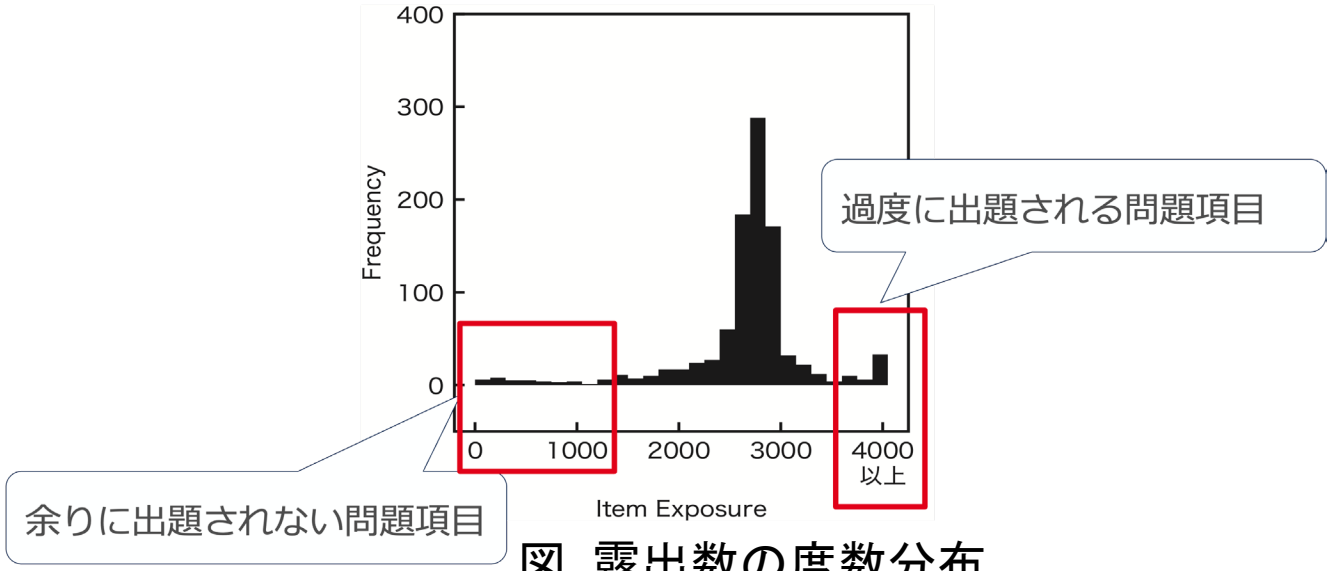


図. 露出数の度数分布

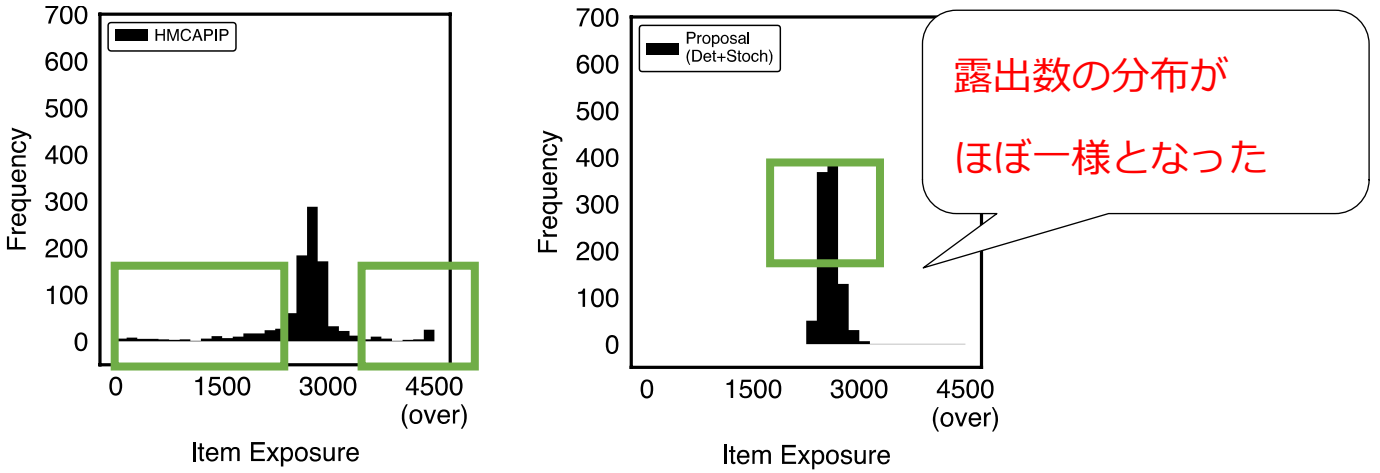
<u>最も深刻な問題</u>

- 問題流出により受験対策され、
  問題項目やテストの信頼性が低下[Wainer 2000]

Wainer, H (2000). CATS: Whither and Whence, Educational Testing Service

$$\textbf{Maximize}\ \sum_{i=1}^{n}\left(\lambda_i\, x_i - \underbrace{\frac{1}{1+e^{-z_i}}}_{①} - \underbrace{\mathrm{M}_{\{i,p\}}}_{②}\right) x_i$$

- 露出数の大きさに応じたロジスティック関数による
  ①決定論的ペナルティおよび②確率的ペナルティを与えた
  整数計画法を用いて逐次的にテスト生成



Fuchimoto et al. [2022]　　提案手法

渕本 壱真, 植野 真臣. 項目露出ペナルティを用いた整数計画法による自動並行テスト構成, 統計数理 （印刷中）

eテスティングの実用化

- 異なる問題から構成されるが、同一受検者が受検すると同一得点を返す複数の異なるテストの構成が実現される。
- これらはeテスティングと呼ばれ、世界標準になり、日本最大の国家試験：情報処理技術者試験、医学部共用試験、などに導入されている。

まとめ

・信頼性とは　同一能力の受験者が何度テストを受けても同一の結果を返せる性能

・信頼性の高いテストを生成するには、各テストの誤差を予測し、その誤差がすべてのテストで等しくなり、さらになるべくその誤差を小さくするように構成（アイテムバンクより抽出）する

・各テストの誤差を予測するために項目反応モデル（Item Response Theory;IRT)であらかじめ項目データベースの項目の特性値を求めておく必要がある