多肢強制選択型心理測定の
回答負荷を軽減するための項目提示法

分寺 杏介

神戸大学　経営学研究科

✉ bunji@bear.kobe-u.ac.jp

杉山 剛

株式会社リクルートマネジメントソリューションズ

日本行動計量学会第52回大会

2024年9月12日

心理学でよく使われる質問の形式

リッカート尺度

一つ一つの項目について「程度」を答える

e.g., 自分に当てはまる程度，賛成する程度

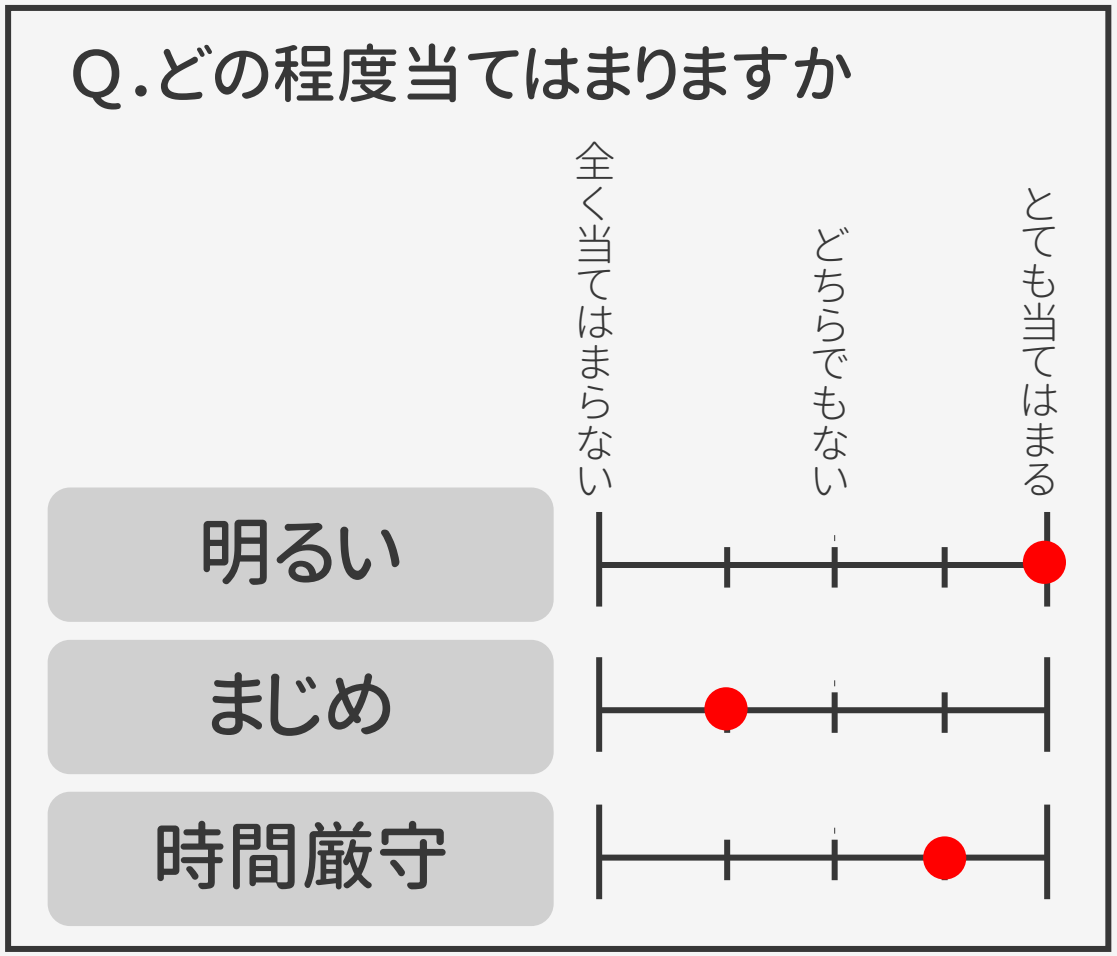
系統的バイアスの影響を受けやすい

中心化傾向…5段階なら3を選びやすい

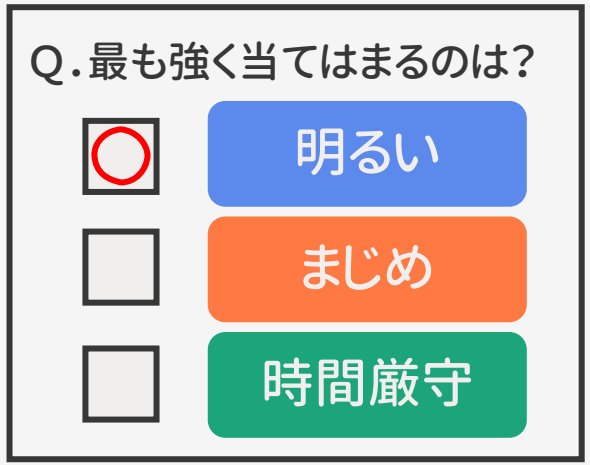
黙従傾向（寛大化傾向）…内容とは無関係に「あてはまる」を選びやすい

フェイキング…自分がよく見えるように意図的に回答を変える　など

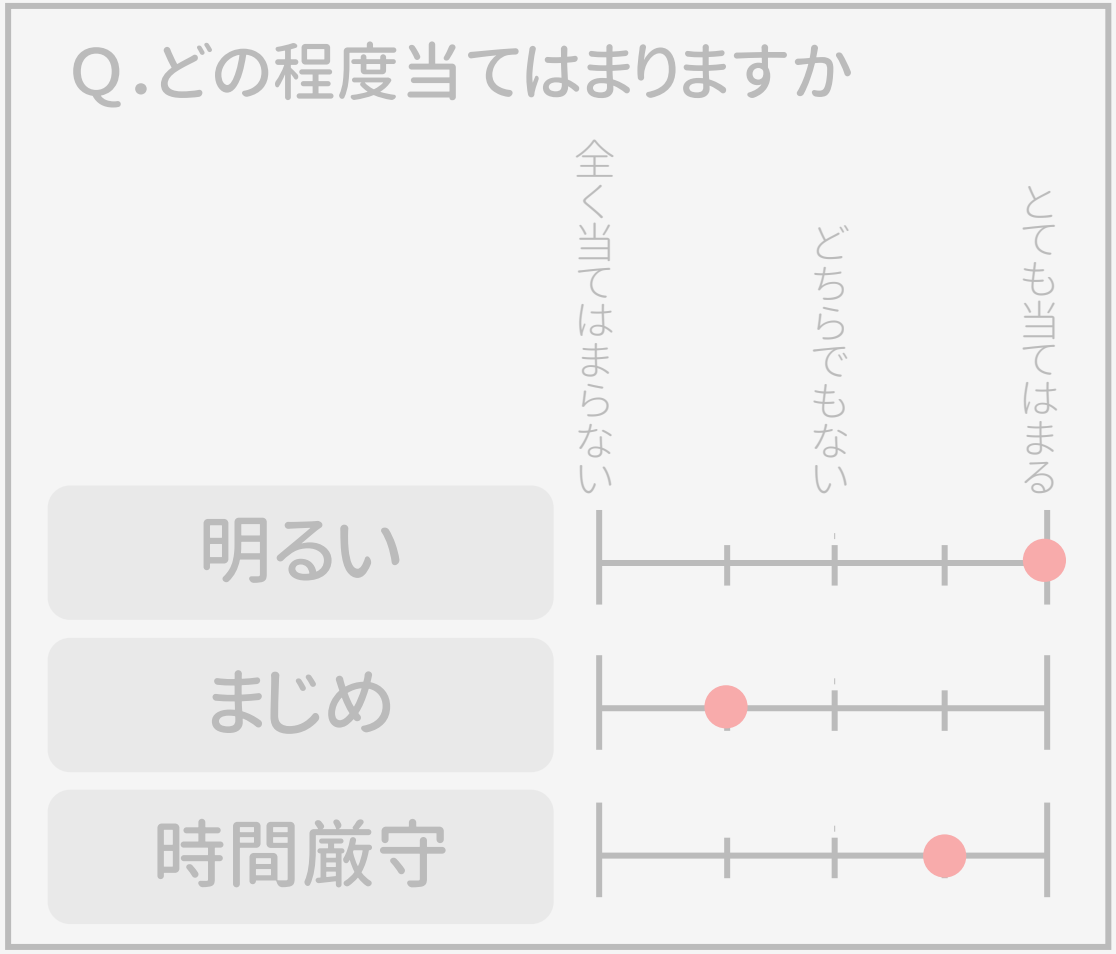
Single-Stimulus (SS); Likert scale



バイアスが発生しにくい質問方法



Single-Stimulus (SS); Likert scale



（多肢）強制選択式(Forced-Choice; FC)

複数の選択肢の中から回答する形式

- 最も当てはまる／らないもの
- 当てはまる順にランキング

▶ 中心化傾向などは起こり得ない
フェイキングもかなり抑えられる

FCのほうが妥当性が高いという先行研究あり (e.g., Christiansen et al., 2005 Salgado & Táuriz, 2014)

多肢強制選択型測定の問題点

答えるのが大変

例：OPQ32
（職務行動に関するアセスメントツールとして
最も有名なものの一つ）

	Most	Least
I like to discuss abstract concepts	●	○
I enjoy interpreting statistics	●	○
I feel that people are honest	●	○

ざっくり日本語訳

	Most	Least
抽象的な概念について話し合うのが好きだ	●	○
統計データを解釈するのは楽しい	●	○
人はみな正直であると思う	●	○

普通のリッカート尺度だとしたら

抽象的な概念について話し合うのが好きだ

Tourangeau et al.(2000)
のフレームワーク

- 理解 → まずは文章の意味を理解して
- 想起 → 過去の記憶などを思い出したりして
- 判断 → どの程度当てはまるかを考えて
- 回答 → 選択肢に落とし込む

「ややあてはまる」
くらいかなぁ

多肢強制選択型測定の問題点

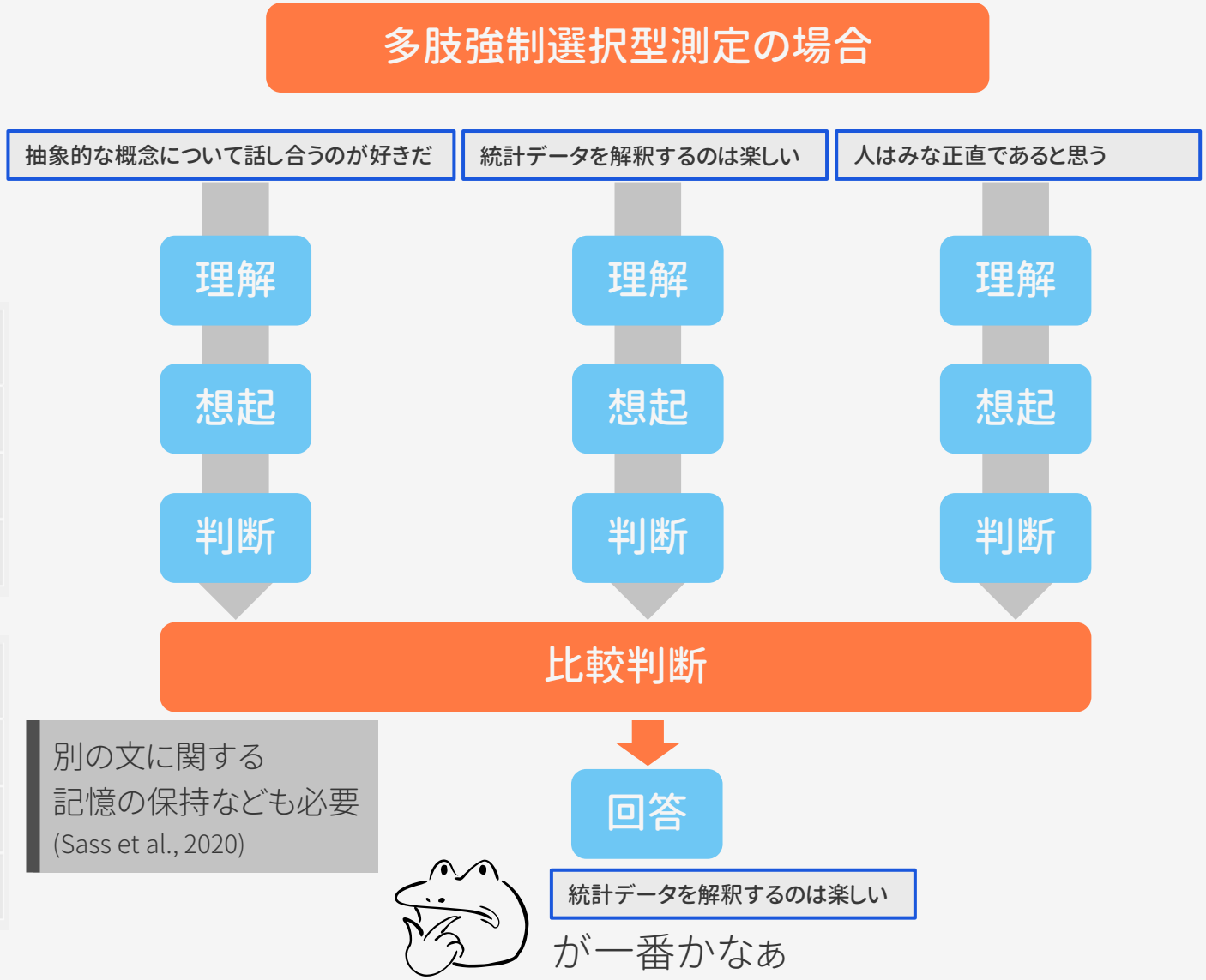
答えるのが大変

例：OPQ32
（職務行動に関するアセスメントツールとして
最も有名なものの一つ）

	Most	Least
I like to discuss abstract concepts	○	○
I enjoy interpreting statistics	○	○
I feel that people are honest	○	○

ざっくり日本語訳

	Most	Least
抽象的な概念について話し合うのが好きだ	○	○
統計データを解釈するのは楽しい	○	○
人はみな正直であると思う	○	○



研究の目的

回答に必要な認知負荷を軽減できそうな項目提示法を提案

一言でいうと「同じ文を使い回す」

情報量の低下をカバーするためのデータ拡張法を提案

「同じ文を使い回す」ために情報量が低下する

▶「推移律を満たす」と仮定したデータ拡張を行う

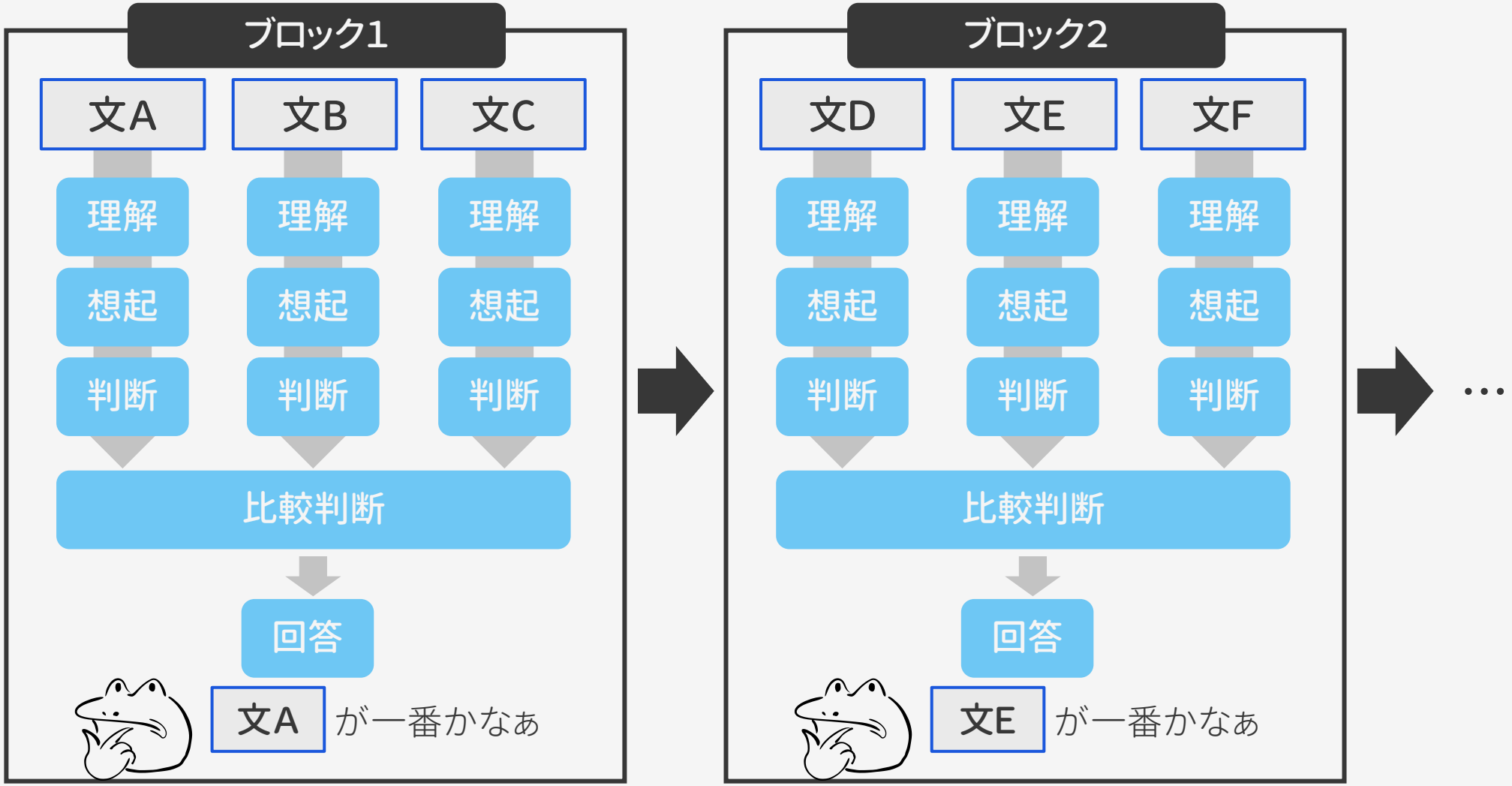
提案手法の性能を検証する

回答負荷が軽減されているかを確認（今回は回答時間のみ）

データ拡張法による推定精度の変化を確認（シミュレーション）

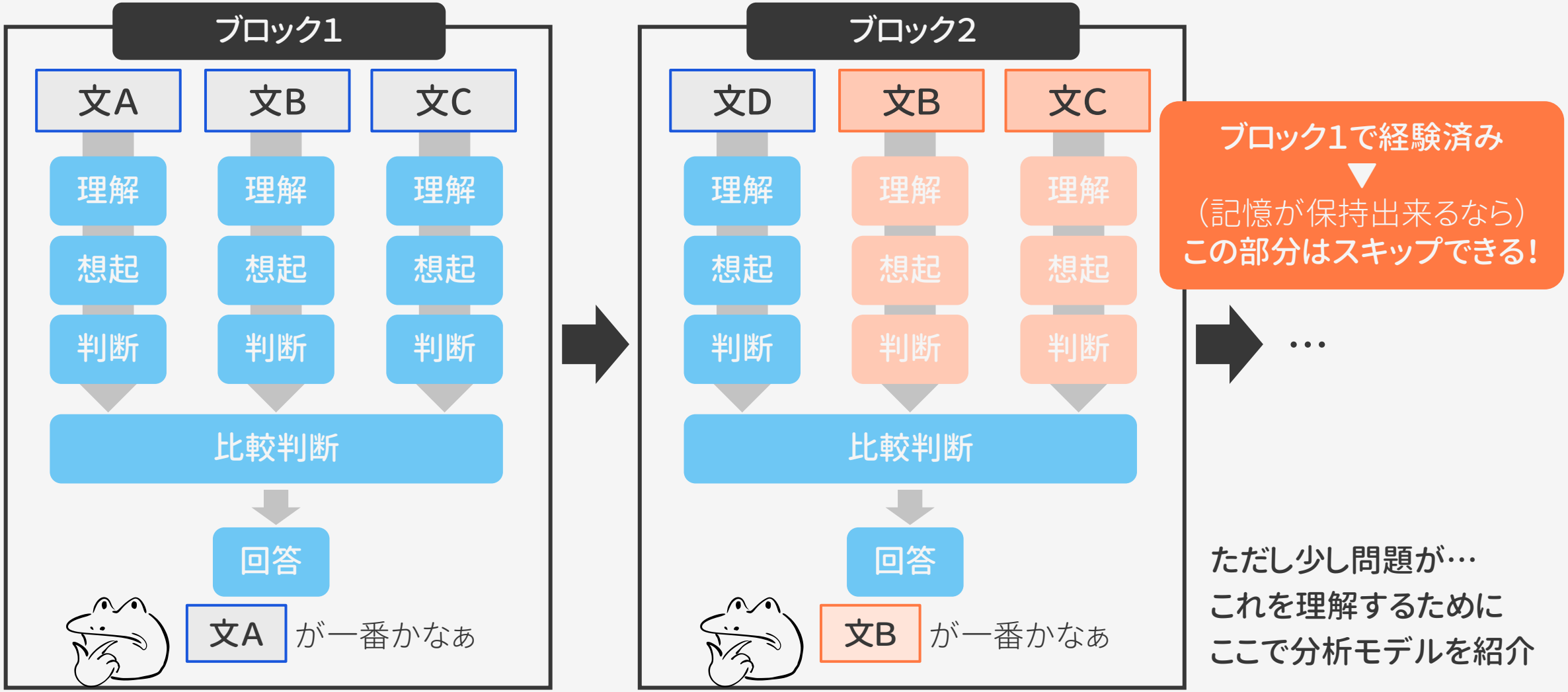
通常の多肢強制選択型

問題は，ひとつひとつの回答に対するプロセスが重すぎること



少しでも負荷を軽減するために

毎回全てを新規の文にするのではなく，一部は使い回せば良いのでは？



多肢強制選択型に対する回答を分析するモデル

Thurstonian IRT（TIRT）モデル (Brown and Maydeu-Olivares, 2011)

【多肢強制選択】

Q．最も強く当てはまるのは？

- ○ 明るい
- □ まじめ
- □ 時間厳守

μ: 選択肢の選好の平均

β : 因子負荷

η : 因子得点（特性値）

■ 2つの文 (j_1, j_2)はそれぞれ異なる因子(f_{j_1}, f_{j_2}) を測定する

$$X_{ij}^* = u_{ij_2} - u_{ij_1} \qquad X_{ij} = 1 \text{ if } X_{ij}^* \geq 0$$

■ 一方の潜在変数u_{ij_2}について見ると

$$u_{ij_2} = \mu_{j_2} + \beta_{j_2}\eta_{if_{j_2}} + \varepsilon_{ij_2} \qquad \varepsilon_{ij_2} \sim N(0, \Psi_{j_2}^2)$$

■ (j_1, j_2)間の比較において選択肢 j_2 が選ばれる確率は

正規累積モデル

$$P(X_{ij} = 1|\boldsymbol{\eta}_i) = \Phi\left[\left(\mu_{j_2} - \mu_{j_1}\right) + \left(\beta_{j_2}\eta_{if_{j_2}} - \beta_{j_1}\eta_{if_{j_1}}\right)\right]$$

$\Phi(\cdot)$ ▶ 標準正規分布の累積分布関数

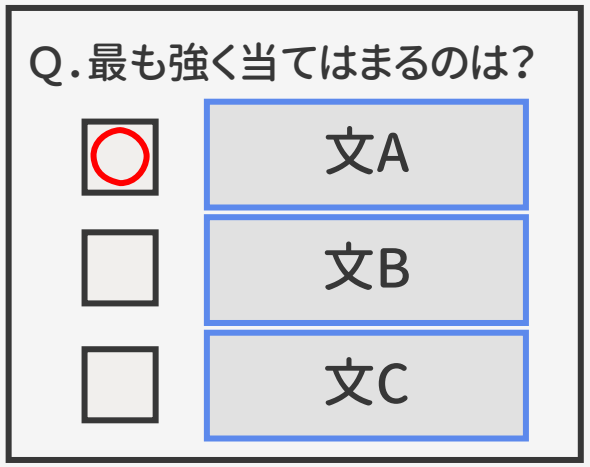
または

ロジスティックモデル

$$P(X_{ij} = 1|\boldsymbol{\eta}_i) = \left(1 - \exp\left[\left(\mu_{j_2} - \mu_{j_1}\right) + \left(\beta_{j_2}\eta_{if_{j_2}} - \beta_{j_1}\eta_{if_{j_1}}\right)\right]\right)^{-1}$$

TIRTモデルで分析する際のポイント

多肢選択の回答を仮想的な2肢選択に分解する



A—B （=文Aと文Bの比較）

$$P(\mathrm{A} > \mathrm{B}|\boldsymbol{\eta}_i) = \Phi\left[\left(\mu_{j_B} - \mu_{j_A}\right) + \left(\beta_{j_B}\eta_{if_{j_B}} - \beta_{j_A}\eta_{if_{j_A}}\right)\right]$$

A—C （=文Aと文Cの比較）

$$P(\mathrm{A} > \mathrm{C}|\boldsymbol{\eta}_i) = \Phi\left[\left(\mu_{j_C} - \mu_{j_A}\right) + \left(\beta_{j_C}\eta_{if_{j_C}} - \beta_{j_A}\eta_{if_{j_A}}\right)\right]$$

B—C （=文Bと文Cの比較）

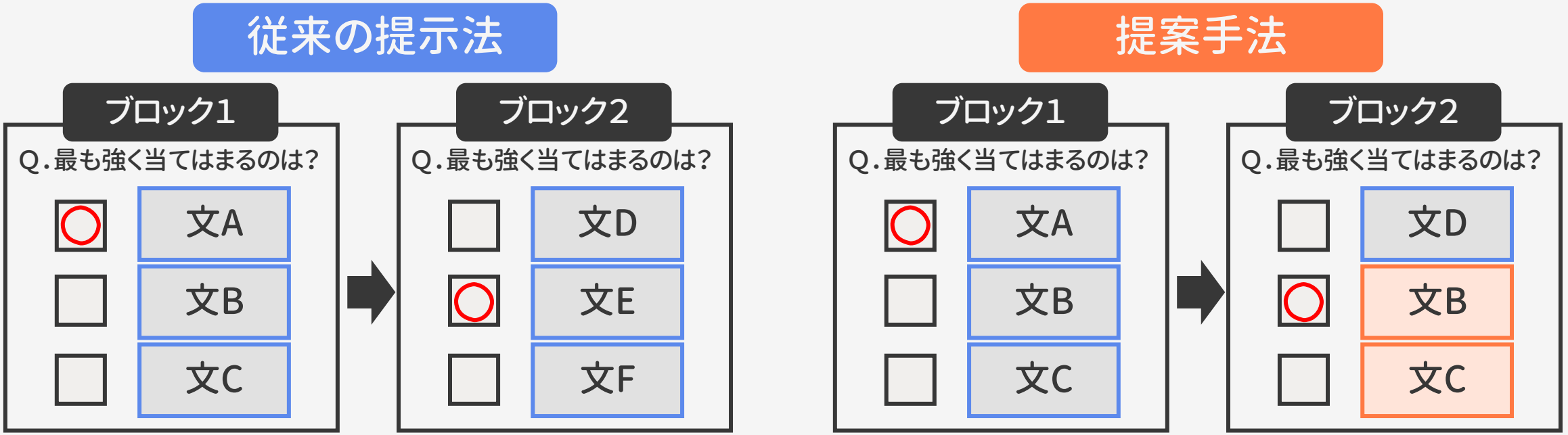
$$P(\mathrm{B} > \mathrm{C}|\boldsymbol{\eta}_i) = \Phi\left[\left(\mu_{j_C} - \mu_{j_B}\right) + \left(\beta_{j_C}\eta_{if_{j_C}} - \beta_{j_B}\eta_{if_{j_B}}\right)\right]$$

▲ 回答からはわからないので欠測として扱う

提案手法の特徴

提示される文の「種類」数が減ってしまう

選択肢数，回答数をそれぞれ K, J とすると



$K \times J$種類の文が提示される

$K + J - 1$種類の文が提示される

【例】3択で回答数20の場合　60種類　22種類

その特徴がもたらす影響は？

良い（かもしれない）影響

用意する文の数が少なくて済む

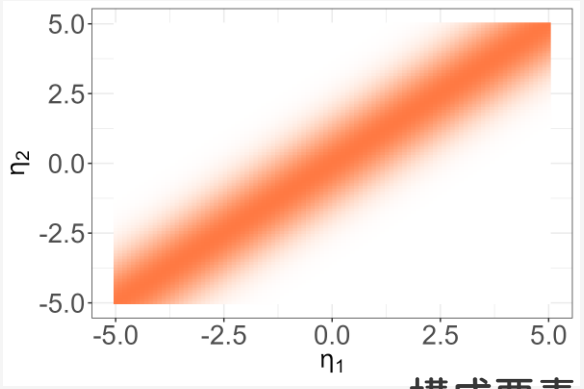
推定する項目パラメータの数が少なくて済む

悪い（かもしれない）影響

個人特性値の推定に与える情報量が安定しにくくなる可能性

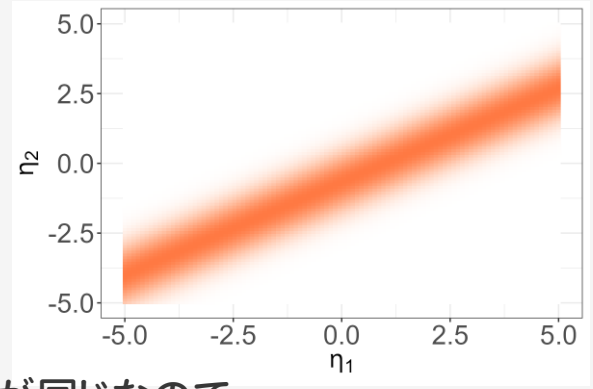
もちろん構成概念の代表性といった
質的な問題もあるのですが

文A 文B の1ペアがもつ
項目情報関数



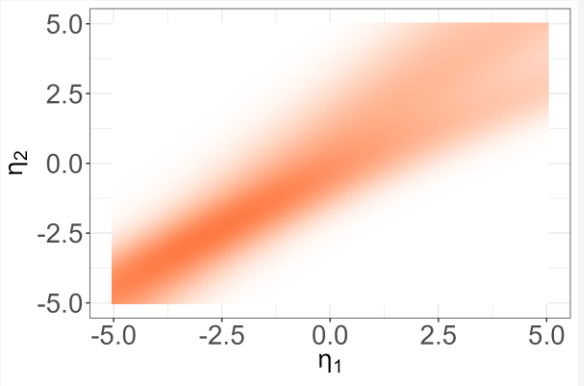
＋

文A 文E の1ペアがもつ
項目情報関数



＝

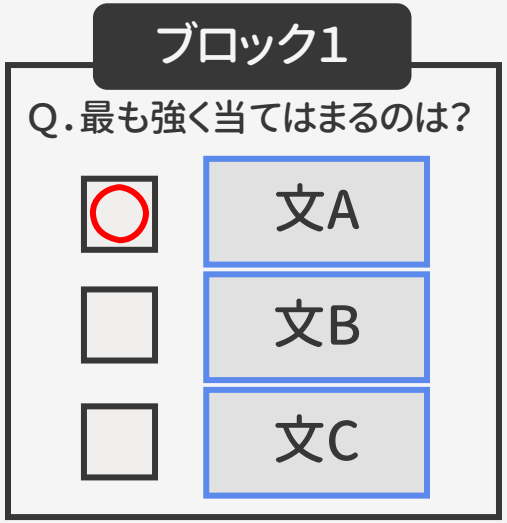
2つのペアの情報関数の和

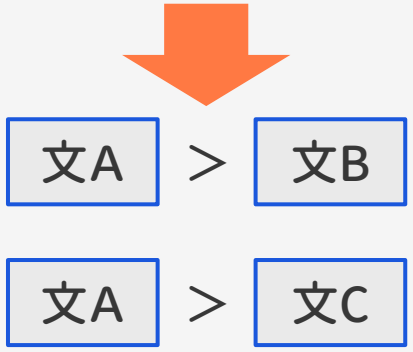


構成要素の半分が同じなので
似たような項目情報量になりやすい

情報量を少しでもカバーするために

回答データを拡張することを考える

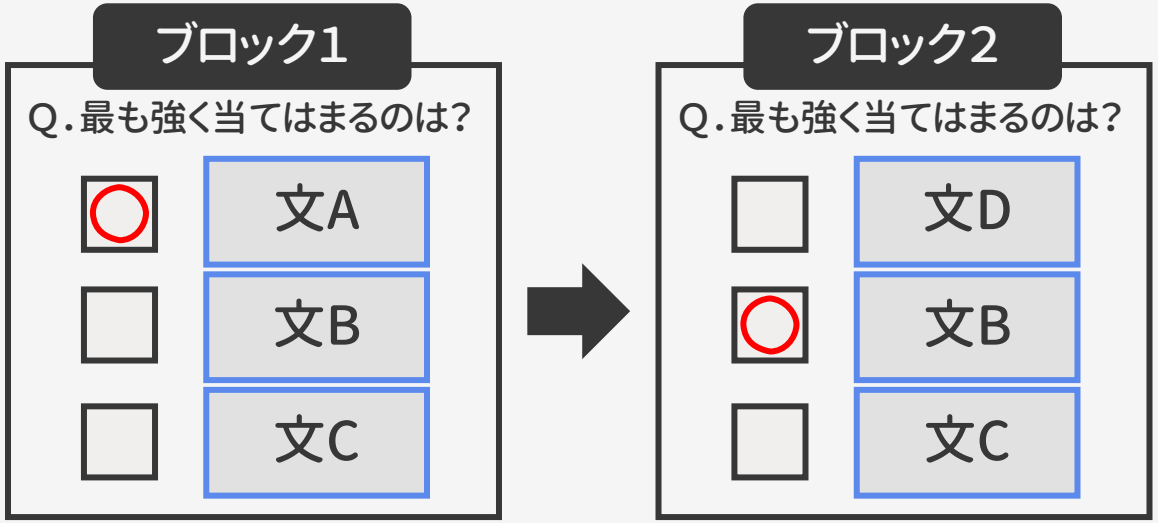




であるため

情報量を少しでもカバーするために

回答データを拡張することを考える



文A ＞ 文B

文A ＞ 文C

であるため

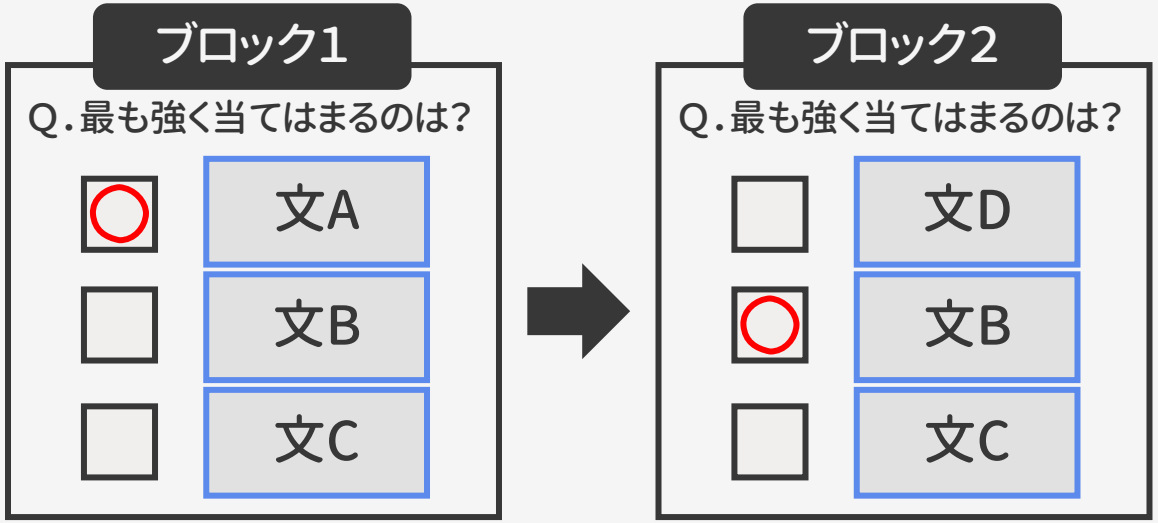
文B ＞ 文D

文B ＞ 文C

であるため

情報量を少しでもカバーするために

回答データを拡張することを考える



文A ＞ 文B

文A ＞ 文C

であるため

文B ＞ 文D

文B ＞ 文C

であるため

データ拡張ルール

推移律が満たされるならば

文A ＞ 文B ＞ 文D

仮に同じブロックに含まれた場合

文A ＞ 文D

となる可能性が高い!

方法

提案手法・データ拡張の効果を確認するため

【文】本調査用に作成した272文

株式会社リクルートマネジメントソリューションズが保有する性格検査をもとに作成

▶ ただし具体的な文は著作権の都合上公開できません

【因子】272文をリッカート式で尋ねたデータを収集（577名）

▶ 探索的因子分析によって9因子構造を設定

【標本】Web調査で1260名から回答を得た

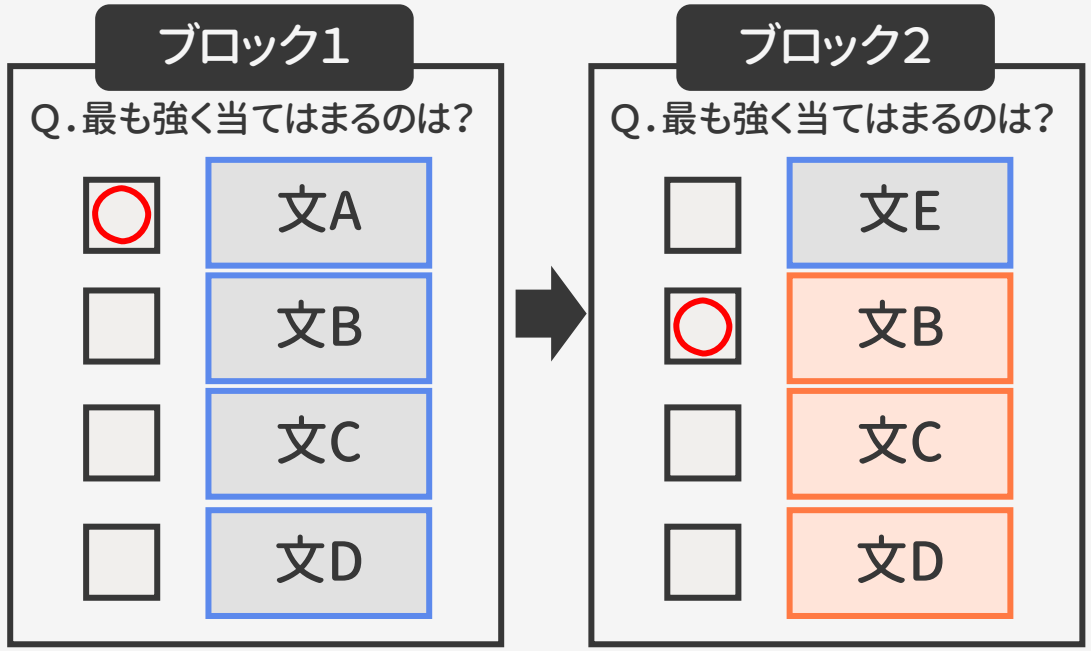
▶ 不適切回答と思われるものを除外して，**最終的に816名のデータを使用**

回答時間，ストレートラインの長さ，マハラノビス距離などにより判断

項目提示法

パターンの実施順はランダムに決定

パターン1：提案手法

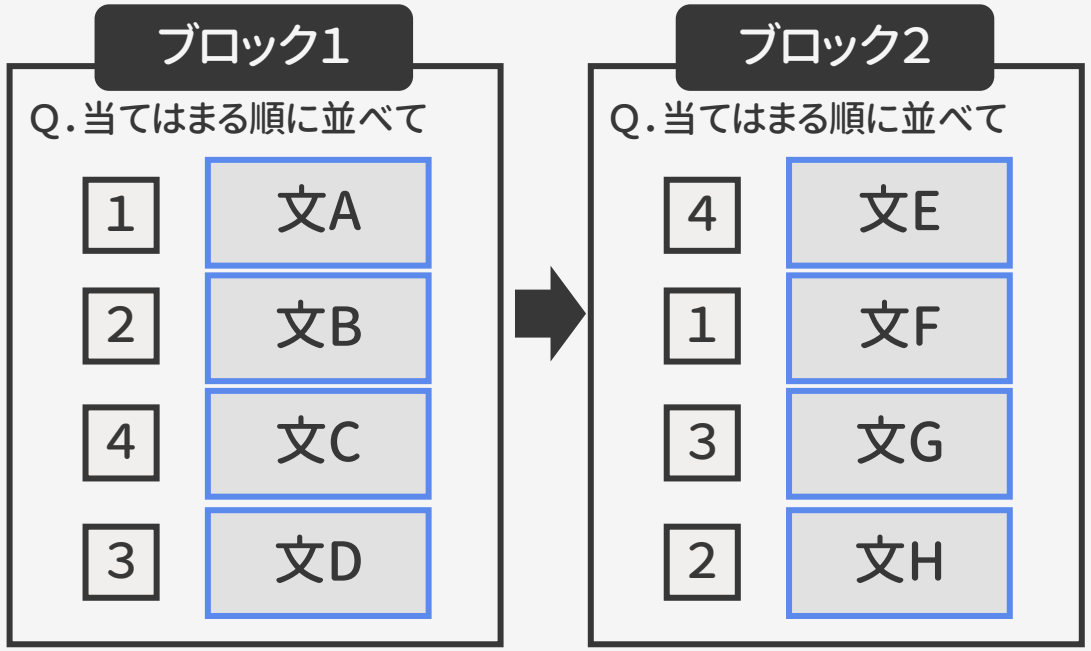


本番では提示できる文がなくなるまで実施

▶ 推定時には最初の70回答のみ使用

後半の適当回答が多すぎたため

パターン2：従来手法



8ブロックを事前に用意

その他

データサイズ（パターン1）

拡張前：168330ペア ▶ 拡張後561720ペア

MCMC (cmdstan) によるベイズ推定

データの設計などの理由で?どうしても一部のηが収束しなかった

事前分布

$$|\beta_j| \sim \text{normal}(1,3), \quad \mu_j \sim \text{normal}(0,3), \quad \Psi_j \sim \text{normal}(1,0.3),$$
$$\nu_j \sim \text{normal}(0,\Psi_j), \quad \boldsymbol{\eta} \sim \text{multi\_normal}(0,\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \sim \text{lkj\_corr}(1),$$

結果｜回答負荷

回答時間ベースで見てみる ※パターンごとに回答方法が異なる点に注意

パターン1

1つ選択するまでの時間が記録されている

Q.最も強く当てはまるのは?

○	文A
□	文B
□	文C
□	文D

パターン2

3つ目を選ぶまでの時間が記録されている

Q.当てはまる順に並べて

1	文A
2	文B
4	文C
3	文D

比較できるのは
「1位を選ぶまでの時間」

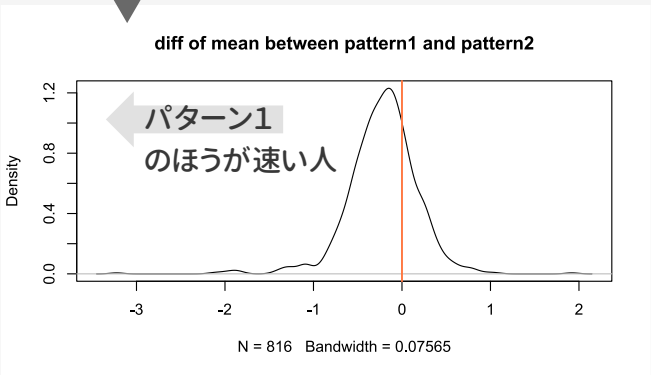
↓

今回はシンプルに
「3で割った値」を使用
※個人ごとの平均値・中央値

結果

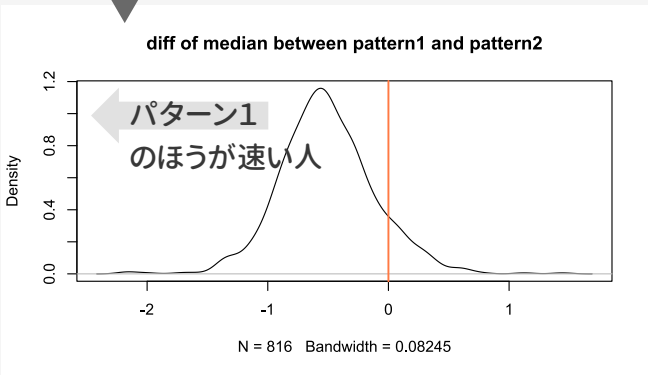
log(平均値)
パターン1-パターン2

パターン1：2876ms
パターン2：3580ms



log(中央値)
パターン1-パターン2

パターン1：1901ms
パターン2：3172ms

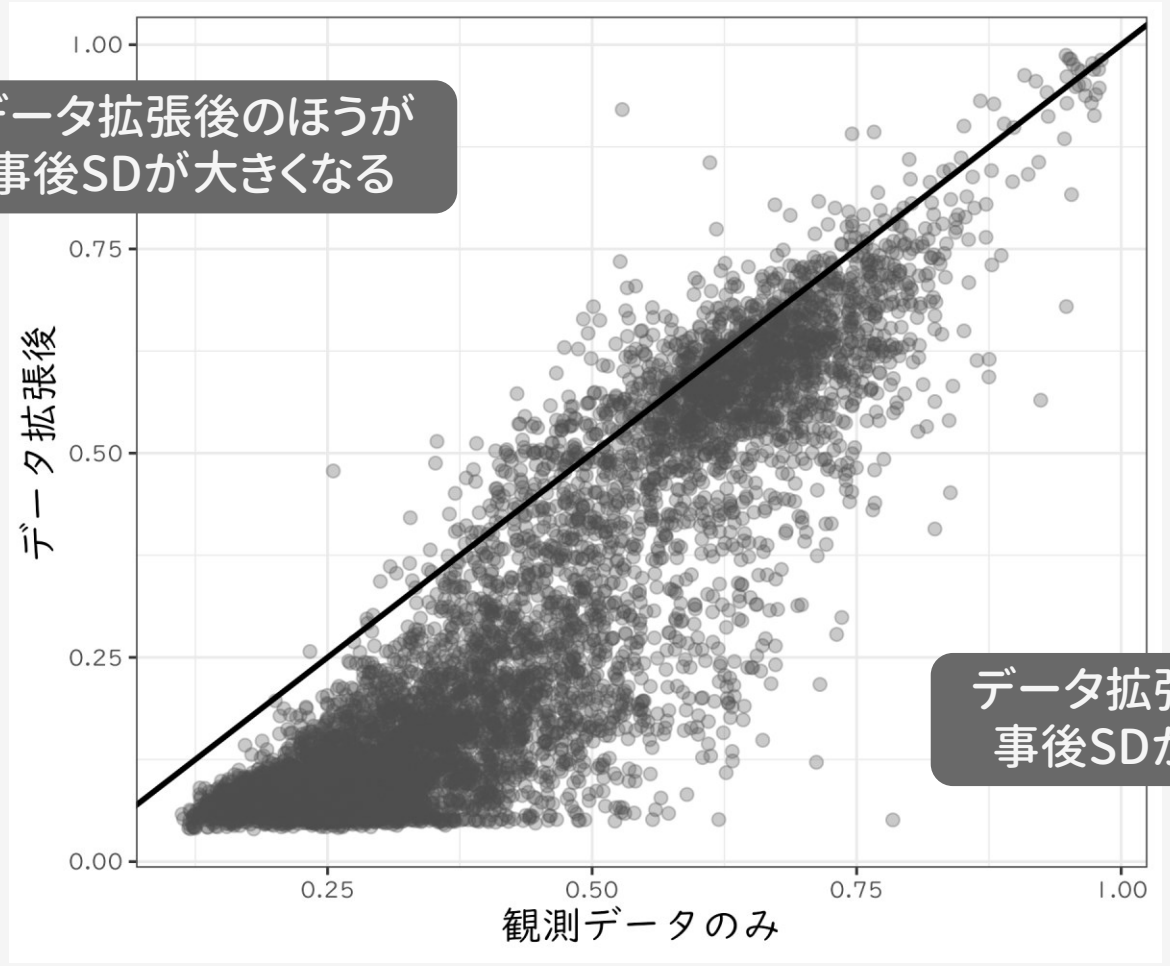


（人によるが）
平均的にはパターン1のほうが回答時間は短くなっていた

結果｜事後分布の標準偏差

事後SDの散布図

$y = x$



抄録に掲載したプロットは
収束していないものを含んでいました

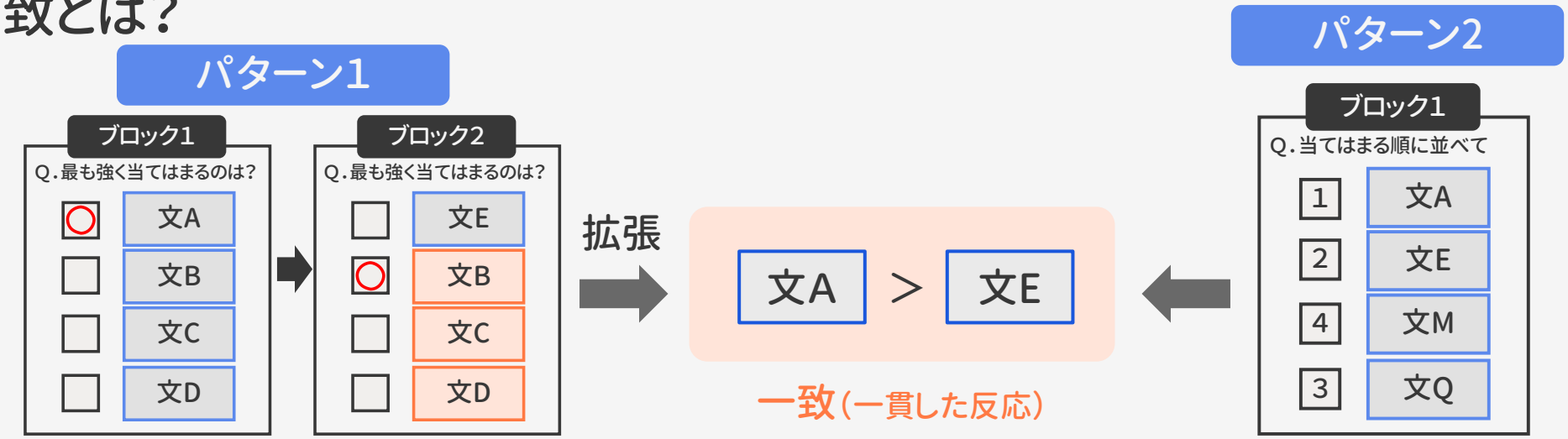
▼

どう頑張っても収束しなかったので
「$\hat{R} < 1.1$のηのみ」を表示しています

全体として事後SDは小さくなることが多いが
一部では拡張後のほうが大きくなることも

結果｜回答の一致率

一致とは？



結果

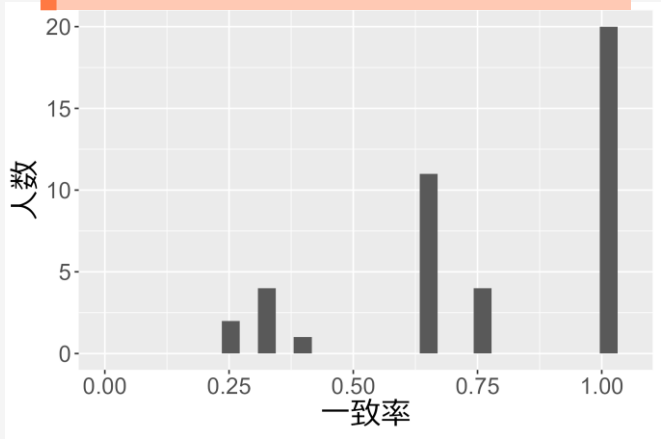
比較可能なペアの総数

	N	一致率	p
観測データのみ	224	**80.8**%	.417
データ拡張のみ	597	**77.9**%	

チャンスレベルは50%

（当然ながら）一致率は
人によって大きく異なる

3件以上比較可能であった回答者



考察

1回答あたりの回答負荷は多少軽減されているはず

▲ 回答時間が短くなっていることから判断

データ拡張によって推定の安定性（事後SD）も向上している

▶ データを水増ししているので当然といえば当然

推定精度については場合による

どれだけ**推移律に沿った（つまり全体として整合的な）回答をしているか**による

- 全体としては「実際に同じ比較を2回行った場合」に近い割合で一致
- ただし人によって大きく異なる一致率

今後の展望

データ拡張によって推定の誤差はどうなるか

回答の一貫性などを操作したシミュレーションでの検証

出題アルゴリズムの改良

「負け残り」なので，どうしてもストレートライン的回答が起こりやすくなる

▶ 提示順に工夫を加えたり，一定のストレートラインが見られたら全入れ替えなど

よりフェアな形で比較できるようにデータを収集する

【推定精度】回答時間が同じになる回答回数で比較する

【回答負荷】回答回数あるいは「仮想的な二値」の数を揃えて尋ねる？

あるいは情報量が同じになるときの回答負荷を問えるか？