【セミナー】

Computer Based Testingの過去・現在・未来

Past, Present, and Future of Computer Based Testing

分寺 杏介

神戸大学　経営学研究科

✉ bunji@bear.kobe-u.ac.jp

日本テスト学会 第17回学会賞記念講演会

2024年3月16日

本講演内容の大半は以下の論文に基づいています。

分寺 (2023) コンピュータを用いたアセスメントに関する研究トピックの整理と最新の動向.
日本テスト学会誌, *19*, 191-225. https://doi.org/10.24690/jart.19.1_191

展望論文

コンピュータを用いたアセスメントに関する研究トピックの整理と最新の動向

分寺杏介 [1]

[1] 神戸大学
[1] ベネッセ教育総合研究所

要約

本稿では，コンピュータを用いたアセスメント（computer-based testing［CBT］）の理論的側面に関する各領域の研究を概観するとともに，最新の研究動向を紹介する。CBTに関する主要な研究トピックのうち「紙筆式（PBT）による得点との比較可能性」「適応型テスト」「新しい形式の項目」「オンライン試験における不正行為とその対策」「ログデータの活用」「特別な配慮」の 6 点について，これまでの知見および最新の動向を紹介した。また，CBTの発展に関する先行研究の予測に従い，CBT に関する研究の今後の方向性についての展望を「妥当性」「テスト不安」「自動化」という 3 つの観点から論じた。

キーワード： computer-based testing，モード効果，適応型テスト，technology enhanced items，不正行為，ログデータ

ちなみに本セミナーのタイトルは
もともとこの論文につけようか迷って
結局つけなかったものです。

※この論文で引用していない
文献のみリンクを貼っています。
リンクのない文献はこの論文
をご参照ください。

Outline

1 **イントロダクション**

CBTとは何者か

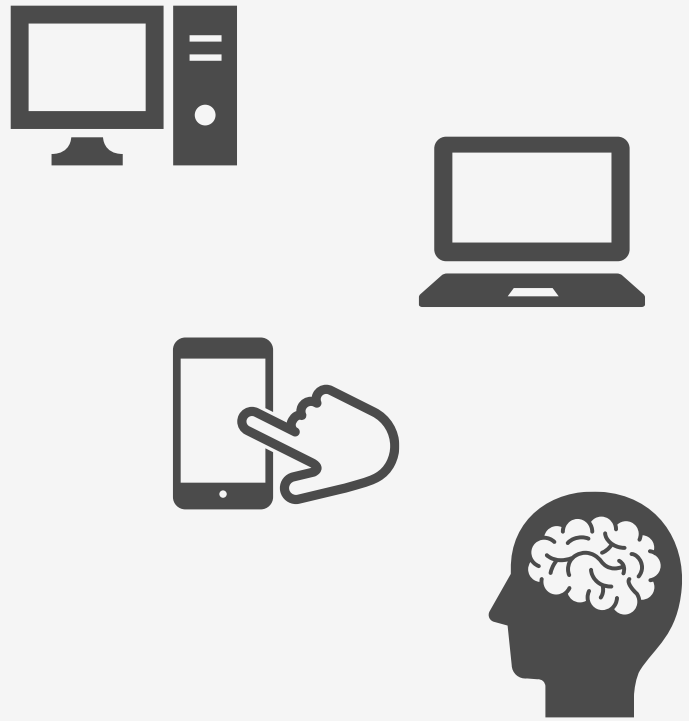
CBTの歴史を概観する

2 **CBT研究の過去・現在**

主要な研究トピックを概観する

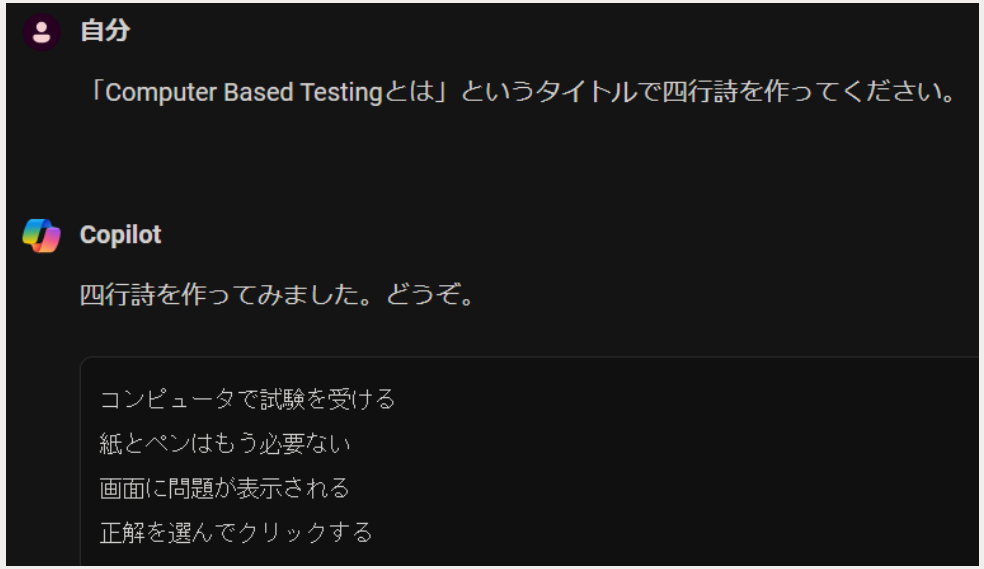
3 **CBT研究の未来**

テストを「より良い」ものにするために



目次にもどる

1
イントロダクション



- CBTとは何か
- CBTの歴史を概観

Computer Based Testing

その名のとおり「コンピュータ」上で実施されるテスト全般を指す

テストはコンピュータを用いた形式へ移行が進んでいる

大規模テストではNAEP, PISA, PIRLS, TIMSSがCBT化 (Alrababah, S. A., & Molnár, G., 2021)

国内でも様々な資格試験などでCBTが導入されている

e.g., アクチュアリー試験、情報処理推進機構の試験、統計検定

文科省は公的CBTシステム(MEXCBT)の運用を開始 ▶ MEXCBT メクビット

従来の紙筆式 (PBT) にはない様々なメリットがある

紙を印刷しなくても良い，受験者を一箇所に集めなくても良い，
様々な「新しい出題形式」ができる，解答内容以外のログデータが取れる etc.

（ちなみに）CBTはいつ頃からあったのか？

Web of Scienceで見つかる最古の論文 の一つ

▶ Churchill, S., Naess, L., & Olivier, W. P. (1971). CAN-4, an advanced author language for CAI, computer-based testing and psychological experimentation: PDP-9 implementation. Behavior Research Methods & Instrumentation, 3(2), 95–99. https://doi.org/10.3758/BF03206998

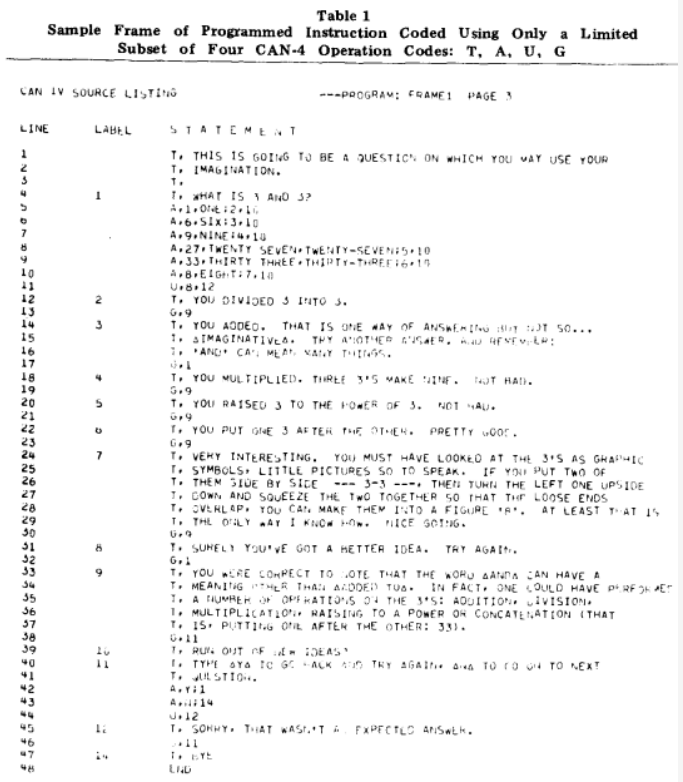


Table 1
Sample Frame of Programmed Instruction Coded Using Only a Limited Subset of Four CAN-4 Operation Codes: T, A, U, G

```
CAN IV SOURCE LISTING                    ---PROGRAM: FRAME1  PAGE 1

LINE      LABEL     S T A T E M E N T

1                   T, THIS IS GOING TO BE A QUESTION ON WHICH YOU MAY USE YOUR
2                   T, IMAGINATION.
3                   T,
4         1         T, WHAT IS 3 AND 3?
5                   A,1,ONE;2,10
6                   A,6,SIX;3,10
7                   A,9,NINE;4,10
8                   A,27,TWENTY SEVEN,TWENTY-SEVEN;5,10
9                   A,33,THIRTY THREE,THIRTY-THREE;6,10
10                  A,8,EIGHT;7,10
11                  U,8,12
12        2         T, YOU DIVIDED 3 INTO 3.
13                  G,9
14        3         T, YOU ADDED.  THAT IS ONE WAY OF ANSWERING BUT NOT SO...
15                  T, IMAGINATIVE.  TRY ANOTHER ANSWER, AND REMEMBER:
16                  T, 'AND' CAN MEAN MANY THINGS.
17                  G,1
18        4         T, YOU MULTIPLIED. THREE 3'S MAKE NINE.  NOT BAD.
19                  G,9
20        5         T, YOU RAISED 3 TO THE POWER OF 3.  NOT BAD.
21                  G,9
22        6         T, YOU PUT ONE 3 AFTER THE OTHER.  PRETTY GOOD.
23                  G,9
24        7         T, VERY INTERESTING.  YOU MUST HAVE LOOKED AT THE 3'S AS GRAPHIC
25                  T, SYMBOLS, LITTLE PICTURES SO TO SPEAK.  IF YOU PUT TWO OF
26                  T, THEM SIDE BY SIDE  --- 3-3 ---, THEN TURN THE LEFT ONE UPSIDE
27                  T, DOWN AND SQUEEZE THE TWO TOGETHER SO THAT THE LOOSE ENDS
28                  T, OVERLAP, YOU CAN MAKE THEM INTO A FIGURE 'B'.  AT LEAST THAT IS
29                  T, THE ONLY WAY I KNOW HOW.  NICE GOING.
30                  G,9
31        8         T, SURELY YOU'VE GOT A BETTER IDEA.  TRY AGAIN.
32                  G,1
33        9         T, YOU WERE CORRECT TO NOTE THAT THE WORD 'AND' CAN HAVE A
34                  T, MEANING OTHER THAN 'ADDED TO'.  IN FACT, ONE COULD HAVE PERFORMED
35                  T, A NUMBER OF OPERATIONS ON THE 3'S; ADDITION, DIVISION,
36                  T, MULTIPLICATION, RAISING TO A POWER OR CONCATENATION (THAT
37                  T, IS, PUTTING ONE AFTER THE OTHER: 33).
38                  G,11
39        10        T, RUN OUT OF NEW IDEAS?
40        11        T, TYPE 'GO' TO GO BACK AND TRY AGAIN, 'END' TO GO ON TO NEXT
41                  T, QUESTION.
42                  A,*,1
43                  A,#,14
44                  U,12
45        12        T, SORRY, THAT WASN'T AN EXPECTED ANSWER.
46                  U,11
47        14        T, BYE
48                  END
```

CBTを実施するための言語に関する論文
コンピュータに問題を表示し，**入力された解答に応じて異なるフィードバック**を表示できたりする
【左のコードによる例】問：想像力を働かせて答えてね。What is 3 and 3?

1 →「割り算だね」
6 →「足し算だね。いいんだけどもっと想像力を出してみよう」
9 →「掛け算だね。悪くないよ。」
27 →「累乗だね。悪くないよ。」
33 →「3を並べたんだね。かなりいいよ。」
8 →「ひっくり返した3と重ねてみたんだね。素晴らしい。」
その他 →「想定していなかった解答です。」

CBTの進化

CBTの進化に関する主要な論

Bennett (1998, 2015)

Computer-Based Tests — **第1世代**

従来のPBTを
CBTに焼き直す

Electoronic Tests — **第2世代**

出題形式の変化
（マルチメディア・TEIs）

適応型テスト
（個人ごとに異なる項目）

Web上での実施

Generation “R” — **第3世代**

より「リアル」な出題形式
（シミュレーション・VRなど）

複雑な項目の
自動採点

自動作問

「学習のためのテスト」へ

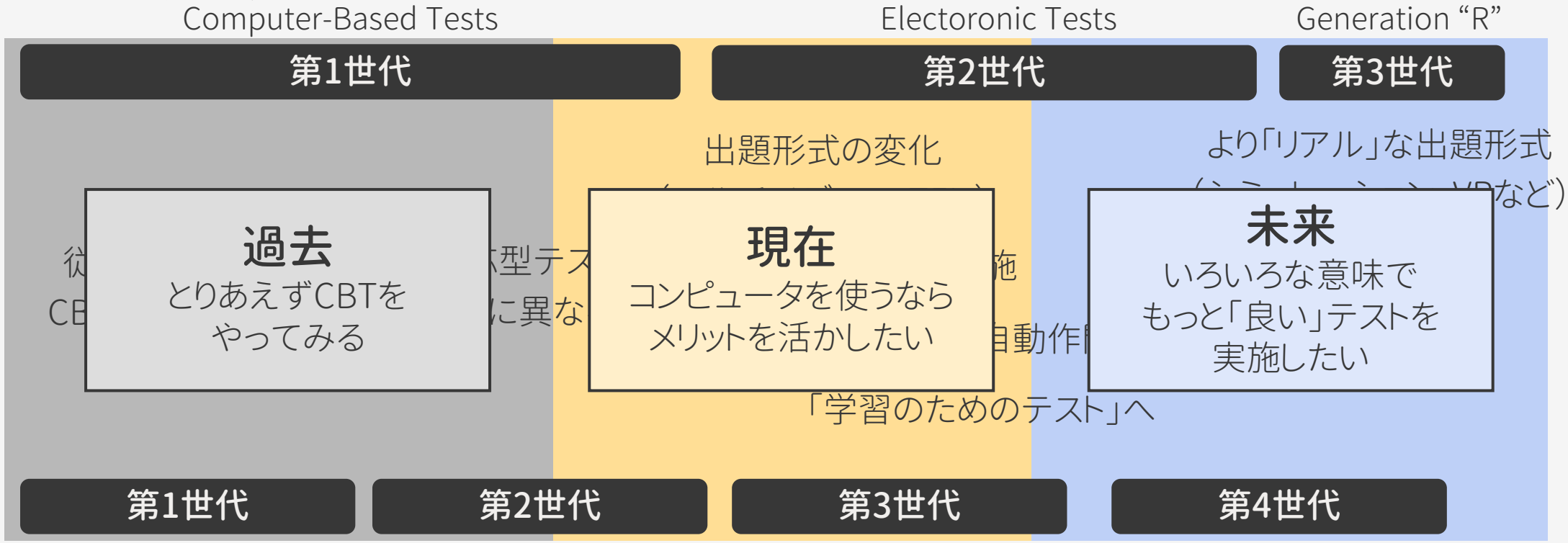
Bunderson et al. (1988)

第1世代	第2世代	第3世代	第4世代
Computerized Testing	Adaptive Testing	Continuous Measurement	Intelligent Measurement

CBTの進化

「過去・現在・未来」を当てはめるなら…（私見）

Bennett (1998, 2015)



Bunderson et al. (1988)

本講演の目的

CBTに関する主要研究トピックの紹介

主にハード面ではこれまでにどのような研究が行われてきたか
最新の研究動向を交えて紹介

CBT研究の展望を想像（あるいは妄想）してみる

Bennett (1998, 2015)やBunderson et al. (1988)の予測に沿って

今回は「能力測定」を目的としたCBTに限定

心理尺度やアンケート（いわゆるウェブ調査）など「正解がない」ものは除外
（ただそのような調査にも適用できる情報もあります）

主要な研究トピック (分寺, 2023)

CBTの進化に関する主要な論

Bennett (1998, 2015)

Computer-Based Tests　　Electoronic Tests　　Generation “R”

第1世代　　**第2世代**　　**第3世代**

6 特別な配慮

出題形式の変化
（マルチメディア・TEIs）

3 新しい形式の項目

「リアル」な出題形式
（シミュレーション・VRなど）

5 ログデータの活用

従来のPBTを
CBTに焼き直す

2 適応型テスト

（個人ごとに異なる項目）

Web上での実施

4 不正対策

自動作問

複雑な項目の
7 CBTの自動化

1 CBTとPBTの比較

「学習のためのテスト」へ

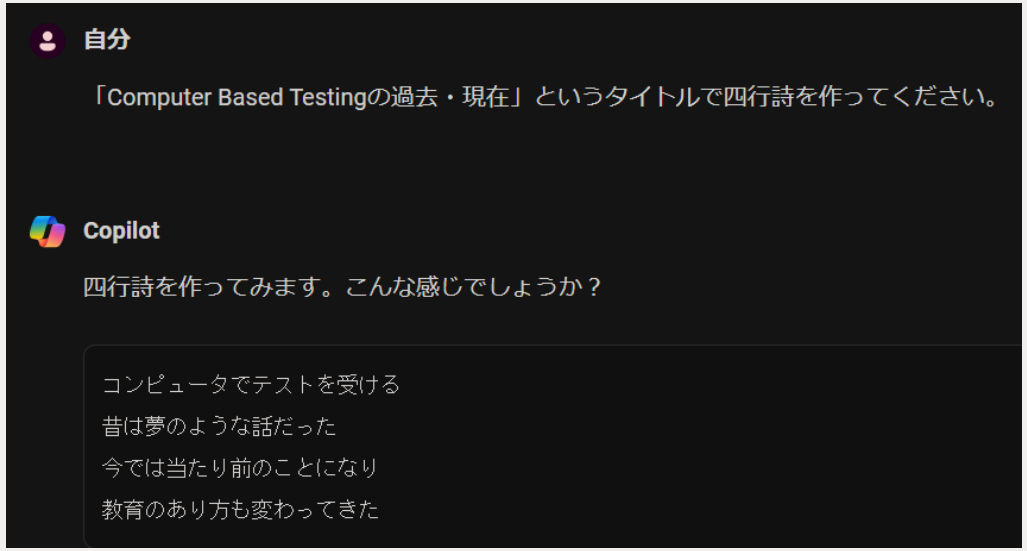
第1世代　　**第2世代**　　**第3世代**　　**第4世代**

Computerized Testing　　Adaptive Testing　　Continuous Measurement　　Intelligent Measurement

Bunderson et al. (1988)

目次にもどる

2
CBT研究の過去・現在



- 以下の6トピックのご紹介

1. CBTとPBTの比較
2. 適応型テスト
3. 新しい形式の項目
4. 不正対策
5. ログデータの活用
6. 特別な配慮

Ch. 2目次にもどる

1

CBTとPBTの得点の比較

PBTからCBTへ

CBT黎明期の主要なトピックのひとつ

CBTのつくりかた

大きく分けると2パターン

第1世代

既存のものをCBT向けに作り変える

- 「とりあえずCBTで」
- 作成済みのリソースを生かせる
- PBTとCBTを共存させることも可能

こちら側のお話

新世代

イチからCBT用のアセスメントを作成する

- 様々な問題タイプが開発可能に
  - Technology Enhanced Items
  - 音声や動画を駆使した項目

PBTのCBT化に関する懸念点：公平性の問題

2つのモードを並行して実施する場合、**受験モード間で不公平があってはならない**

CBTに関するガイドライン

国際テスト委員会（ITC）のガイドライン（2005）

> Where the CBT/Internet test has been developed from a paper and pencil version, **ensure that there is evidence of equivalence.**

AERA, APA, NCMEのガイドライン (2011) でも同様のことが言われている

具体的には↓

1. 同じ程度の信頼性を持つ
2. 信頼性推定により期待されるレベルで，互いに相関する
3. 他のテストや外部基準とは，同じ程度に相関する
4. 同じ程度の平均値と標準偏差値を持つ，あるいは**同じ程度のスコアを生み出すように適切に調整がなされている**

モード効果 mode effect

受験形式の違いが結果に与える影響

すでにレビューはたくさんある(e.g., Alkhadher et al., 1994; Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; S. Wang et al., 2008)

モード効果を引き起こす要因(一例)

受験者側

- 人種・性別などのデモグラフィック属性
  (Gallagher et al., 2000; Parshall & Kromrey, 1993)
- コンピュータの経験値・親和性
  (J. A. Lee, 1986; Taylor et al., 1999)
- 読解の認知過程(Mayes et al., 2001; Noyes & Garland, 2003)
- テストストラテジーの変化(Alkhadher et al., 1994)

ページめくりが面倒だと「一旦飛ばす」が起こりにくくなる?

テスト側

- 画面表示(Ziefle, 1998; Bridgeman et al., 2003)
  (画面サイズ、解像度、フォント、一画面に提示する項目数など)
- ユーザーインタフェース(e.g., Revuelta et al., 2003)
  (スクロール、ページ送り、見直しの可否)
- 制限時間(Alkhadher et al., 1994; Mead & Drasgow, 1993)

どの要因がいつどんな効果をもたらすかは「場合による」
▶ 手元の環境下でモード効果が生じるかは逐一チェックするのがよさそう

検証のためのデザイン・データ分析方法はかなり多様です

モード効果検証のためのデザイン(Arce-Ferrer & Bulut, 2019)

主に実験的な環境

1. 同一受験者に複数のモードに取り組んでもらう
2. 各受験者にどちらか一方のモードを割り当てる（ランダム, カウンターバランス）

主に実データ環境

3. モード間で差がないことを検証済みのアンカーテストを用いる
4. 受験者が自由にモードを選択した結果を利用する

⟶ 選択自体がモード効果の要因と関係する可能性がある

▶ **準実験的アプローチを用いる**（傾向スコアなど: J. Liu et al., 2016; Seo & De Jong, 2015）

得点の比較可能性の条件

異なるモード・試験における得点が比較可能と言えるためには (Pommerich, 2016)

1 Distributional equivalence ➡
対応付けた後の得点の分布が同じ

平均点の比較（t検定，回帰分析など）
SDの比較（分布の等質性を見るため）
IRTによる方法（項目特性曲線の比較: Raju et al., 1995）
▲ テストおよび項目ごとにチェック(Arce-Ferrer & Bulut, 2019)

2 Construct equivalence ➡
構成概念が同じ

SEM（多母集団同時分析）(Schroeders & Wilhelm, 2011)
IRT（多母集団モデル）(Buerger et al., 2016)
モード間での得点の相関が十分にあるか(Buerger et al., 2016)

3 Predictive equivalence ➡
外的指標の相関が同じ

外的指標との相関をとる
（ここまで行っている研究は現状かなり少ない）

Distributional equivalence

Lord (1980)→Morris (1982) による*equity*の定義

真の能力値を θ としたときに，2つのテストで得られる得点 $s_{\mathrm{PBT}}, s_{\mathrm{CBT}}$ の関係が

$$E[s_{\mathrm{PBT}}|\theta] = E[s_{\mathrm{CBT}}|\theta], \text{for all } \theta$$

▶どんな能力の個人でも，2つのテストを受けたときの得点の期待値が等しい

同様にequal precision

$$\sigma(s_{\mathrm{PBT}}|\theta) = \sigma(s_{\mathrm{CBT}}|\theta), \text{for all } \theta$$

個人レベルだけでなく集団レベルでも

θ の分布自体が等しくなっていてほしい (equal distribution property)

※サブグループ単位で確認するのが望ましい

Ch. 2目次にもどる

2

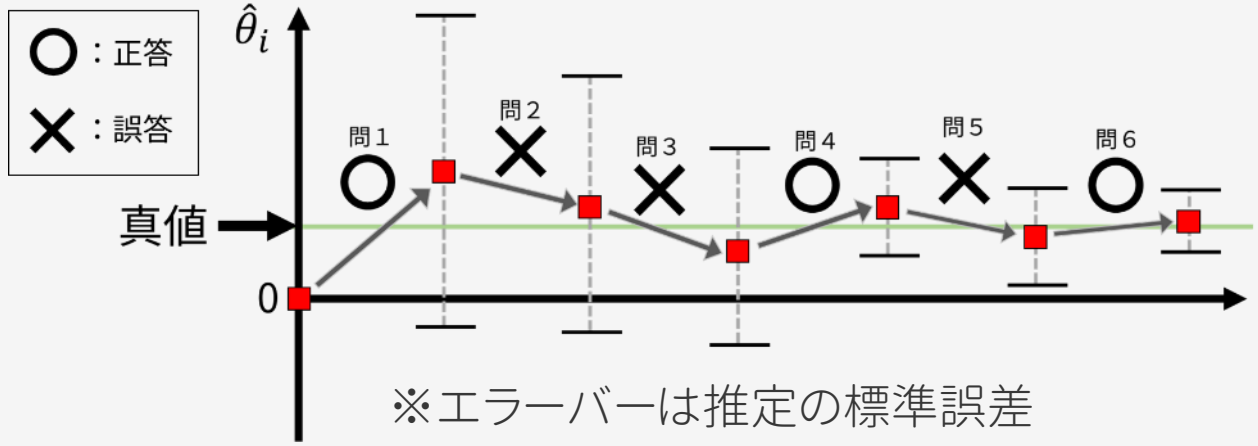
適応型テスト

Computerized Adaptive Testing

適応型テストとは

個人ごとに異なる項目を提示する

（視力検査のようなイメージ）



適応型テストの流れ

1. 特性値の初期値を決定する
2. 項目プールから出題項目を選択する
3. 項目を出題して解答データを得る
4. 特性値を推定（更新）する
5. 終了基準を満たしているかを判定し，満たしていなければ2.に戻る

現在では幅広く実用化されている

PISAなどもすでに導入済み(Yamamoto, Shin, & Khorramdel, 2019)

厳密にはMulti Stage Testing (MST)

適応型テスト黎明期

アイデア自体はCBTとは無関係に存在していた

世界初の適応型測定はビネー式知能検査(Binet & Simon, 1905)と言われている

コンピュータ発達以前にもいくつか研究は見られる

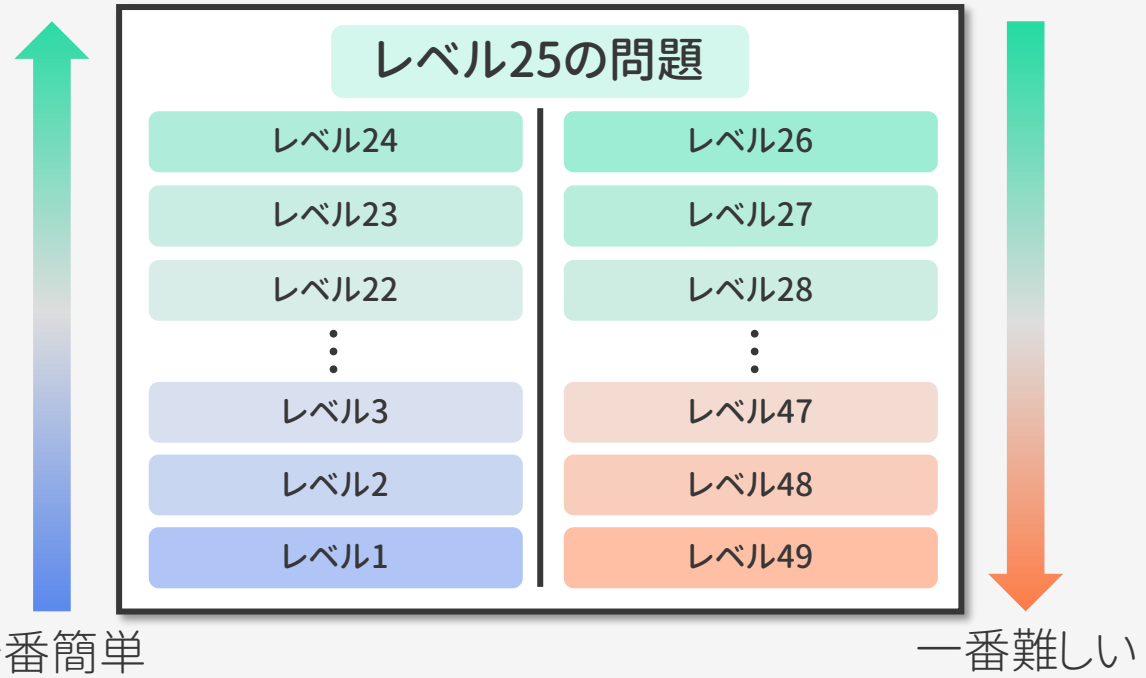
例｜Lord(1971) のFlexilevel test（紙で実施可能）

最初に真ん中の項目に答える
▶ 正解したら右の一番上
▶ 不正解なら左の一番上

次以降も…
▶ 正解したら右の**未解答の**一番上
▶ 不正解なら左の**未解答の**一番上

基本的な考え方は今も同じ

（全49問のテストの場合）



適応型テストのポイント

適応型テストの流れ（再掲）

1. 特性値の初期値を決定する
2. 項目プールから出題項目を選択する
3. 項目を出題して解答データを得る
4. 特性値を推定（更新）する
5. 終了基準を満たしているかを判定し，満たしていなければ2.に戻る

どうやって初期値を決める？
初期の推定は？（e.g., 最尤法が使えない）

（いちばんだいじ）
どうやって次の項目を選ぶ？

推定法は？（e.g., 最尤法・ベイズ）

終了基準は？（e.g., 項目数・推定精度・制限時間）

項目選択アルゴリズム

3つの要素(Han, 2018)

推定精度を高める項目は
多くの人にとって「良い」ため
出題回数が増えがち

推定精度を高める

- 項目情報量ベース
- ベイズ推定

出題回数の偏りを抑える

- 乱数を用いる
- 条件付き確率で考える
- 項目パラメータで層化する

出題回数の偏りを抑えるためには
あまり性能の良くない項目も
出す必要が出てくる

人によって出題領域の
バランスが異なる
かもしれない

出題領域など他の要素を考慮する

- 出題数の少ない領域から選ぶ
  (Kingsbury & Zara, 1989)
- 仮想テスト版をつくる

推定精度を高める代表的な項目選択法

項目情報量

（IRT）特性値θの推定値の分散はテスト情報量の逆数（＝フィッシャー情報量） $V(\theta_i) = I(\theta_i)^{-1}$

テスト情報量は項目情報量の和 $I(\theta_i) = \sum I_j(\theta_i)$

期待正答率が50%に近いほど
項目識別力が高いほど
} **項目情報量は多くなる**

→ **項目情報量$I_j(\theta_i)$が最大の項目を選べばよい**(Birnbaum, 1968)

尤度関数で重みづけたり(Veerkamp & Berger, 1997)
代わりにKL情報量を使ったりもする(Chang & Ying, 1996)

ベイズ事後分布

事後分布の分散の期待値が最小になる項目を選べばよい(Owen, 1975; Thissen & Mislevy, 2000)

出題回数の偏りを抑える代表的な項目選択法 レビューはGeorgiadou et al. (2007)

Randomesque法(Kingsbury & Zara, 1989)

候補を複数用意し、その中から等確率で選択する

条件付き確率を用いた方法

出題される確率　選ばれる確率　選ばれた項目が出題される確率

$P(A) = P(S) \times P(A \mid S)$

ここを操作して出題確率を調整してあげる(Sympson & Hetter, 1985)

自動的に決まる(van der Linden, 2003)

項目プールや項目選択基準、受験者のレベルなどによって

▲それまでの受験者での出題割合をもとに調整する方法も(Barrada et al., 2009; van der Linden & Veldkamp, 2004)

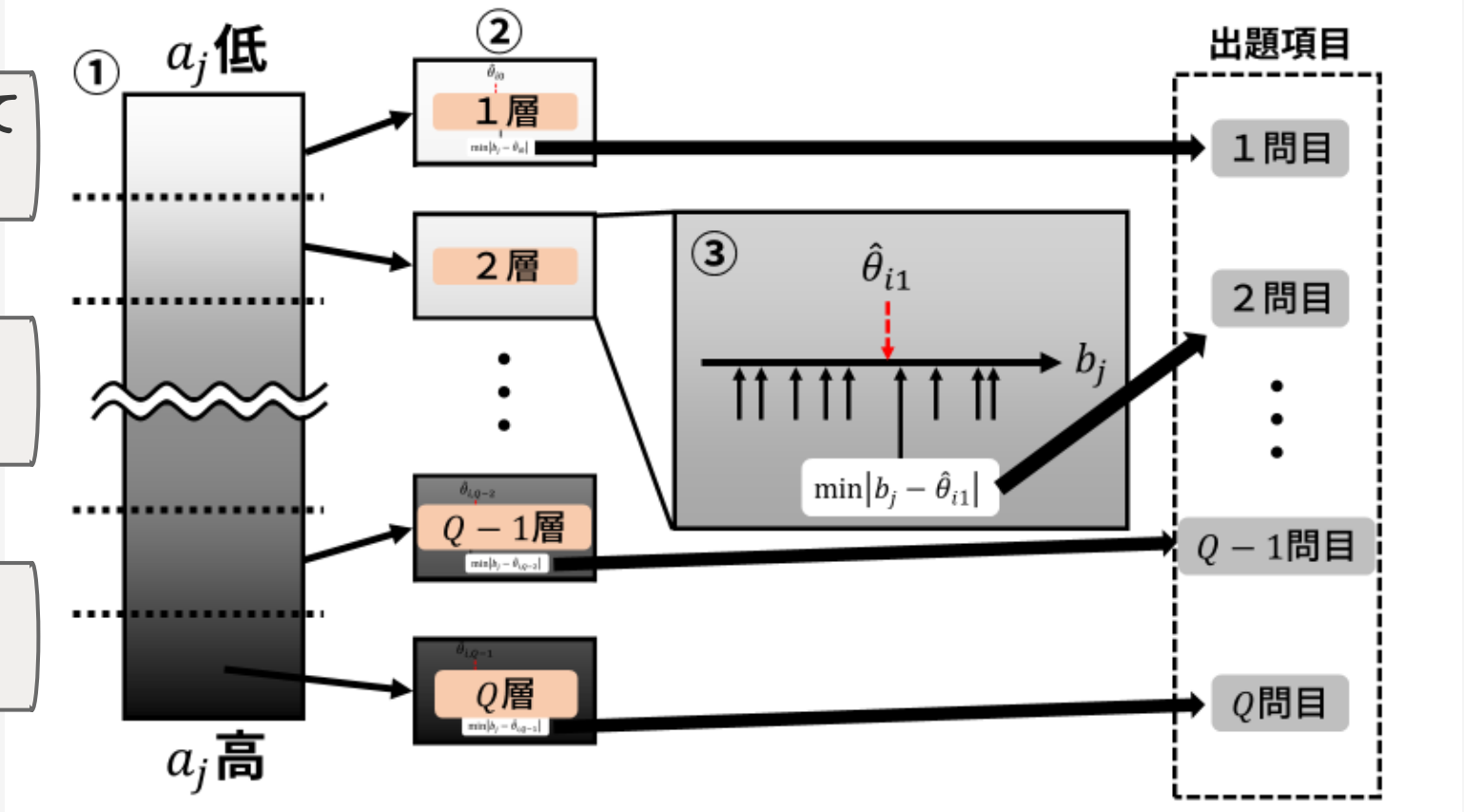
出題回数の偏りを抑える代表的な項目選択法　レビューはGeorgiadou et al. (2007)

a-層化法(Chang & Ying, 1999)

項目情報量ベースでは識別力 (a) が低い項目ほど出題回数が減る

▶ 識別力が低い項目も意図的に出題することで偏りを抑えられる

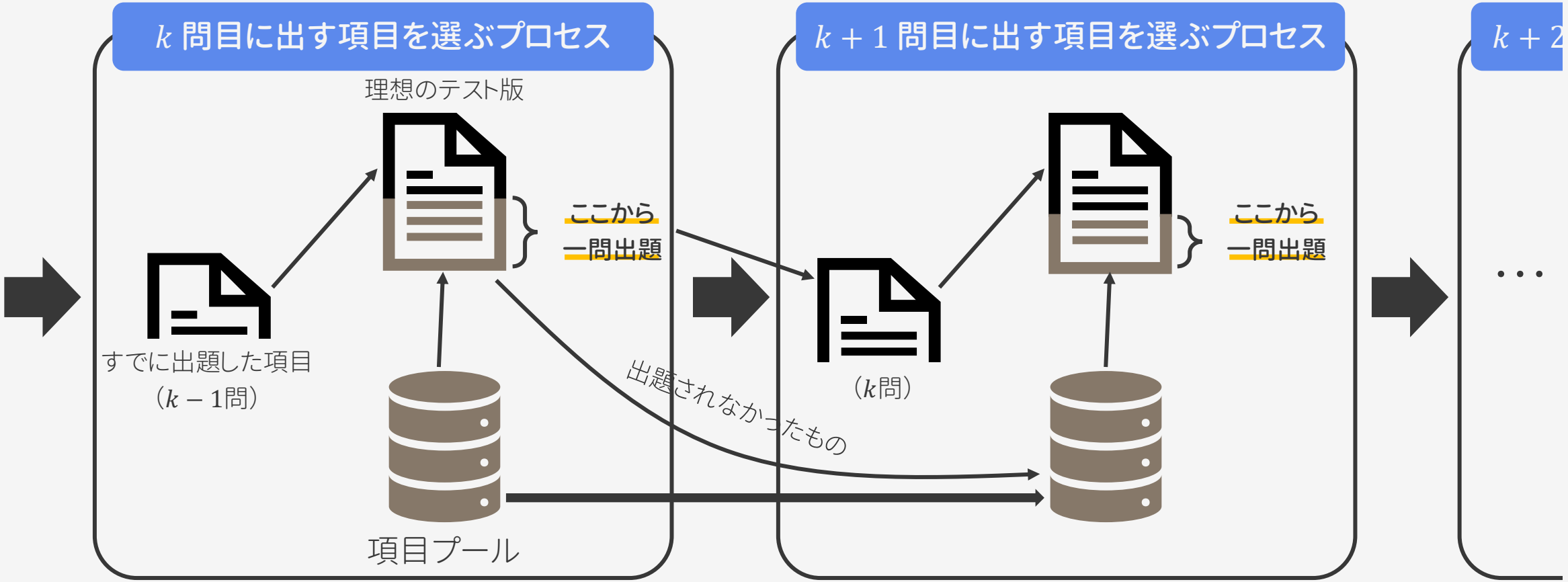
1. 全項目を識別力の昇順に並べて等分割する
2. 識別力の低い層から順に各層から1問ずつ出題していく
3. 各層からは，暫定の推定値と困難度が最も近い項目を選ぶ



出題領域などいろいろな要素を同時に考慮する項目選択法

Shadow test (van der Linden & Veldkamp, 2004)

整数計画法を用いて「条件を満たすテスト版」の中でテスト情報量が最大の版を作る



様々な目的に応じた項目選択法

ClassificationのためのCAT (AMT: Kingsbury & Weiss, 1979; CCT: Parshall et al., 2002)

▶ 個人の能力（連続変数）を推定するよりも**「合格／不合格」等のカテゴリを精度良く識別できれば良い**

カットオフポイント前後での正答確率のオッズ比が最大の項目を選ぶ(Lin, 2000)

カテゴリ所属確率を最もよく識別するような項目を選ぶ(Rudner, 2009)

ノンパラメトリックなCAT

▶ 小規模なテストなど，**IRTが使いにくいような状況においても適応的に選択したい**

決定木ベースで「プール内の全項目に解答した場合の正答数」を最小項目数で予測する (Yan et al., 2004)

認知診断モデル (CD-CAT; Wang et al., 2012)

▶ 個人の能力を離散変数（の組み合わせ）として考えるので，それに合わせた選択法を使うべき

各アトリビュートに関する項目の出題数を揃える (MPI)

異なるアトリビュートパターンでの正答率が最も異なる項目を選ぶ

アルゴリズム比較の指標

いろいろな項目選択法が提案されている中で，「どれを使えばいいの?」を考える

基本的な指標は2種類

推定精度を高める ⇄ 出題回数の偏りを抑える

RMSEなど

Overlap Rate (OR)
(Way, 1998)

多くの場合，ORは
- テストセキュリティの重要度
- 項目開発と使用回数のペース

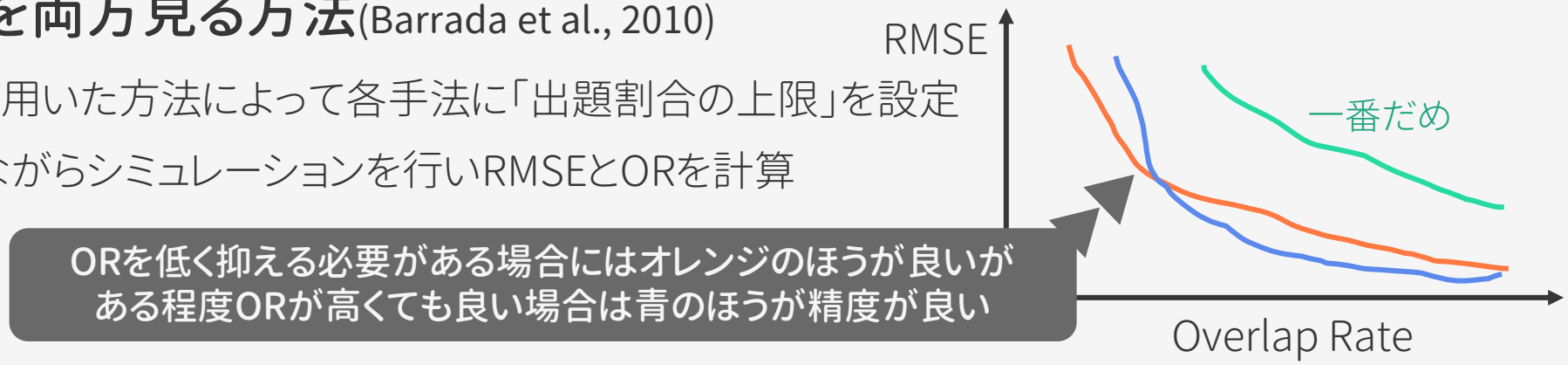
等に基づいてテスト実施者が調整

任意の二人の受験者に出題された項目のうち
共通して出題された項目の割合の期待値

RMSEとORを両方見る方法(Barrada et al., 2010)

条件付き確率を用いた方法によって各手法に「出題割合の上限」を設定

▶ 上限を変えながらシミュレーションを行いRMSEとORを計算



ORを低く抑える必要がある場合にはオレンジのほうが良いが
ある程度ORが高くても良い場合は青のほうが精度が良い

Adaptivity (e.g., Reckase et al., 2019; Wyse & McBride, 2021)

特にプール内の項目数が少ないとき

その適応型テストは本当に「適応」的に項目を出し分けているのか？

▶ 適応的ならば、個人の θ に応じて異なる項目が出ているはず

Adaptivityが低い場合は
もっと項目を用意する必要があるかもしれない

【Adaptivityの指標】

1. **Reckase et al. (2018)の考え方**

   推定値 $\hat{\theta}$ と出題された項目の困難度の平均 $\bar{b}$ の相関が高い、など

2. **Ju and Reckase (2019)の考え方**

   各時点での推定値 $\hat{\theta}$ と出題された項目の困難度の平均 $\bar{b}$ の差が小さい、など

3. **Wyse and McBride (2021)の考え方**

   項目選択法的に理想的な項目 (target) と実際の項目の困難度の差が小さい

Ch. 2目次にもどる

3

新しい形式の項目

Technology-Enhanced Items (TEIs)

コンピュータ技術が項目にもたらすイノベーション

1 PBTでは表示できない要素を含んだもの

映像，音声，3Dグラフィック，アニメーション，など

2 選択肢を選ぶ・解答を入力する以外の解答方法

テキストのハイライト，画像の一部をクリック，オブジェクトを動かす，など

▶ **これらをひっくるめてTechnology Enhanced Items (TEIs)と呼んでいる**

Russell (2016) などでは②のみをTEIsと呼ぶなど，細かい定義は人によって多少異なる

コンピュータ技術が項目にもたらすイノベーション

1 PBTでは表示できない要素を含んだもの

佐賀大学CBT

動画を見て解答する

安野 (2021)

3Dオブジェクトを操作して解答する

2.3.4 事物・現象の観察

本研究では，受験者の ICT スキルの影響をできるだけ軽減させるために，タブレット端末を利用した CBT を提案している。タブレット端末は，基本操作を「タップ」で行うが，特徴的な機能として，「ピンチイン」及び「ピンチアウト」によって，拡大・縮小が容易に行える。そこで，この機能も化学の事物・現象の観察で役立つと考え，様々な方向から撮影した蒸留装置の画像を拡大・縮小して観察する問題も開発した。2.3.2 での 3D モデルでも同様のことが行える。

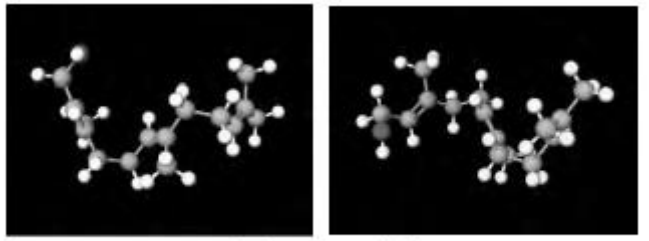


図 8 3D の例

TEIsで利用できる情報とは

1 PBTでは表示できない要素を含んだもの

映像，音声，3Dグラフィック，アニメーション，など

▶ 文字情報以外が含まれるのでmultimediaとも呼ばれる (e.g., Bennettほか, 1999; Dirkxほか, 2021)

具体的に用いられる要素

視覚情報

- 画像：CBTならば拡大・回転が自由自在
  ▶ PBTよりも細かい情報が載せられる(Drasgow, 2002)
- 文字だけよりも画像や動画を組み合わせると
  テストの成績が向上する
  （multimedia effect: e.g., Lindnerほか, 2017; Mayer, 1989; She & Chen, 2009）

聴覚情報

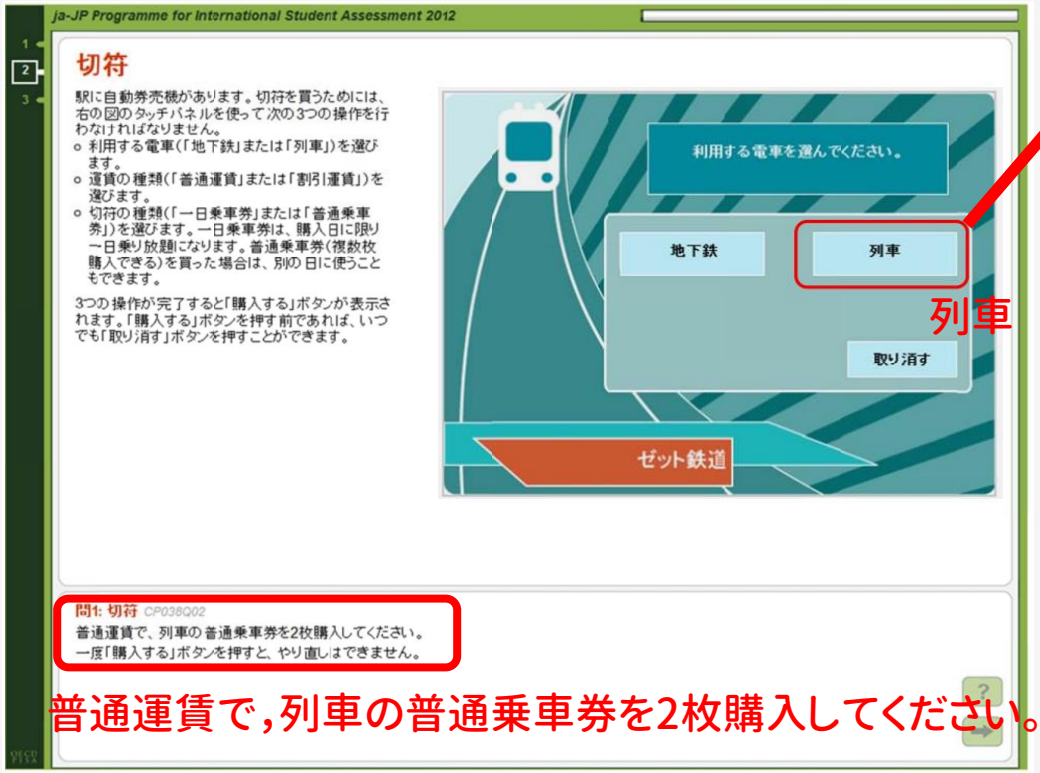
- 4種類の利用方法
  - primary / supplemental
    ▶ supplementalな情報は
    インタフェースの理解に役立つ
  - speech / nonspeech

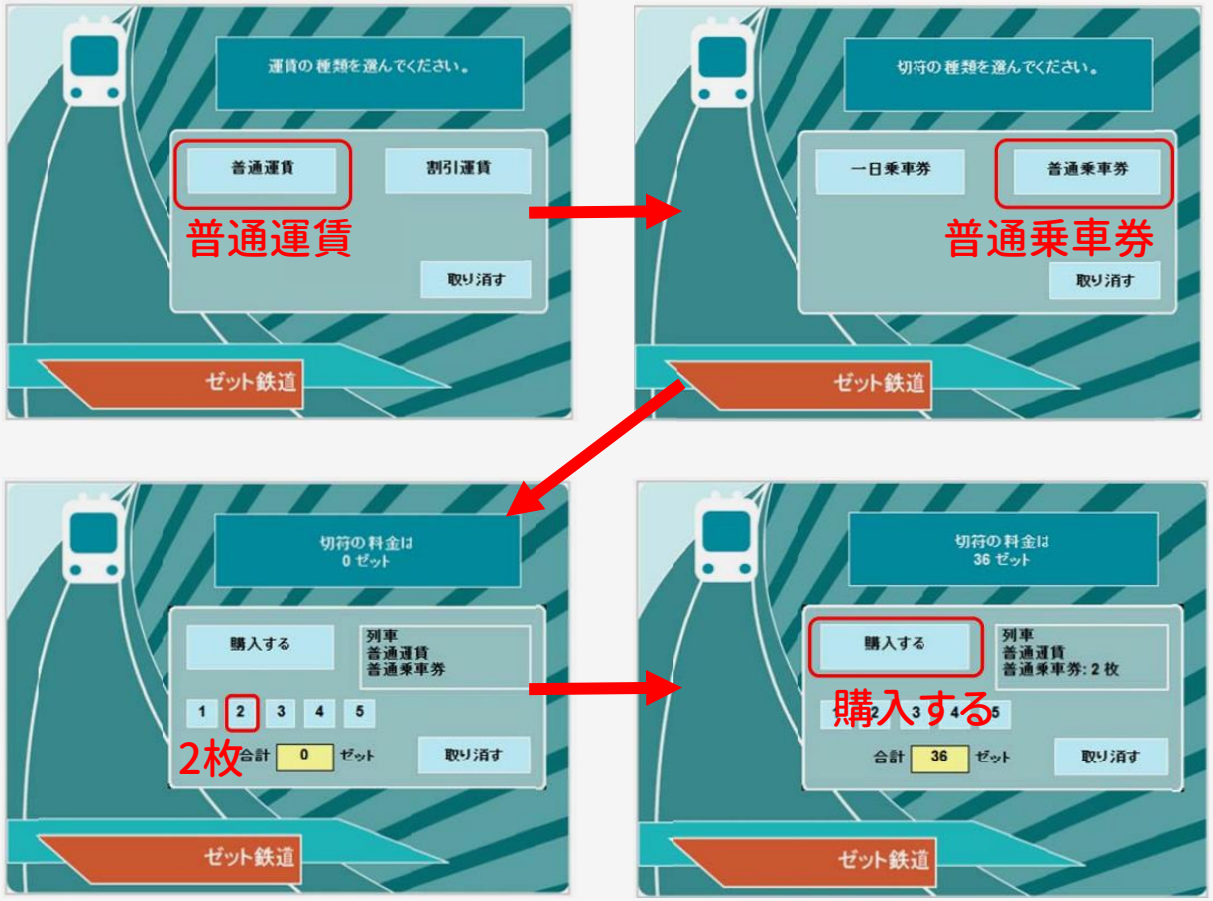
コンピュータ技術が項目にもたらすイノベーション

② 選択肢を選ぶ・解答を入力する以外の解答方法

PISA2012「切符」

券売機を模した画面を操作して解答する





新しい解答方法がもたらすメリット・デメリット

2 選択肢を選ぶ・解答を入力する以外の解答方法

従来の項目では測定できなかった構成概念や能力を測定できる (Boyle & Hutchison, 2009)

解答行動をより現実場面に近づけることで測定の妥当性を高める (Bryant, 2017)

一方でTEIsには懸念点も

内容や解答手順が複雑になる

▶ 本来測定したい能力以外（e.g., タイピング速度やマウス操作の精度など）の影響

▶ **適当に作ってしまうとかえって妥当性が低下する危険性もある**

新規性の高い項目の作成は慎重に行われるべき

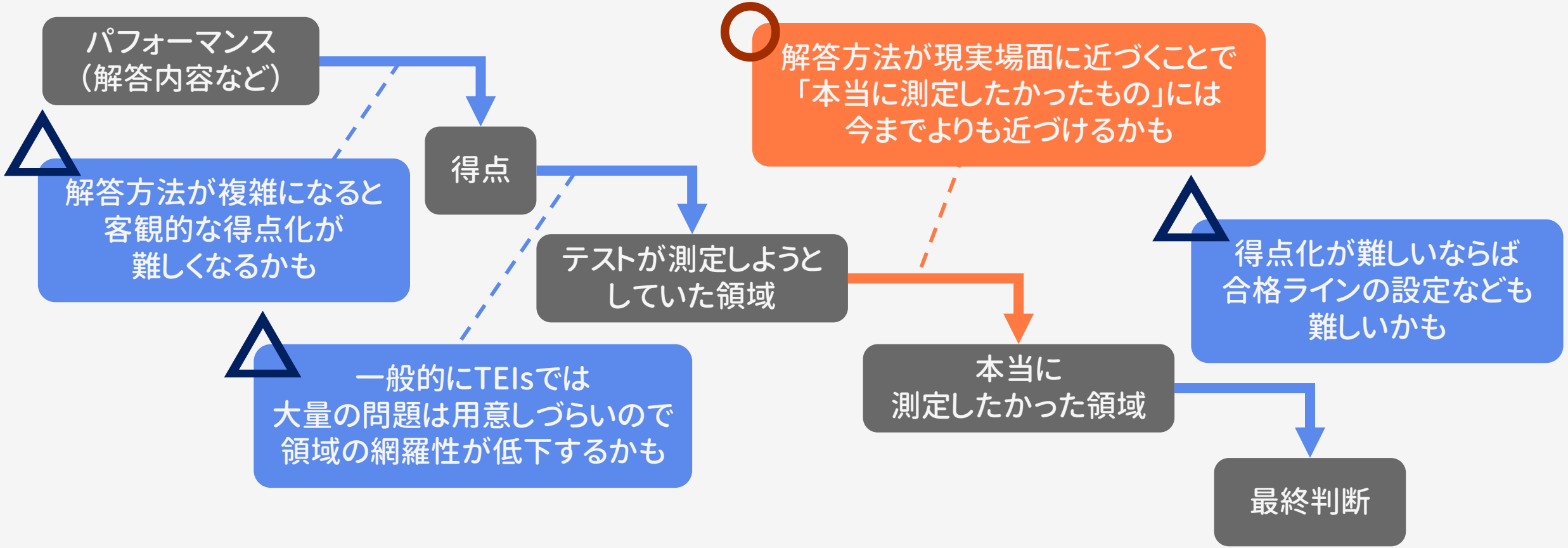
通常の作問プロセスに加えてhuman-computer interactionに沿ったユーザビリティの検証を行うべき (Parshall & Harmes, 2014)

fidelity, usability, accessibilityが満たされて初めてTEIsは意味を成す (Russell, 2016)

TEIsが妥当性に及ぼす影響 (Wools et al., 2019)

Kaneのフレームワークに基づいて考えると

解答内容から意思決定が行われるまでには様々な推論が含まれているとして
TEIsの場合それぞれの推論にはどう影響するか考えてみると…



より真正性の高い測定

Simulation-Based Assessment

従来のアセスメントであれば多肢選択問題の連続で問うていたが
SBAでは解答に応じてその先の結果が変わってくる

例｜アメリカの医師国家試験の一部

患者が来る ▶ 様子を見る ▶ 何をするか決める ▶ 結果を見る ▶ …

バイタルサイン
過去の記録 etc.

検査
様子見
病棟の移動 etc.

選択した行動に
応じた結果が
返ってくる

Game-Based Assessment

アセスメントにゲームの要素を取り入れて
エンゲージメントや満足度を高めること、という感じ(Grelle & Ellinikakis, 2022)

いわゆるシリアスゲームによるアセスメント（≠ゲーミフィケーション）(Grelle & Ellinikakis, 2022)

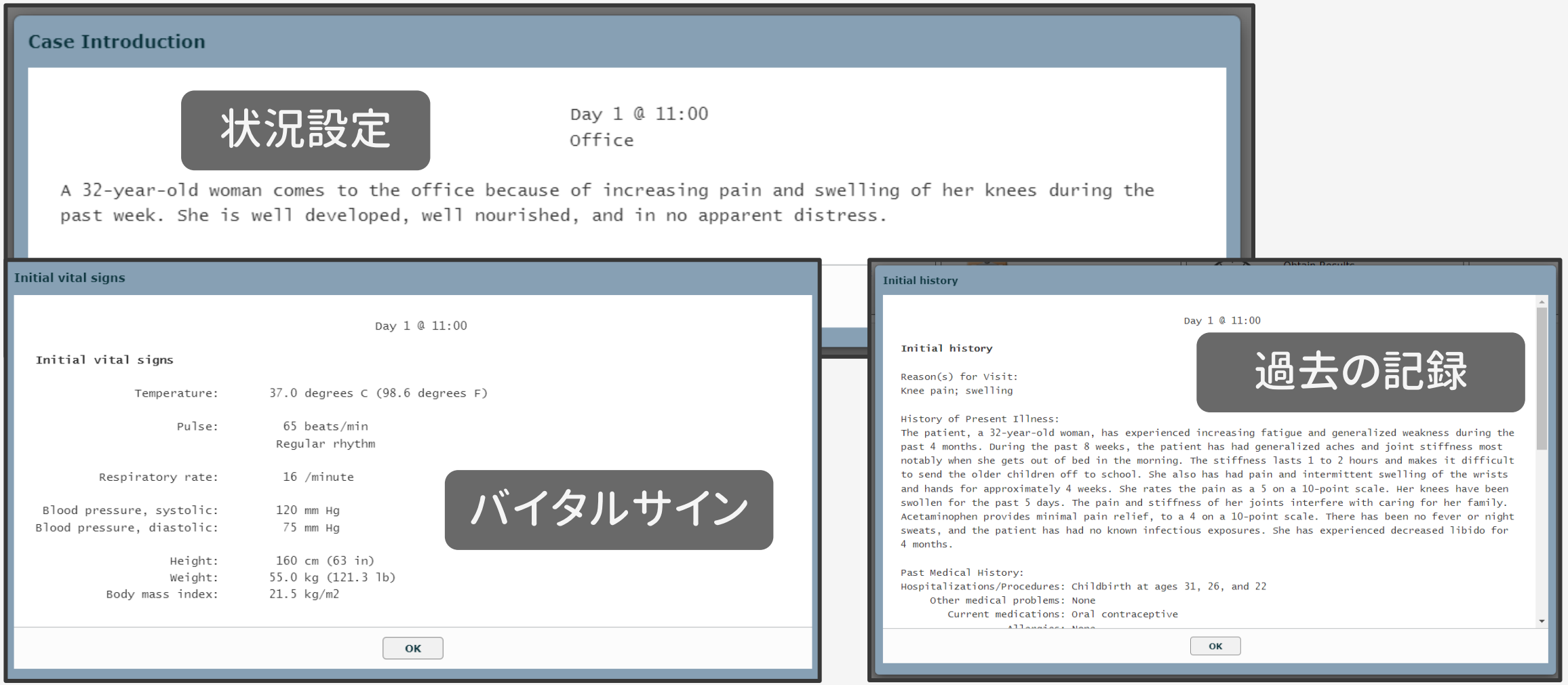
▲ 主目的が「能力を測定すること」で、ゲーム全体がその目的のために設計されている

【GBAにおける評価方法】(Mislevy et al., 2016, SBAでも同じ気がする)

1. プレイヤーがゲーム外で発揮したパフォーマンス (big-G Game: Gee, 2008)
2. ゲーム内で発揮したパフォーマンス（成果物）
3. 課題解決の過程で発揮したパフォーマンス（ログデータ・プロセスデータ：後述）

United States Medical Licensing Examination (USMLE)

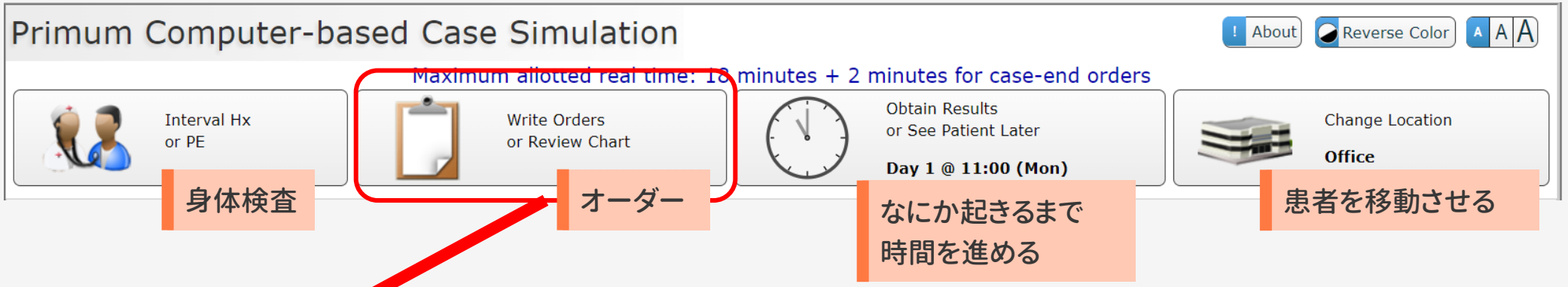
Step3：ケースシミュレーション

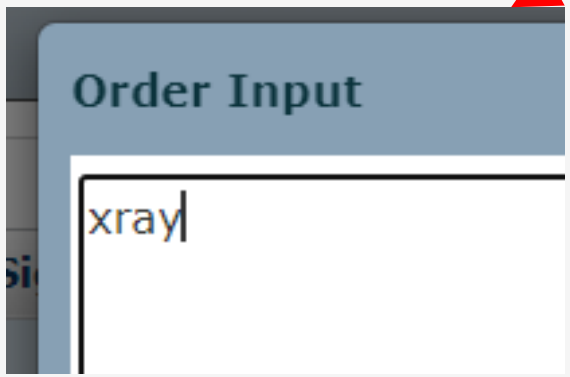


United States Medical Licensing Examination (USMLE)

ケースシミュレーション

何をするか選ぶ





自由記述で入力
（検索する）

オーダーを選ぶ
（2300種類ほど）

Clauser et al. (2016)

更に進んだTEIs

ゲームの空間で行う (Peters et al., 2021)

どちらかというとアセスメント＋ゲーミフィケーション?

https://www.sciencedirect.com/science/article/pii/S0747563221000236

Minecraftの中で空間把握能力を測定するための課題を出している

例｜同じ形のオブジェクトを複製する課題

仮想現実(VR)の空間で行う (Miskowiak et al., 2022)

Simulation-based assessmentの発展形

https://www.sciencedirect.com/science/article/pii/S0022395621007056

VR空間上のキッチンでお料理課題 ▶ 日常生活での認知機能をテストしている

アセスメントの実施環境をより現実に近づけるために
「モニターに映し出される問題」から形を変えていく?

更に進んだTEIs

ゲームの空間で行う (Peters et al., 2021)

どちらかというとアセスメント＋ゲーミフィケーション？

Minecraftの中で空間把握能力を測定するための課題を出している

▲ 与えられたブロックを使って左と同じ形を作りましょう。

仮想現実(VR)の空間で行う(Miskowiak et al., 2022)

Simulation-based assessmentの発展形

VR空間上のキッチンでお料理課題 ▶ 日常生活での認知機能をテストしている

Ch. 2目次にもどる

4

Web上でのCBTの実施

Web Based Testing / Internet Based Testing

CBTの実施形式

テストセンターなどでの実施

【長所】

- セキュリティが高い
- 実施環境をコントロールしやすい
  （デバイスなど）

【短所】

- 実施時間に制約がある
- 物理的に人を集める必要がある
  例｜情報処理技術者試験（国家試験）は2020年度中止になった（コロナのせい）

インターネット上での実施 (IBT; WBT)

(Roever, 2001)

【長所】

- 実施時間・場所の制限が少ない

【短所】

- 不正行為が起こりやすい(Harton et al., 2019)

  従来のテストでも見られた不正行為

  外部資料を盗み見るなど

  WBT特有の不正行為(Noorbehbahani et al., 2022)

  顔や声を偽装した替え玉受験(Vegendla & Sindre, 2019)

  RDPによる画面共有(von Grunigen et al., 2018)

不正対策の強化が重要

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

現状用いられている方法は大きく分けて以下の3種類

1 本人確認の強化

替え玉受験を防ぐためには従来よりも強力な方法が必要

2 オンライン試験監督 Online proctoring system

受験者の周囲がすべて見えるわけではない状況でカンニングなどを防ぐ

近年ではAIを利用した検出方法も様々に提案されている

3 試験後の検証

リアルタイムで全て検出するのは厳しいので，あとから録画を見てじっくりチェック

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

1 本人確認の強化

替え玉受験 (impersonation) を防ぐために必須

試験途中での入れ替わりを防ぐため，断続的&強力な本人確認が必要

本人確認に用いられる要素(Vegendla & Sindre, 2019)

知識情報（*know*：パスワードなど）
所持情報（*have*：ICカードなど）
} 意図的な悪意には無力（他人に渡せるため）

生体情報（*is*：指紋など）→ **これを使うしかない**

身体特性 カメラで顔を定期的に撮影する，動画を撮り続ける(Sabbah, 2017)

(Eude & Chang, 2018 ;
Monaco, 2018; Shabliy et al., 2021)

行動特性 キーストロークの癖による本人識別（キー入力継続時間や間隔など→SVM, HMM, 異常検知）

タッチペンの傾きや筆圧(林 & 赤倉, 2018)，タップ時の手の形(安田 & 小方, 2021)

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

2 オンライン試験監督 Online proctoring system

ベンダーによって提供されている機能(Hussein et al., 2020)

本人確認（前ページ）

コンピュータの機能制限（ロックブラウザなど）

遠隔での管理（怪しい受験者にフラグを立てる，中止させるなど）

怪しい行為に関するレポート作成（あとからチェックを楽にするため）

基本的には追加デバイスを使わずにできる方法が用いられる(Nigam et al., 2021)

カメラ、マイク、画面共有/録画、生体認証、視線追跡など

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

2 オンライン試験監督 Online proctoring system

受験人数の増加や受験可能時間帯の増加

▶ 近年では機械学習や人工知能を利用した試験監督システムも(Nigam et al., 2021)
AI proctoring system

近年提案された手法の例

実際に企業が提供しているものがどのような手法を用いているかは不明：企業秘密のため

Garg et al. (2020)

Haar-like特徴量を用いてWebカメラの画角から受験者が消える、別人が映り込むなどの異常を検出

Indi et al. (2021)

顔の向きを推定するFSA-Netなどを用いて受験者の視線が画面外に向いたことを検出

Prathish et al. (2016)

Webカメラに音声やスクリーンキャプチャを組み合わせて，周囲の別人や怪しい行動を検出

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

2 オンライン試験監督 Online proctoring system

AIだけで完結させるべきではない

検出率は絶対100%にはならないので，最終判断の責任は人間が持つべき(e.g., Li et al., 2015)

OPS導入には負の側面も

- 例｜静かな環境で受験できない人がいたら

  【テスト本来の目的からすると】不利になる要因は考慮して評価できると良い

  【OPSでは】騒音は「第三者」として不正行為扱いとしてしまうのか？

- OPSの侵襲性

  多面的に監視されているという事実それ自体が受験者に不安を与える(Duncan & Joyner, 2022)

  ▶ OPS導入時のほうが，テスト不安が成績により強い悪影響をもたらす (Woldeab & Brothen, 2019)

不正行為を防ぐ方法(Noorbehbahani et al., 2022)

3 試験後の検証

録画などを用いてあとから検証

解答データなどを用いた検出方法

技術の進歩（DeepFakeなど）▶ 従来の検出方法は欺かれる可能性

WBTではBYODで実施されることも多い ▶ CBTソフトウェア外部からの攻撃も(Dawson, 2016)

不正検出のための統計モデルの多くは解答内容＋解答時間を用いている(Man et al., 2019)

例｜C. Wang et al. (2018)

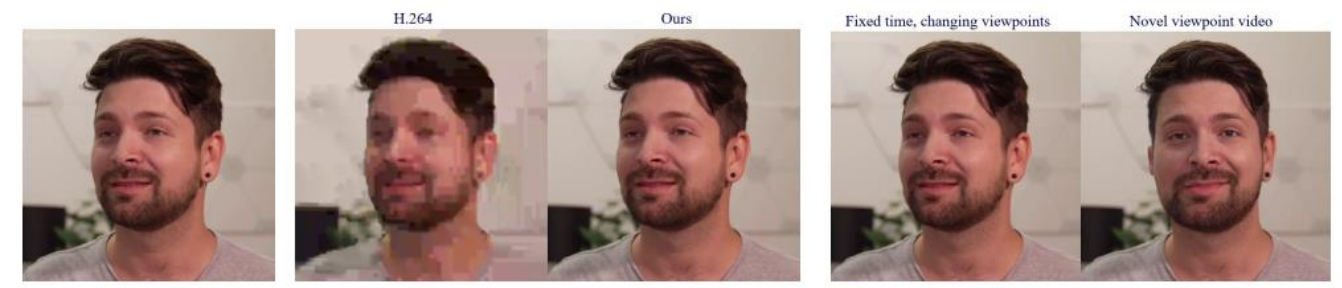
正常な解答行動と異常な解答行動（当て推量・カンニング）は
正答確率および解答時間の分布が異なることを利用して混合モデルで異常行動を検出する

不正行為に使われそうな最新技術

2024年時点ですでに「最新」ではないかもしれませんが…

Fan et al. (2022)
リアルタイムで顔を変える

Wang et al. (2021)
顔の向きや視線を変える



Yang et al. (2023)
動画から特定のオブジェクトを消す

Ch. 2目次にもどる

5

ログデータの活用

CBTで取得できるデータ

解答以外の情報がたくさんとれる

ログファイル＝CBTシステムで取れるすべての情報(Provasnik, 2021)

▲ 社会調査におけるパラデータとよく似た概念と言えそう

パラデータは調査の改善のために用いられる
ログファイルは個別の解答者の評価などにも関心がある

ログファイル（パラデータ）の分類 (Kröhne & Martens, 2011)

access-related ▶ **主に不正行為の検出に使える**

（ログイン時間、位置情報、デバイス情報など受験環境の情報）

process-related

（項目間の移動、テスト全体の所要時間などテスト実施過程の情報）

response-related

（解答選択・変更の記録、解答時間、マウス・キーボード操作の記録など解答行動の情報）

access-relatedデータ

主に不正行為の検出に用いられる

Komosny & Rehman (2022)

IPアドレスから位置情報を推定
▶ 試験前後の操作ログにおける位置（IPアドレス）の変化とタイムスタンプの差分によって不自然な速度での移動（すなわち、なりすましの可能性）を検出している

Balderas et al. (2021)

異なる受験者間のテスト開始／終了のタイムスタンプの関連を用いる
▶ すでに受験を終了した人が別の受験者の手助けをしている可能性を検出

Diedenhofen & Musch (2017)

試験を実施しているウィンドウからフォーカスが外れた（すなわち、他のソフトウェアなどを利用しようとした）ことを検出（JavaScriptで実装）

response-relatedデータ

- **不正行為の検出にも用いられる** (e.g., van der Linden, 2009)
- **測定の精緻化にも用いられている**

  特に解答時間 (Response Time; RT)は長い歴史を持っている(Schnipke & Scrams, 2002)
- **テストの分類**(Gulliksen, 1987)

  power test … 制限時間の無いテスト（解答者が解けるかどうかを試す）

  speed test … 制限時間内にどれだけ解けるかを試すテスト

  → 現実的なテストの大半は**その両側面を持っている** (Hambleton & Swaminathan, 1985)

  → 正誤情報だけを用いるIRTなどはpower test的側面のみに着目

  → **本来は解答時間の情報も使わないといけないのでは？**(van Breukelen, 2005)

response-relatedデータ｜解答時間 (RT) について

統計モデルの分類(van der Linden, 2009)

speed testでは「時間が無限にあれば必ず正解できる」
▶ 誤答は使われないことが多い(Schnipke & Scrams, 2002)

- RTの分布のみを単独で用いるもの ▶ speed test的側面しか見ていない
- 項目反応の情報と組み合わせたもの

▶ 解答者の特性値パラメータ θ とは別に**RTや「解答速度パラメータ」を加えたモデル**が多い

【さらなるモデルの分類】by Tuerlinckx et al. (2019)

1. 特性値 θ 推定のための付加情報として用いる

   例｜階層モデル (van der Linden, 2006; 2007)

2. スコアリングルールなどの要素

   例｜正答時（誤答時）に項目ごとの残り時間が加点（減点）されるという設定に基づくIRTモデル (G. Maris & van der Maas, 2012)

3. 認知プロセスをモデリングする

   例｜Diffusionモデルに基づくIRTモデル(Tuerlinckx & Boeck ,2005; van der Maas et al., 2011)

van der Linden の階層モデル

項目反応 U_{ij} の尤度 ▶ 普通のIRTモデル

$$U_{ij} \sim f(u_{ij}; \theta_j, a_i, b_i, c_i)$$

▲ 二項分布

$$p_i(\theta_j) \equiv c_i + (1 - c_i)\Phi\left[a_i(\theta_j - b_i)\right]$$

▲ 二項分布の正答率パラメータ

RT T_{ij} の尤度

$$T_{ij} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i)$$

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij} - \beta_i + \tau_j)\right]^2\right\}$$

▲ 対数正規分布

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau),$$

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}$$

項目パラメータも
同じように

$$\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta),$$

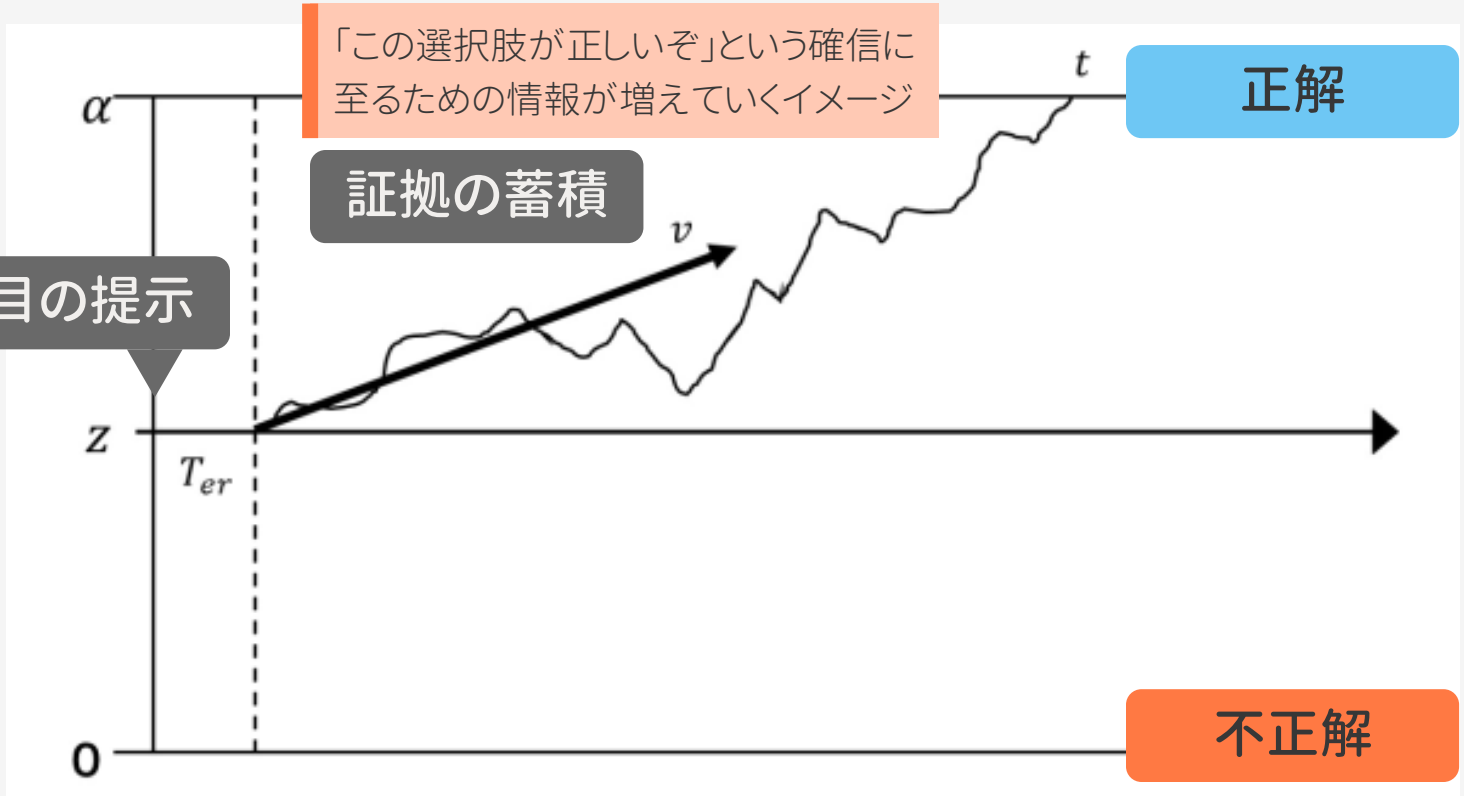
$$\boldsymbol{\Sigma}_I = \left(\begin{array}{ccc|cc} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ac} & \sigma_{bc} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \hline \sigma_{a\alpha} & \sigma_{b\alpha} & \sigma_{c\alpha} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{a\beta} & \sigma_{b\beta} & \sigma_{c\beta} & \sigma_{\alpha\beta} & \sigma_\beta^2 \end{array}\right)$$

フルモデルの尤度

$$f(\mathbf{u_j}, \mathbf{t_j}; \boldsymbol{\xi_j}, \boldsymbol{\psi}) = \prod_{i=1}^{I} f(u_{ij}; \theta_j, a_i, b_i, c_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i)$$

$$f(\mathbf{u}, \mathbf{t}; \boldsymbol{\xi}, \boldsymbol{\psi}) = \prod_{j=1}^{J} \prod_{i=1}^{I} f(\mathbf{u_j}, \mathbf{t_j}; \boldsymbol{\xi_j}, \boldsymbol{\psi}) f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) f(\boldsymbol{\psi}_i; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$$

Diffusionモデル (Ratcliff, 1978)



RT利用時の注意点

RTデータは項目反応データと同じ構造であるべし (van der Linden, 2011)

- 各回答者×各項目のRTを独立に扱う
- 観測されたRTは何かしらの確率分布から発生したものだと考える

データ収集の時点でデザインしておく

One-Item-One-Screen (OIOS) デザイン(Reips, 2021)が良い

▲ Web調査でも推奨されている（文脈効果が軽減できる，脱落のポイントが明確など）

ただ日本のテスト文化では組問が多い（例｜共通テストの数学など）

⟶ 操作ログともうまく組み合わせると良いのかもしれない

process-related (プロセス)データ

Provasnik (2021; 猿田訳, 2021)

テスト問題に取り組む過程(プロセス)を反映した実証的なデータ

解答完了までの過程の把握がメイン

主にTEIsでの解答行動を扱う

プロセスデータの例

キー入力ログ(e.g., Almond et al., 2012; Chan, 2017; Uto et al., 2020)

入力の間隔(短い間を連続入力とみなす)やカーソルの位置(文の途中にある場合は修正、最後尾にある場合は追加とみなす)などの情報を組み合わせて解答者の状態を推定

アイトラッカーなどCBTシステム外のデータ　CBTシステムで取れない情報は含まない派の人もいる

例|Yaneva et al. (2022) ▶ テスト得点の妥当性の証拠としてアイトラッキングを利用している

PISAや国際成人力調査(PIAAC)の複雑な問題解決課題の解答ログ

解答完了までに複数の操作が必要になるため，その状態推移の情報を用いている

プロセスデータ|分析例

Liu et al. (2018)

「各時点でどの線分が選択されているか」の情報を用いて、途中に選択された経路の解答に至るまでの過程の能力 (process-level ability) を評価

▶正解からかけ離れた道を途中で選択した人は能力が低く推定された

正解

Class 1

誤答だがまだマシ

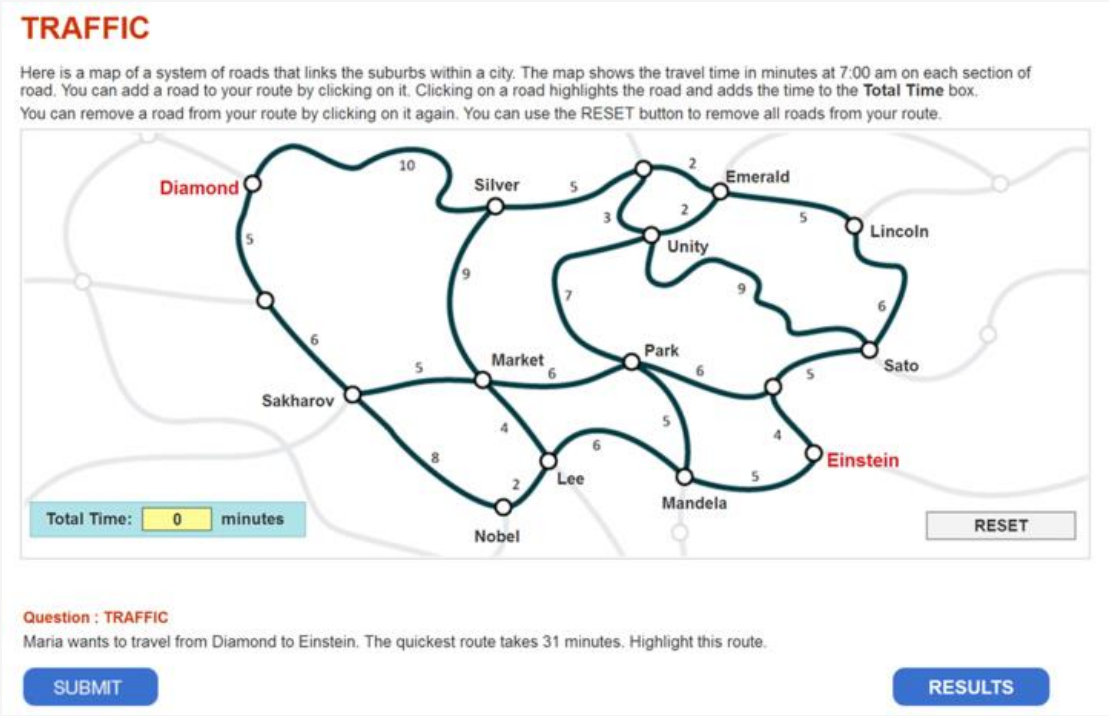
Class 2 & Class 6

【PISA 2012の最短経路問題】

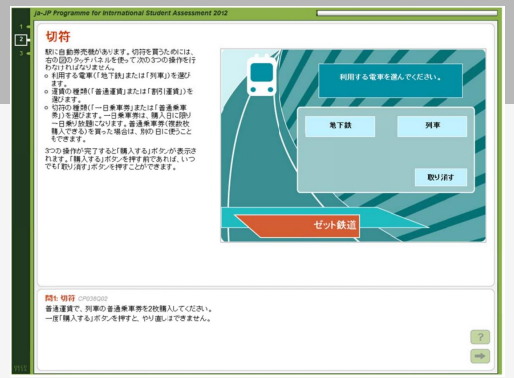
線分をクリックして2地点の最短経路を求める

クリックすると「現在の合計距離」が表示される

▶試行錯誤して正解にたどり着けるか？



プロセスデータ|分析例



Han et al. (2022)「切符の問題」の操作ログを用いた分析

状態遷移を組み込んだIRTモデルによる分析の結果…

状態遷移パターン　正誤　θ の推定値(事後平均)

the marginal p… trib… nd its corresponding response sequences than 100 for th… l re… data.

Response pattern	Frequency	Finish	Mean	Median	Standard deviation	95% Highest posterior density interval
AGHIJK[1]	1609	No	−1.184	−1.142	0.653	(−2.449, 0.099)
AGHIK	4804	No	−1.067	−1.021	0.681	(−2.425, 0.244)
AGHIAGHIK	165	No	−0.764	−0.735	0.503	(−1.783, 0.175)
ABHIJK	604	No	−0.749	−0.733	0.521	(−1.798, 0.263)
ABHIK	855	No	−0.601	−0.592	0.559	(−1.689, 0.519)
ABCIK	1544	No	−0.298	−0.306	0.510	(−1.298, 0.708)
AGABCIK	105	No	−0.198	−0.202	0.449	(−1.132, 0.631)
ABCDFK	457	No	−0.178	−0.186	0.478	(−1.104, 0.78)
AGHIJABCDEK	177	Yes	0.069	0.065	0.376	(−0.635, 0.842)
ABCDK	708	No	0.096	0.072	0.560	(−0.96, 1.245)
AGHIABCDEK	661	Yes	0.236	0.231	0.411	(−0.577, 1.035)
ABCDFEK	329	Yes	0.371	0.348	0.515	(−0.637, 1.383)
ABABCDEK	172	Yes	0.374	0.339	0.511	(−0.562, 1.432)
ABHIABCDEK	152	Yes	0.386	0.370	0.437	(−0.444, 1.271)
AGHABCDEK	287	Yes	0.415	0.404	0.456	(−0.47, 1.305)
ABCDEFEK	259	Yes	0.427	0.400	0.507	(−0.566, 1.44)
ABCABCDEK	140	Yes	0.447	0.413	0.507	(−0.491, 1.501)
ABCDEABCDEK	209	Yes	0.542	0.510	0.478	(−0.395, 1.474)
AGABCDEK	543	Yes	0.591	0.563	0.511	(−0.362, 1.621)
ABCIABCDEK	239	Yes	0.594	0.561	0.488	(−0.321, 1.581)
ABHABCDEK	155	Yes	0.654	0.626	0.511	(−0.31, 1.689)
ABCDEK	10545	Yes	0.903	0.856	0.687	(−0.388, 2.27)

[1]AGHIJK is short for sequence A → G → H → I → J → K, with the arrowswere omitted for brevity. The same below.

同じ正解状態(K)に
到達した人でも
最短で到達した人が
最も θ が高くなる

プロセスデータを利用する場合

RTのように「項目単位」で使うことはあまりない

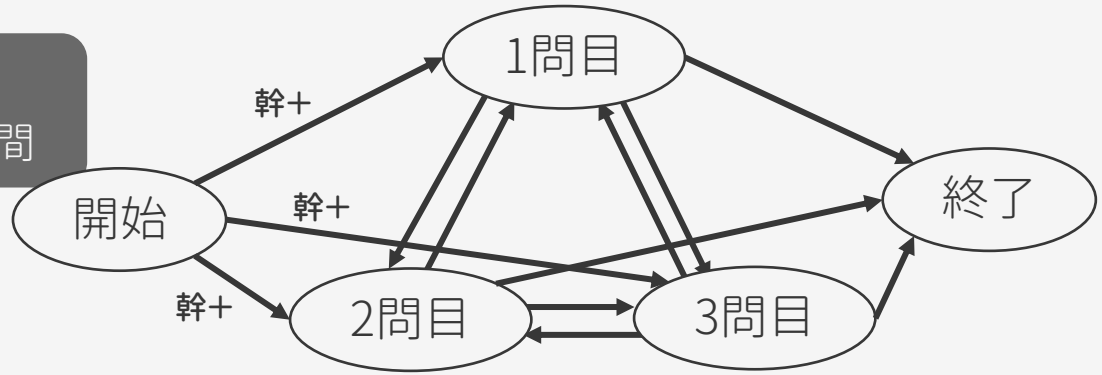
前述の例のように項目内での状態推移を見ることが多い

組問の場合には項目間でも「状態」を考えることがある

例｜3問セットで1画面の場合 (Kroehne & Goldhammer, 2018)

タイムスタンプ
画面オープン
1問目に解答入力
2問目に解答入力
3問目に解答入力
解答提出

×1問目の解答時間
○幹を読む時間＋1問目の解答時間



「意味のある」プロセスデータの再構築が必要(Provasnik, 2021)

テスト設計時から測定したいプロセスを意識して項目を作成する必要がある

Ch. 2目次にもどる

6

特別な配慮

特別な措置の現状

CBTに限らず重要な話

障害者差別解消法の施行 ▶ 試験でも合理的配慮の提供の必要性 (都築, 2018)

2023年度大学入学共通テストでは3,165名に配慮措置が行われた

具体的な配慮としては試験時間の延長，点字や文字での解答など

ではCBTによる支援の形は？(Hansen et al., 2004)

画面の拡大表示，スクリーンリーダー，音声認識，代替キーボード・マウス

拡大表示はCBTの場合自由に倍率を変えられるというメリット

ただし一画面内の情報量が減少する

▶ デバイス操作スキルやワーキングメモリなどはより高水準に要求される可能性(Kamei-Hannan, 2008)

スクリーンリーダーはTTS(Text-to-speech)か可変点字ディスプレイ

複数のメタ分析でLD(学習障害)を持つ受験者に対して中程度の効果(Perelmutter et al., 2017; Wood et al., 2018)

可変点字ディスプレイ Refreshable Braille Display

支援機能の難しさ

PBTのほうが良い?

LD受験者に対する先行研究の多くではPBTのほうが成績が良いという結果(Gelbart, 2018)

支援機能のせいではなく,支援機能への慣れの問題?(Flowers et al., 2011)

CBTシステムはそのCBT用に作られるものが多い

▶ (コンピュータ自体には慣れていても) 日常的にトレーニングする機会があまりない

支援機能の限界の可能性も

スクリーンリーダー向けに構造化されていない素材がある (Hansen et al., 2004)

図表や下線部などの装飾表現は苦手(Kamei-Hannan, 2008)

点字には装飾表現などのために拡張された8点点字がある(Argyropoulos et al., 2020)

▶ 日常的に使われているものではないので練習が必要

支援機能の難しさ

支援機能自体がテストの性質に与える影響も要検討

合理的配慮の本質的な目的

▶ テストが測定したい構成概念とは無関係な障壁 (construct-irrelevant barriers) の除去

合理的配慮がテストの構成概念を変容させる例

- スクリーンリーダー向けに構造化できる範囲内での作問を行う
- 数学のテストで計算障害 (dyscalculia) を持つ受験者に電卓を用意

▶ 受験者は単純計算能力の有無を問われなくなり，数学的問題解決能力のみを問われる

もしこのテストが単純計算能力を含む数学力を測定しようとしていたら，電卓は不適切?

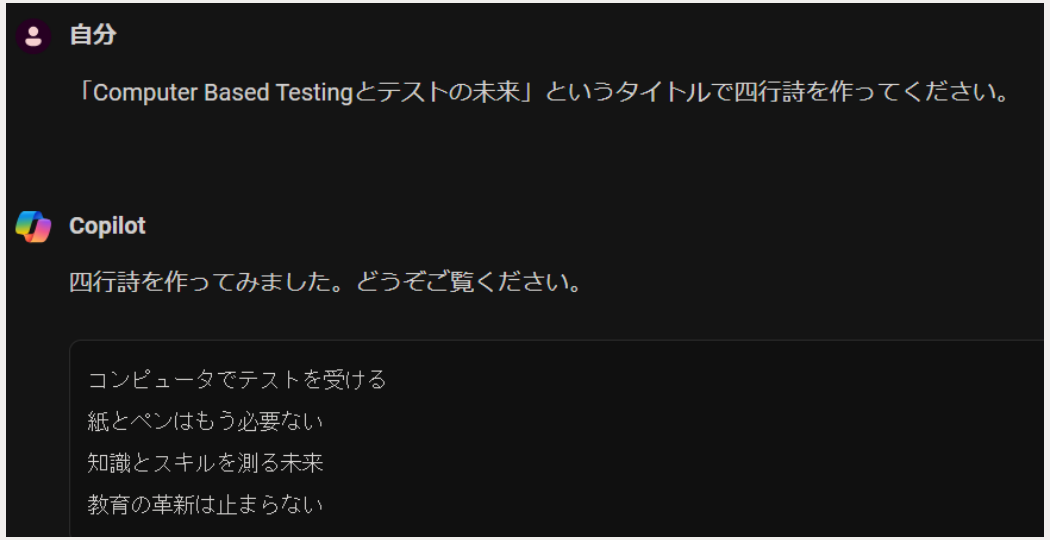
Construct / Predictive Equivalence

合理的配慮によるモード効果の検証もしたほうが良い

その際はPBTとの比較のときと同じように (e.g., Flowers et al., 2011; Kamei-Hannan, 2008)

目次にもどる

3
CBT研究の未来（やや妄想）



- CBTによって「テスト」はどう進化していくか

Why CBT?

CBTを作るのは大変なことです

「なんか良さそう」で作ると痛い目を見るかもしれない

② 適応型テスト ▶ 項目を大量に作り続けるだけの体力が必要

③ 新しい形式の項目 ▶ 作成から評価まで，精緻な計画が必要

④ 不正対策 ▶ 最新技術まで視野に入れて，対策の継続的な更新が必要

一方でCBTには夢があります

いろいろな意味で，**テストを「より良い」ものにしてくれる**技術（だと思っています）

CBTの未来を「良さ」から考えてみます

「より良い」CBTを目指して

テストの目的は「(人の)能力などをうまく測定すること」

1 測定および評価の方法の妥当性 (validity) 真正性 (authenticity) をもっと高めたい

TEIsは更に進化する

妥当性の検証フレームワークの確立

より真正性の高いTEIs開発の方法論

TEIsによる妥当な・透明性のある評価

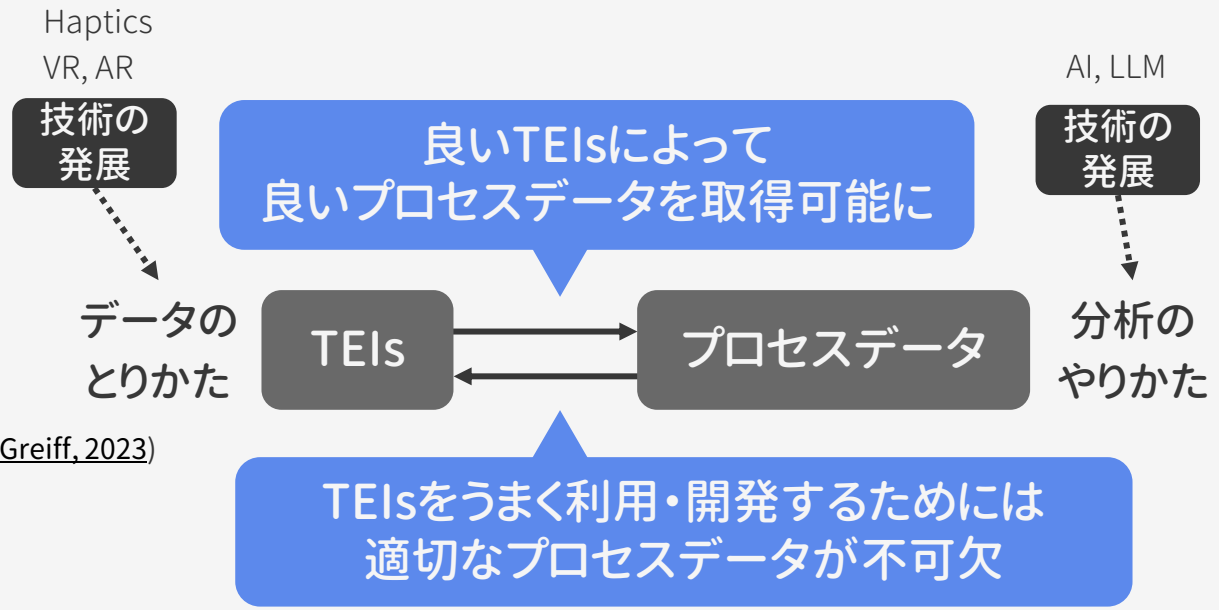
プロセスデータ利用は発展途上 (e.g., Linder & Greiff, 2023)

プロセスデータによる評価の妥当性

プロセスデータで何ができるか

倫理的な側面についての議論

何でもかんでもデータを収集するのは
プライバシーの問題につながる



「より妥当な」CBTを目指して

テストの目的は「(人の)能力などをうまく測定すること」

② 測定したいものと無関係 (construct-irrelevant) なものを取り除きたい

WBTは地域など受験機会の格差解消に役立つかも

例｜大学入学共通テストの受験会場 (朝日新聞, 2024/3/8閲覧)

> 愛媛、佐賀、熊本の3県は試験会場が県庁所在地にしかない。

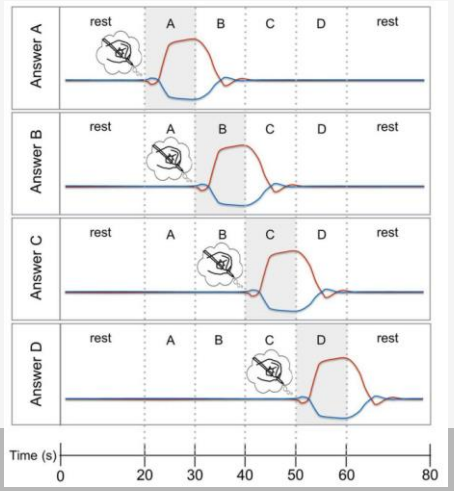
> 県最南端の愛南町にある県立南宇和高校の生徒たちが、車で約2時間半かかる100キロ超も離れた会場に向かう姿は毎年恒例で、修学旅行さながらだ。

技術革新によって特別な配慮も変わるかも

例｜Brain-Computer interface (e.g., Nagels-Coune, 2021)

▶「書く動き」を想像した際の脳波を読み取って解答できたり

Neuralink社がすでに臨床試験に向けて動いている



海外ではWBTが公的に利用され始めている

カナダの医師を目指す人が受ける試験の一つ

Medical Council of Canada Qualifying Examination (MCCQE)

2021年よりリモート試験監督を利用した受験が本格始動

もともとはCOVID-19対策として導入されたようだが，**成果をあげたためそのまま継続**

アメリカの産科などの看護師への国家認定プログラムの試験

National Certification Corporationが実施する試験

24時間いつでも自宅のコンピュータから受験可能らしい

ただし事前に手持ちのPCの
Compatibility testが必要

ただしテスト文化の違いもあるので日本ではなかなか厳しいのかも?

特に大学入試などは一斉受験が根強いので，根底から考え方を変える必要がありそう

「完全に自動化された」CBTに向けて

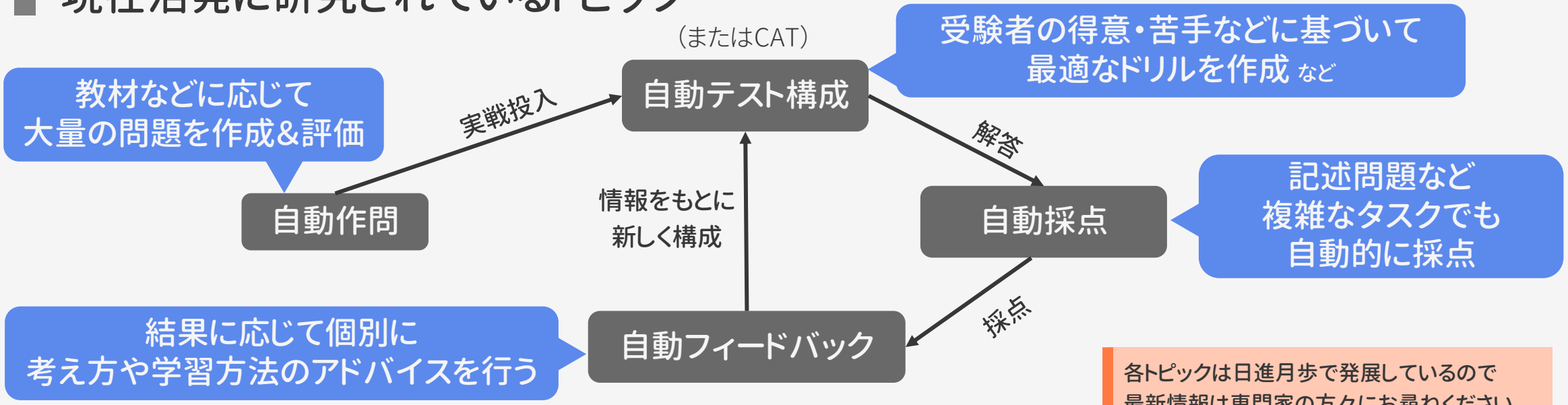
テストの目的は「(人の)能力などをうまく測定すること」だった

➡ 近年では「評価のためのテスト」から「学習の中でのテスト」へ (加藤, 2021)

3 テストをもっと自由に実施・受験できるようにしたい

▶ テスト運用に関する様々なコストを軽減するために，自動化の技術は必要不可欠

現在活発に研究されているトピック



各トピックは日進月歩で発展しているので
最新情報は専門家の方々にお尋ねください。

自動化における生成AIの動き

どんどん研究が進んでいる

1年後，半年後にはどうなっているのでしょうか…？

自動作問
(e.g., Sayin & Gierl, 2024)

Example Prompts

Text generation parameters

prompt = (

"This is an informative text generator."

" This text is four sentences, and each sentence should generate 10 different sentences according to the prompt below. An example text is given below:"

" Sample text: (1.) If there is even the slightest possibility, wouldn't you like to start life all over again, right where you left off? (2.) Enthusiasts of science fiction narratives such as 'Frankenstein' contemplated the potential revival of the human body. (3.) A Turkish scientist who shared this curiosity became one of the 10 researchers who made the best discovery by resurrecting dead nerve cells in the brain in a laboratory at Northwestern University laboratory. (4.) This study, which lasted 60 days and succeeded in bringing dead brain cells to the level of healthy neurons, is a pioneering work for the medical world."

" Prompt for the first sentence: Generate 10 sentences similar to the first sentences. Each sentence should include: A <expression_1> sentence related to <idea_relevant_1>. Make it <structure_1> for the reader. It should have <word count_1> words and a readability score between <readability_1>."

" Prompt for the second sentence: Generate 10 sentences similar to the first sentences. Each sentence should include: A <expression_2> sentence related to <idea_relevant_2>. Make it <structure_2> for the reader. It should have <word count_2> words and a readability score between <readability_2>. Ensure the generated sentence is consistent and maintains continuity with the first sentence."

" And provide me with the sentences in this format. Text:\nPurpose:\nStructure:\nExpression:\nWord Count:\nCount of Sentences:\nReadability:"

" For each sentence, choose five sentences that best match the generation features?"

)

小論文の自動採点
(e.g., Mizumoto & Eguchi, 2023)

I would like you to mark an essay written by English as a foreign language (EFL) learners. Each essay is assigned a rating of 0 to 9, with 9 being the highest and 0 the lowest. You don't have to explain why you assign that specific score. Just report a score only. The essay is scored based on the following rubric.

[*IELTS rubric in a plain text format.*]

ESSAY:

[*Inserting each of the 12,100 essays using a for loop in Python code.*]

フィードバックの効果
(e.g., Escalante et al, 2023)

Appendix B

Sample prompt sent to GPT-4 to generate feedback

You will be a professional language teacher who is an expert on providing feedback on the writing of English language learners. Here is the writing prompt that students are given: [the weekly writing prompt was inserted here].

Below I will share with you a student's writing. Based on their writing, comment on the following:

1. Using simple language, comment on the quality of the topic sentence and if it addresses the writing prompt. Provide suggestions for improvement but don't write a new topic sentence for the student. Provide an example of an improved topic sentence that is about a different topic than the student's writing. Start your feedback with the header in bold "Feedback on the quality of the topic sentence:".

4. Using simple language, comment on the development of ideas throughout the paragraph. Specifically comment on the development of the main idea through supporting ideas and elaborating details such as examples and evidence. Start your feedback with the header in bold "Feedback on the development of ideas throughout the paragraph:".

5. Using simple language, identify language in the paragraph that lowers the para

人間によるフィードバックと
学習効果に有意差なし

おわり

※特に後半の内容は数年後には古くなっているかもしれません。
(2024/03/16)