教育測定におけるAI技術

宮澤　芳光

独立行政法人大学入試センター

本講演の内容

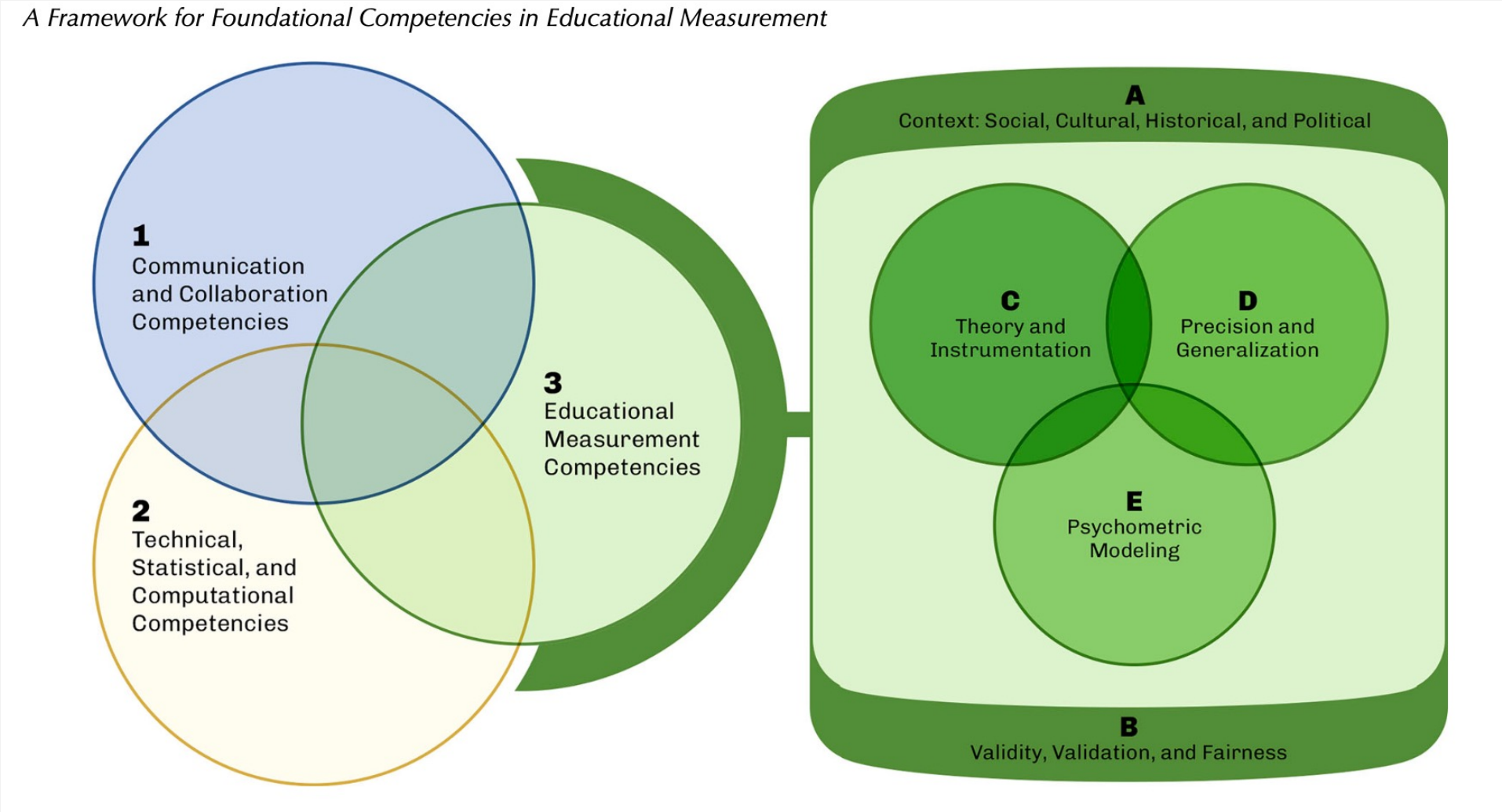
以下の内容を紹介

- 教育測定における基盤的なコンピテンシー
  - Ackerman, T. A. et al. Foundational Competencies in Educational Measurement. Educ. Meas.: Issues Pr. 43, 7–17 (2024).
- 深層強化学習を用いた適応型テスト
  - Wang, P., Liu, H. & Xu, M. An adaptive testing item selection strategy via a deep reinforcement learning approach. Behav. Res. Methods 1–20 (2024) doi:10.3758/s13428-024-02498-x.

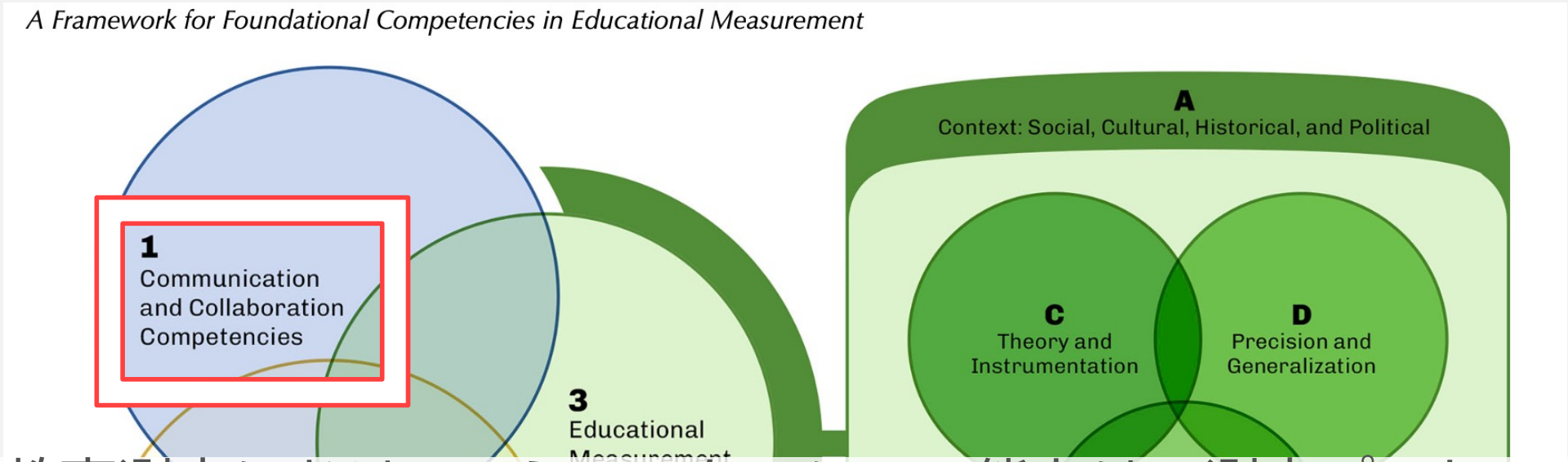
教育測定における基盤的なコンピテンシー

- 論文情報
  - Ackerman, T. A. et al. Foundational Competencies in Educational Measurement. Educ. Meas.: Issues Pr. 43, 7–17 (2024).
- 背景
  - 「教育測定における基盤的なコンピテンシーとは何か」を明らかにする
  - コンピテンシーとは，研究や実践に携わる専門家に必要とされる，知識，スキル，能力，および行動の総体を指す。コンピテンシーを構成する具体的な知識，スキル，能力，行動に関する専門性の程度は，個々の専門家によって異なる。なお，以降では，学習者をこれらのコンピテンシーを高める過程にある学生や専門家とする。
- プロセス
  - タスクフォースメンバー（12名の教育測定専門家）による議論に加え，NCME(全米教育測定評議会)会員からのフィードバックを反映
  - 学会発表や論文の公開を通じて広く意見を募り，学術コミュニティの合意形成を経てフレームワークを確立
  - 既存の教育測定プログラムや実務の要求事項と照らし合わせながら，適用可能性を評価

教育測定におけるコンピテンシーの枠組み



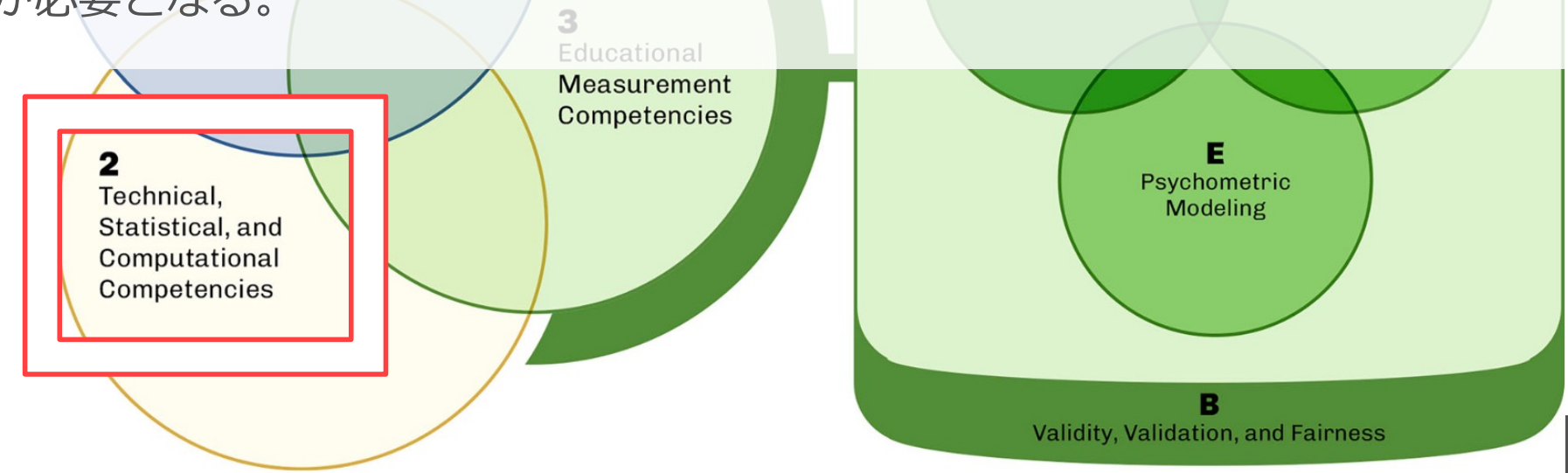
コミュニケーションと協働のコンピテンシー



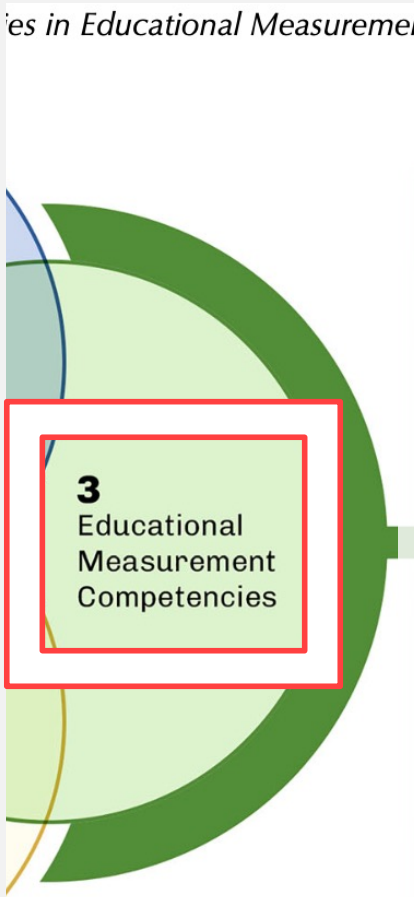
教育測定におけるコミュニケーション能力は，測定プロセスの説明や分析結果の提示，分析内容の解釈を多様な聴衆に伝える能力を指す。協働能力は，教育，心理学，技術開発などの専門家と連携し，測定ツールの設計や評価を行うスキルを含む。効果的な測定には協働が不可欠であり，異分野の専門家と連携するための適切な情報伝達が求められる。実践的な学習機会を通じ，発表やデザイン制作などを経験し，これらの能力を養うことが重要である。

技術的・統計的・計算機的コンピテンシー

技術的（Technical）・統計的（Statistical）・計算機的（Computational）コンピテンシーには，サンプリング理論と方法，探索的データ分析，パラメータ推定の計算手法，多層モデリング，ベイズ統計，因果推論のための実験・準実験手法など，多様な統計および研究手法が含まれる。技術的スキルには，統計ソフトウェアを用いたデータの管理・変換，シミュレーションの設計と実施，レポート作成，仮説の事前登録および検証が含まれる。計算機的スキルには，ソフトウェアコードやプログラムの作成，アルゴリズムの論理と目的の理解が含まれる。測定は通常，数値として表され，不確実性の推定が伴う。これらの推定は確率モデルを用いて形式化される。多くの測定が大規模に行われるため，統計・測定の実践には，R，Python，SASなどの統計ソフトウェア環境に習熟していることが求められる。テクノロジーの進化，データアクセスの拡大，人工知能（AI）や自然言語処理（NLP）の発展により，計算機的コンピテンシーの重要性はますます高まっている。特に，多くの試験がデジタル化される中で，受検者とテスト項目の相互作用を監視・記録・自動採点する評価環境を設計するスキルが必要となる。



教育測定のコンピテンシー



教育測定のコンピテンシー には，以下の5つのサブドメインが含まれる。

(A)測定が行われる社会的，文化的，歴史的，政治的文脈などに関連する包括的なサブドメイン

　測定のプロセス，分析，報告に影響を与える枠組みを提供するため

　「包括的（overarching）」な要素

(B)妥当性，妥当化，公正性に関連する基盤的なサブドメイン

　測定活動の動機づけ，評価，改善の基盤となるため

　「基盤的（undergirding）」な要素

(C)測定理論と計測技法

(D)精度と一般化

(E)心理測定モデリング

サブドメインのコンピテンシーは，教育測定の一般的な取り組みを支えるものであり，測定ツールの設計・開発，反応の採点，スコア精度の推定と報告，パフォーマンス基準の確立，スコアの比較可能性を確保するためのスケーリングや等化手法の適用などが含まれる。教育測定の専門家は，これらのコンピテンシーを活用し，スコアの妥当性，信頼性，公正性を評価・向上させることを目的としている。

コンテキスト：社会的，文化的，歴史的，政治的

コンテキストに関するコンピテンシー は，教育測定の学習者や専門家が，測定に関連する社会的・文化的・歴史的・政治的な要因を理解し，適切に活用する責任があることを強調している。

1. 社会的コンテキスト（Social Context）

受検者が置かれた社会的構造は，その機会，期待，規範に影響を与え，測定ツールとの相互作用にも影響を及ぼす。人種，民族，性別，階級，言語，障害などによって形成される社会的階層（学校，教室，家庭，職業，地域社会など）が関係する。

2. 文化的コンテキスト（Cultural Context）

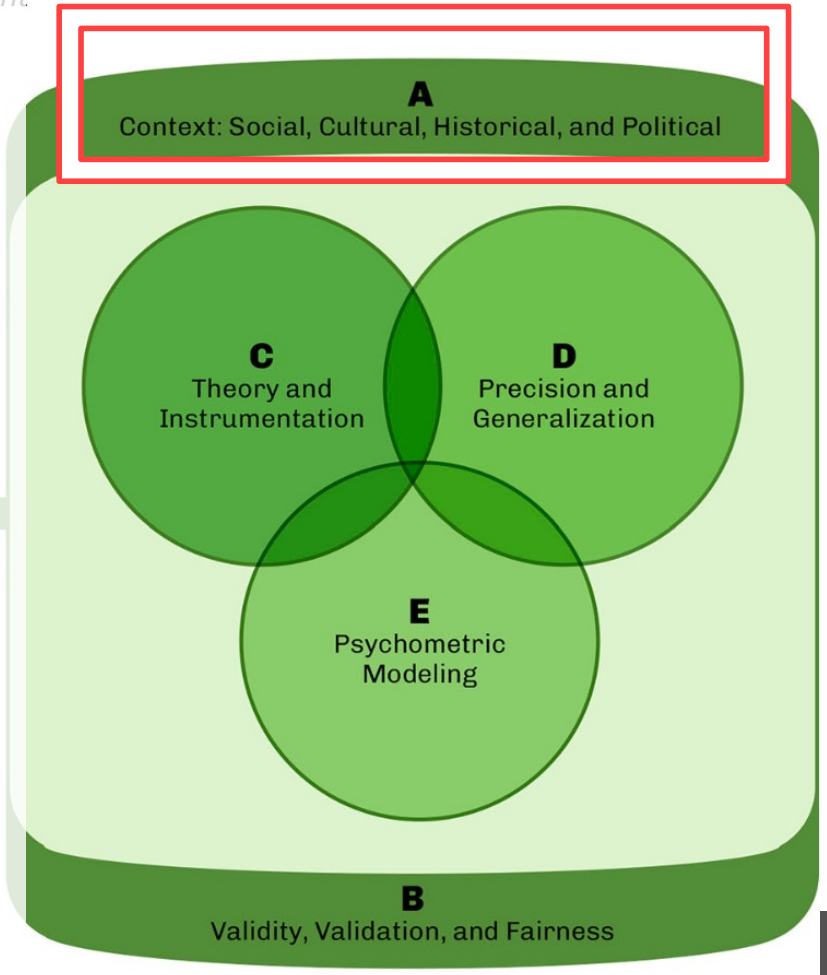
受検者が生活し学ぶ文化は，知識の捉え方，コミュニケーション方法，相互作用の仕方に影響を与え，価値観や世界観を形成する。これらは，測定ツールとの相互作用やスコアの解釈にも影響を及ぼす。

3. 歴史的コンテキスト（Historical Context）

受検者の過去の測定経験や，ある概念に関する信念は，その後の測定ツールとの相互作用に影響を与える。特定の社会集団の測定に関する歴史（例：知能テストの誤用による人種差別的政策の正当化）も，受検者の測定への関与度に影響を与える。したがって，教育測定の設計・実施・報告において，この歴史的背景を考慮することが重要である。

4. 政治的コンテキスト（Political Context）

教育測定は，教育システムのさまざまなレベルで複数の政治的目的を果たすことができる。この政治的文脈は，測定の開発・利用に対して肯定的または否定的な影響を及ぼす可能性がある。



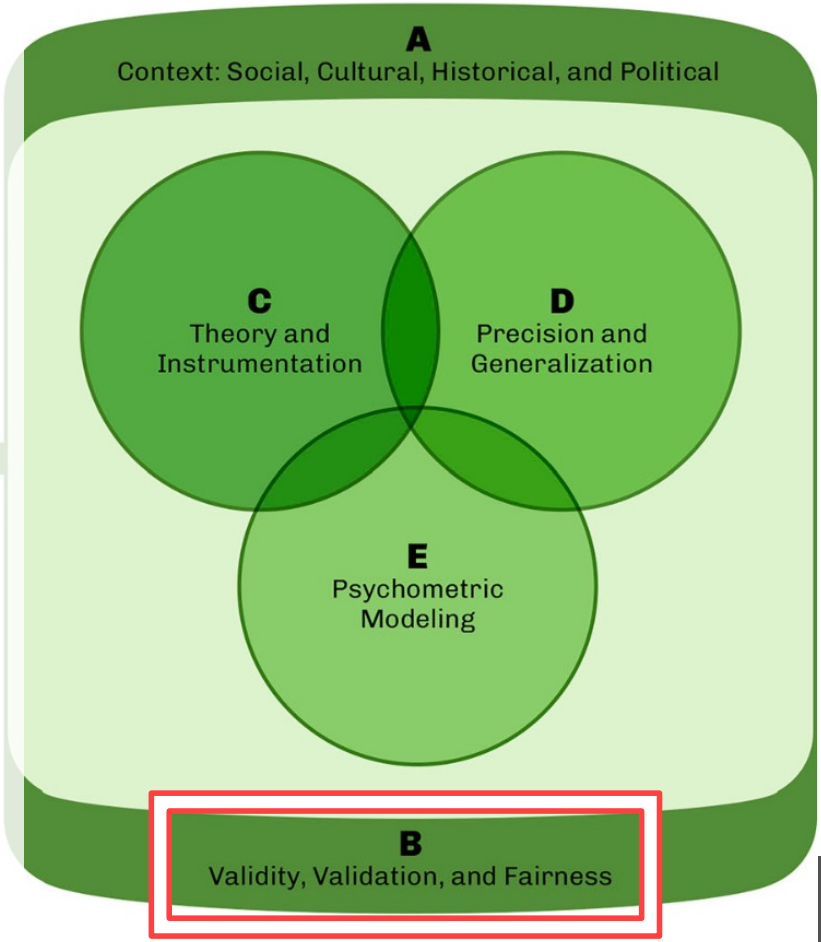
妥当性（Validity），妥当化（Validation/妥当性の確認），公正性（Fairness）

- このコンピテンシーは，学習者がテストスコアの解釈や利用目的を明確にし，それを支える理論や証拠を生成・評価する能力に関連する。妥当性（Validity）は，テストスコアの解釈および使用における最も基本的な考慮事項であり，教育測定の研究者や実務者の主要な取り組みの一つは，妥当性の証拠を生成し，それを評価することである。この妥当性の証拠を構築・検証する活動は妥当化（Validation）と呼ばれる。妥当化には，教育測定のスコアの利用や解釈が，個人や異なるグループに対して公正であり，理論や証拠に基づいて正当化されているかを評価することも含まれる。つまり，妥当性だけでなく，公正性（Fairness）を確保することも，妥当化の重要な要素である。

- 専門的な知識を持つ学習者は，異なる目的に応じて，妥当性と公正性を確保するために複数の証拠源がどのように役立つのかを説明することができる。たとえば，以下のようなケースにおいて，どの証拠が妥当性を支えるのかを適切に説明できる：

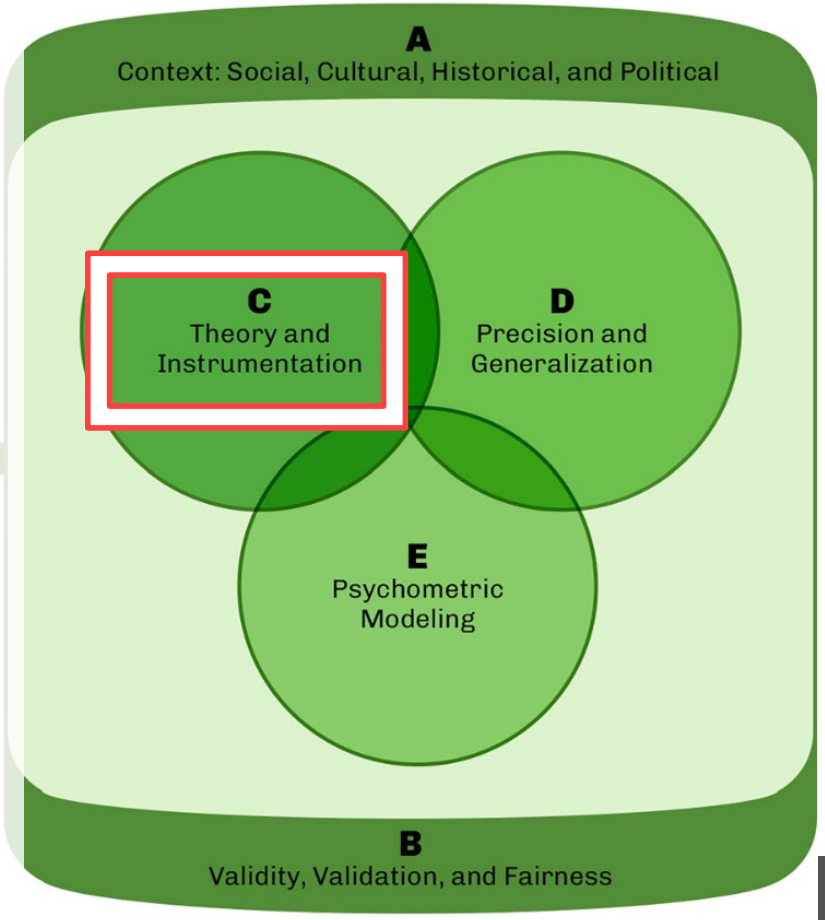
1. 教育アカウンタビリティ（責任検査）においてテストスコアを利用する際，テストの内容と教育カリキュラムの整合性（Content Alignment）が妥当な証拠となるか
2. 大学入試においてテストスコアを利用する際，スコアと大学の成績との相関（Correlations between test scores and college grades）が妥当な証拠となるか
3. 診断スクリーニング（診断的評価）においてスコアを利用する際，内的整合性（Internal Consistency）が妥当な証拠となるか

- また，このコンピテンシーを持つ学習者は，スコアの解釈や使用の妥当性をより強化する追加の証拠を特定し，異なるグループ間でスコアの使用が公正であることを示す証拠を探求することもできる。



理論と計測技法（Theory and Instrumentation）

教育分野である構成概念（construct）を測定するには、まずその構成概念や学習プロセスについての理論が必要である。これらの理論は測定ツールの開発や妥当性検証の指針となり、測定ツールが構成概念のレベルを適切に区別し、意図したスコア解釈や活用を支える証拠を集める役割を果たす。具体的には、測定ツール設計や開発プロセスの理解、デジタルや計算技術を活用した項目生成や採点などにおいて重要である。この分野の専門性を持つ学習者は、学習理論や認知モデルを活用して項目やタスクを設計し、内容適切性やバイアス・アクセシビリティ等のガイドラインに従って測定ツールを開発でき、さらにこれらを理論に基づき評価・改善できる。専門性を養うためには、原理に基づくテスト開発や評価方法、既存テストの仕様書や採点方法の批判的検討を通じた実践的な経験が重要である。

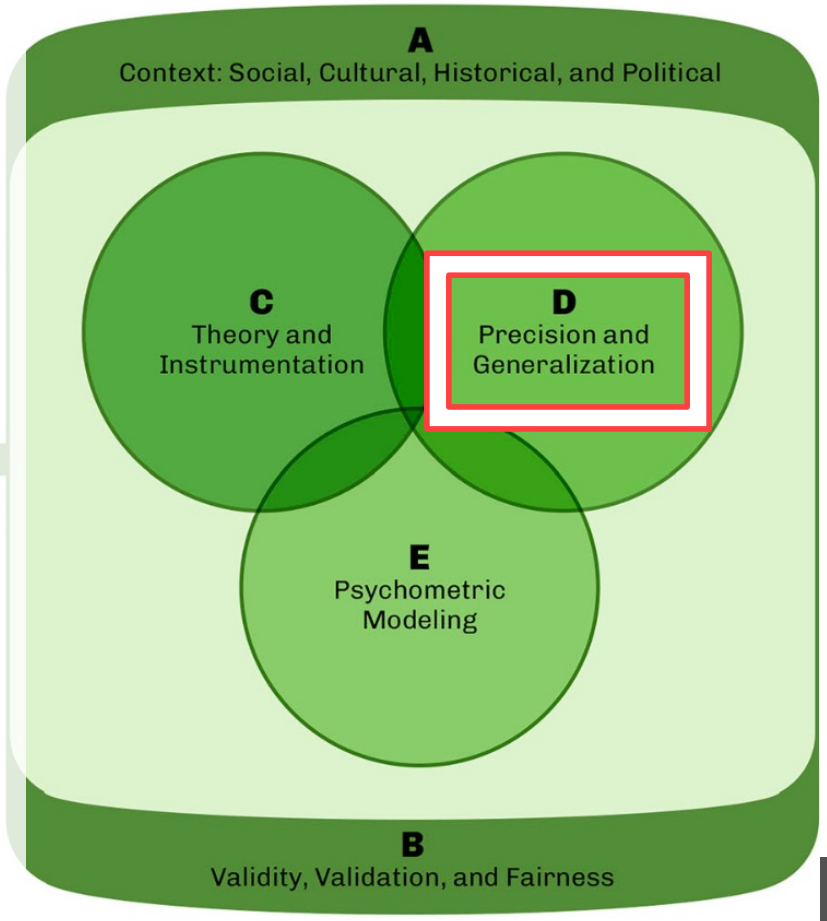


精度と一般化（Precision and Generalization）

このサブドメインに精通した学習者は，テストスコアの一般化の範囲を明確にし，対応する精度の指標を推定・解釈できる。一般化の対象として，他の項目，評価者，実施機会，集約スコアなどを特定し，それに対応する信頼性係数や誤差推定を示すことが求められる。また，一般化の限界を理解し，それを補強するための研究設計を構築する能力も必要とされる。

スコアの特性や統合方法（例：変換，平均化，差分計算）が精度と一般化に与える影響を理解し，適切なスケーリングや等化手法を用いることで，スコアの比較可能性や解釈の一貫性を確保することが重要である。

このコンピテンシーを発展させるためには，信頼性と測定誤差の概念を統計モデルと区別し，異なる信頼性係数や標準誤差を推定・解釈するスキルを習得することが求められる。これにより，異なるグループや熟達度レベルに応じた測定誤差の影響を考慮し，精度の高い測定と適切な情報伝達が可能となる。



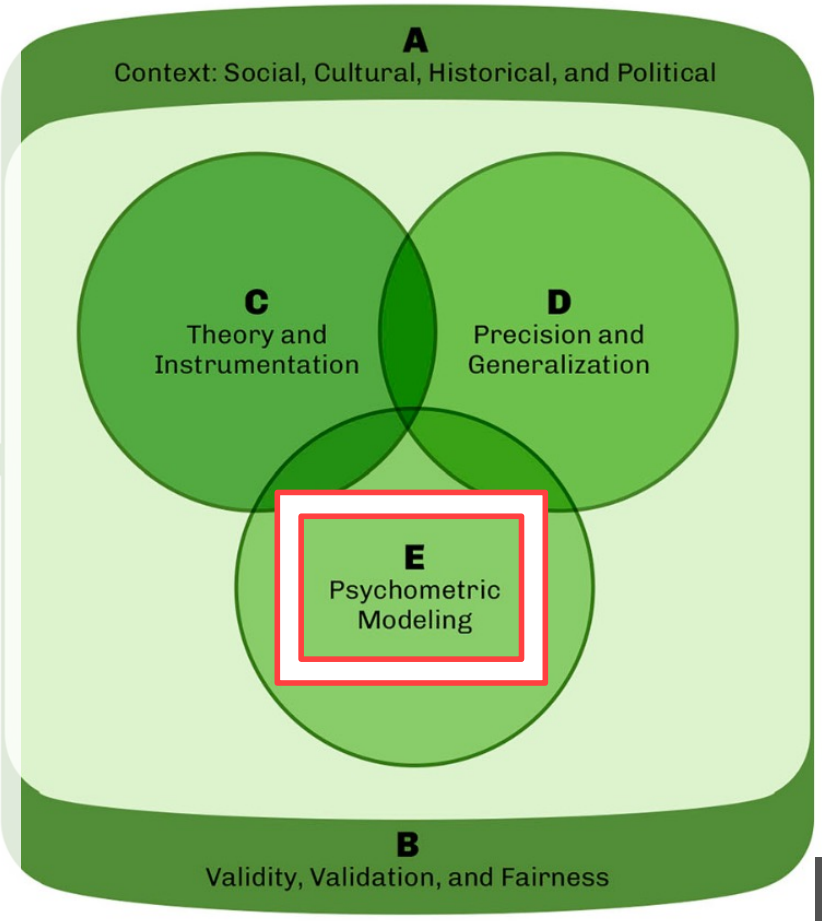
心理測定モデリング（Psychometric Modeling）

心理測定モデルは，測定ツールの設計・開発に不可欠であり，精度，不確実性，信頼性，一般化可能性，不変性，比較可能性などの概念を評価するための重要な手段である。これらのモデルは，測定対象となる構成概念，テスト項目の特性，外部変数との関係を調査するために用いられる。

このコンピテンシーに精通した学習者は，さまざまな統計モデルや心理測定モデルを選択・適合・評価・解釈する能力を持ち，古典的テスト理論（CTT），項目反応理論（IRT），因子分析モデルの違いや前提条件を理解できる。

大規模な教育測定では，スコア解釈のために適切な心理測定モデルを選択し，モデル診断やパラメータ推定を用いてスコアの解釈と活用を向上させることが求められる。また，これらのモデルは混合効果モデルや人工知能（AI），自然言語処理（NLP）などの計算手法とも補完的に利用される。

この分野の専門性を高めるには，心理測定モデリングの基礎を学び，CTTとIRTの相互関係を理解し，テスト項目の評価やスコアリングに適したモデルを選択・評価する能力を身につけることが重要である。



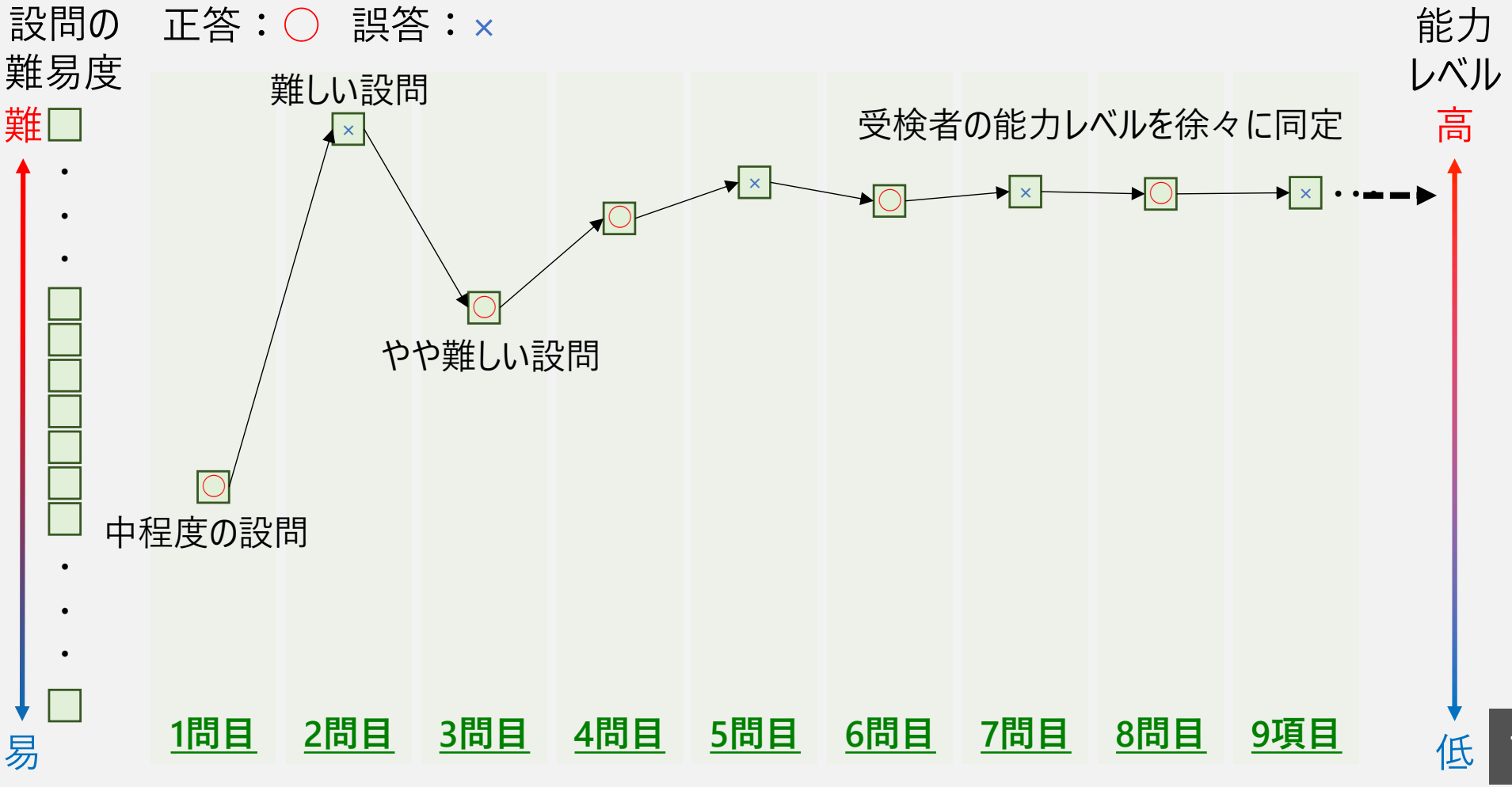
深層強化学習を用いた適応型テスト

適応型テスト

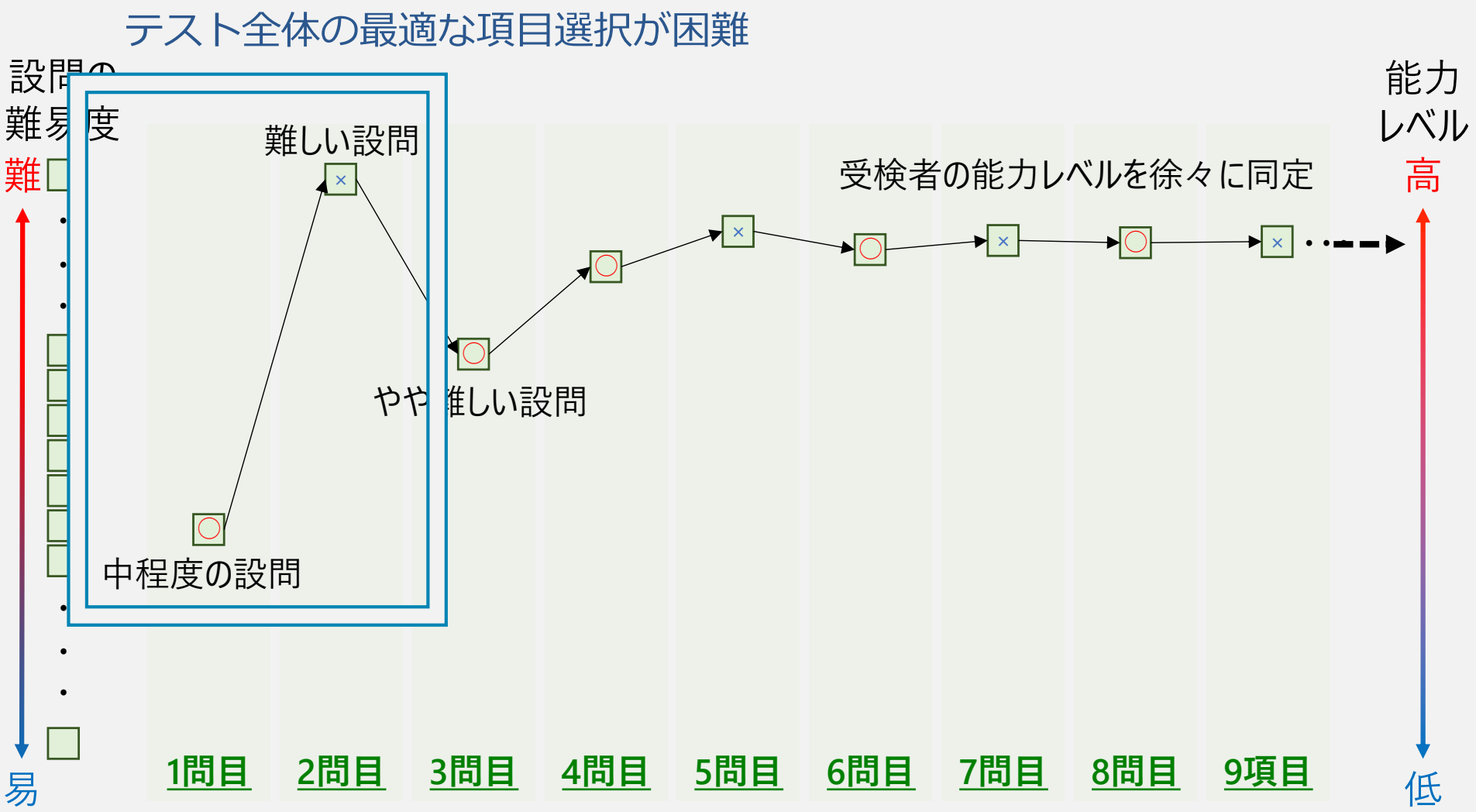
特徴　正答するとより難しい問題を，誤答するとより易しい問題を出題する。

メリット　多様な能力水準の受検者に対応
得点に付随する誤差の低減

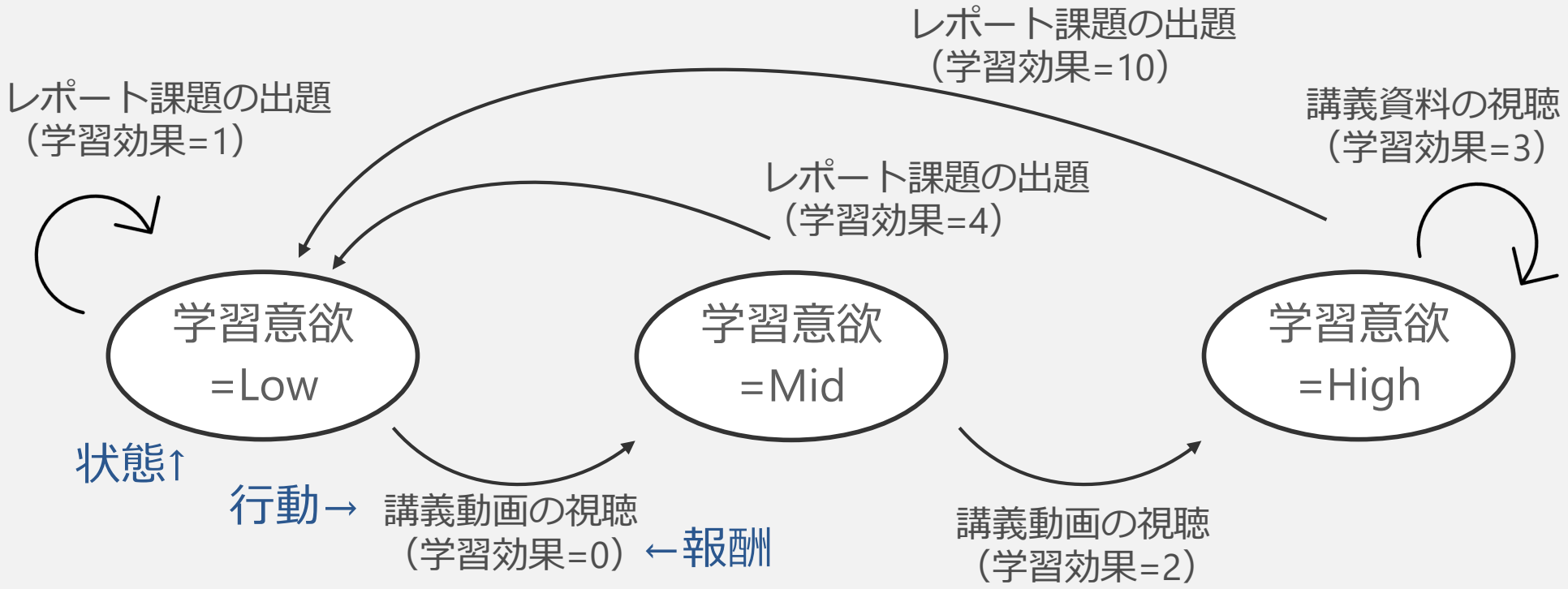
デメリット　自分の解答見直し不可
問題セットの事前配信不可



適応型テストの課題の一つ



逐次的意思決定問題
例：学習の最適化（仮想的な内容）



以前の状態や行動を所与として，次の行動を決定　←方策

A)　レポート課題を出題し続ける。平均学習効果は約1

B)　講義動画を視聴し続ける。平均学習効果は約3

C)　学習意欲=Midでレポート課題を出題する。平均学習効果は約2

D)　学習意欲=Highでレポート課題を出題する。平均学習効果は約4

Mykel J. Kochenderfer 著・ 繁桝 算男 監訳・本村 陽一 監訳，不確定性下の意思決定，共立出版，2020

マルコフ決定過程

マルコフ決定過程とは

　時間tのある状態s_tに基づき，行動a_tを選択して報酬r_tを受け取る

　次の状態は，現在の状態と行動に基づき確率的に遷移（マルコフ性）

マルコフ性とは

現時間ステップtの値が$t-1$の値にのみ依存

サイコロを振った 時間ステップt	1	2	3	4	5	6	7	8	9	10
(a) t回目に出たサイコロの目	3	1	2	2	4	1	6	3	4	2
(b) t回目までの最大値	3	3	3	3	4	4	6	6	6	6
(c) t回目までの中央値	3	2	2	2	2	2	2	2.5	3	2.5

(a)独立同一分布

(b)マルコフ性を満たす

(c)マルコフ性を満たさず，また，独立同一分布でもない

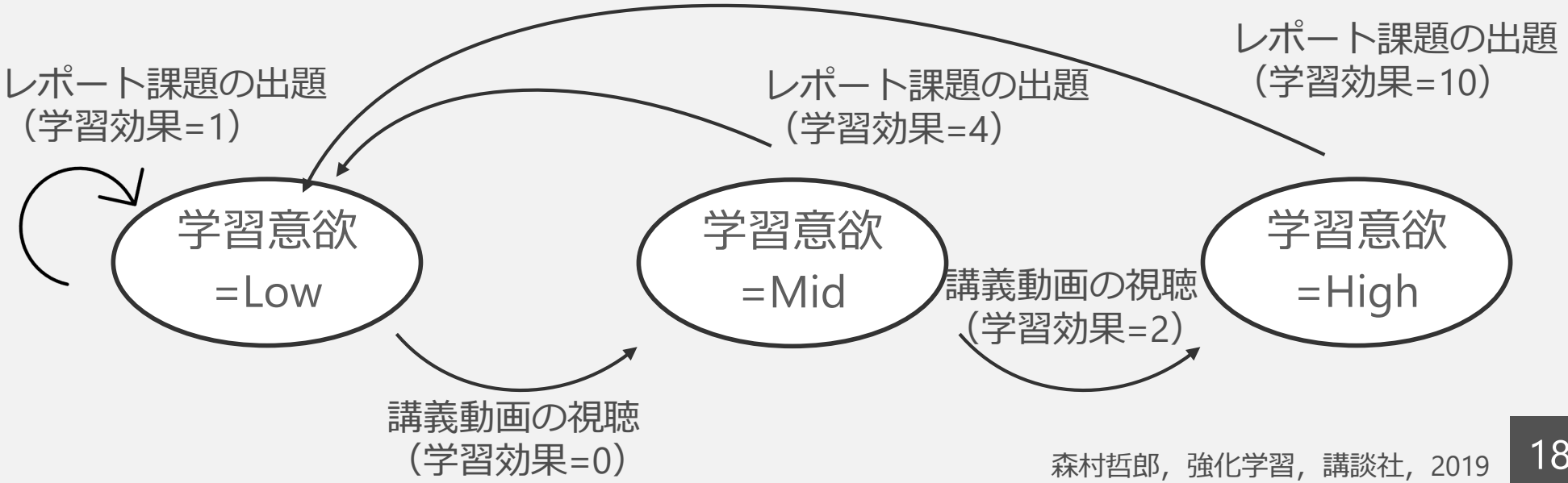
Mykel J. Kochenderfer 著・ 繁桝 算男 監訳・本村 陽一 監訳，不確定性下の意思決定，共立出版， 2020
森村哲郎，強化学習，講談社，2019

逐次的意思決定問題の分類

逐次的意思決定問題

① プランニング問題（環境モデルが既知）
- 学習の個別最適化の例
- 動的計画法を用いて最適解を探索 (後述)

② 強化学習問題（環境モデルが未知，データから学習）
- 例えば，学習の個別最適化において学習効果等が未知
- 環境モデルの探査と知識利用の調整 (後述)



森村哲郎，強化学習，講談社，2019

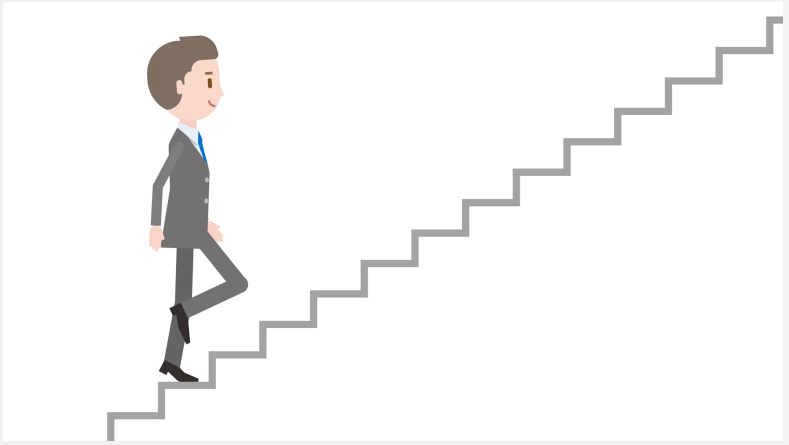
プランニング問題を解く手法の一つ
動的計画法

動的計画法とは

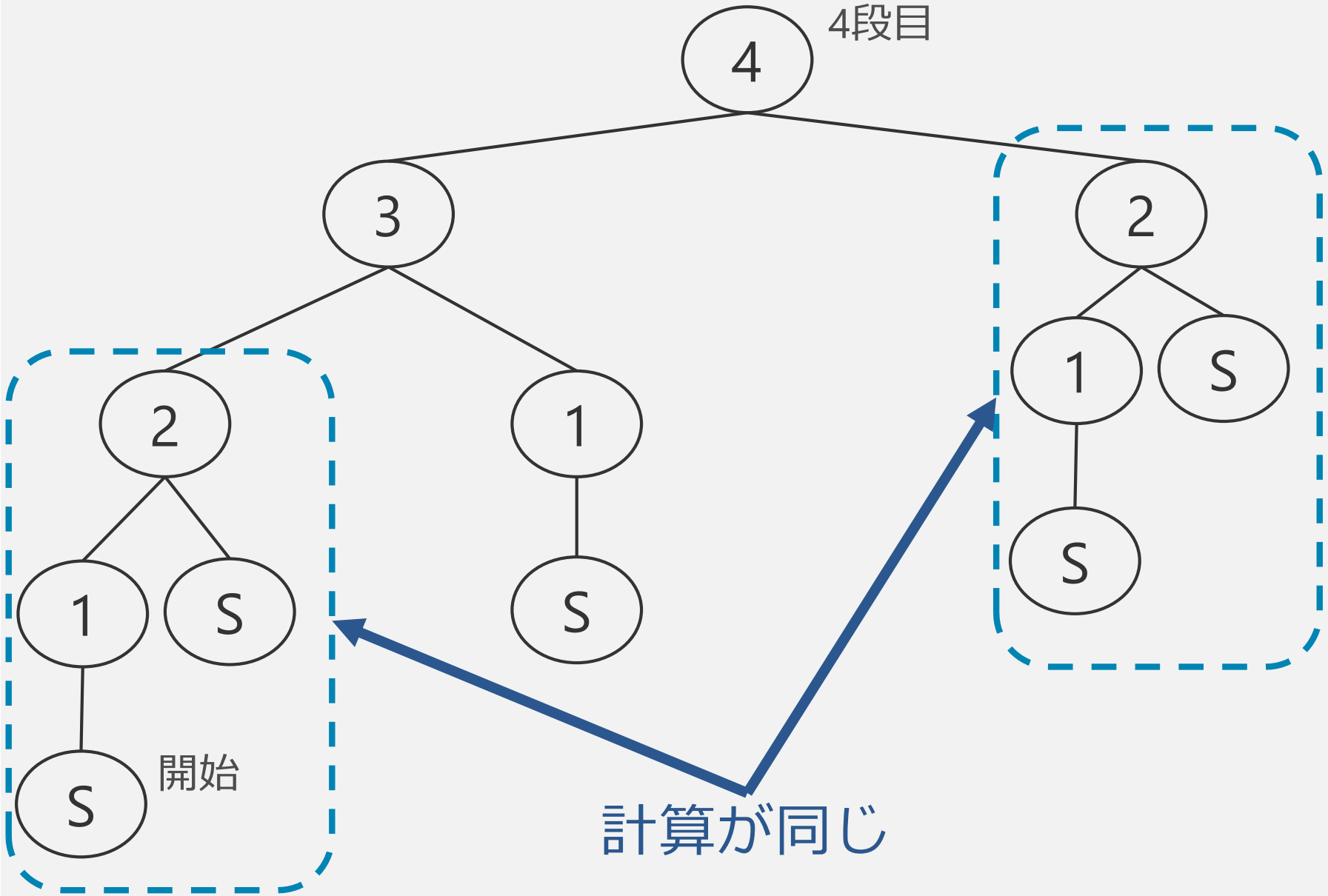
問題を小さな部分問題に分割し、それらを再利用することで効率的に解を求める方法

例：階段を1回で1段上るか，2段上るかを選べるとして「10段の階段をのぼる方法は何通りか？」

- 10段目にたどりつく方法は 「9段目から1だん上がる」 か 「8段目から2だん上がる」 のどちらか
- 「10段目にたどりつく方法の数」＝「9段目までの方法の数」＋「8段目までの方法の数」
- 9段目や8段目にたどりつく方法も同様に考える
- どんどん小さな問題にして、既に求めた答えを再利用して計算を効率化する



4段の階段をのぼる方法は何通りか？



環境の探査と知識利用の調整

- 多腕バンディット問題
  - 当選確率が未知の複数台のスロットマシン
  - スロットのレバーを引く回数：N回
  - 配当を最大化する方策を検討
- 方策
  - (a)ランダムにスロットを一つ選び，N回レバーを引く
  - (b)3台のうち各スロットを$\frac{N}{6}$ずつレバーを引き，← 環境の探査

    当選回数が最大のスロットを$\frac{N}{2}$回レバーを引く　← 知識利用
  - (c)確率εでランダムにスロットを選び，

    確率$(1-\varepsilon)$で当選回数が最大のスロットを引く　← ε-greedy（イプシロン貪欲）法
  - (d)レバーを引くごとにランダムにスロットを選択

スロットA
当選確率：0.1

スロットB
当選確率：0.5

スロットC
当選確率：0.8

Mykel J. Kochenderfer 著・ 繁桝 算男 監訳・本村 陽一 監訳，不確定性下の意思決定，共立出版， 2020

強化学習の一種，Q学習法

Q学習法

Q学習では特定の状態においてある行動をとることで得られる累積報酬の期待値を「Q値（Q-value）」とし，これを更新しながら学習を進める

例：宝探し迷路ゲーム（ゴールまでに宝の数を最大化）

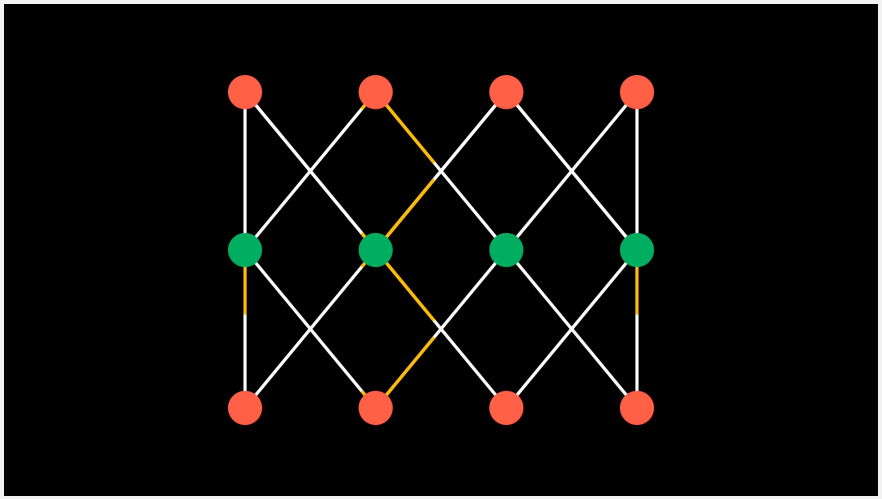
- ランダムに経路を選択 or Q値が高い経路を選択（探査と知識利用の調整）
- 右に進み，宝がある　→　この道はよさそう！（Q値のポイントアップ）
- 左に落とし穴… → この道はやめよう。（ Q値のポイントダウン）

Qテーブル（状態，行動→Q値）

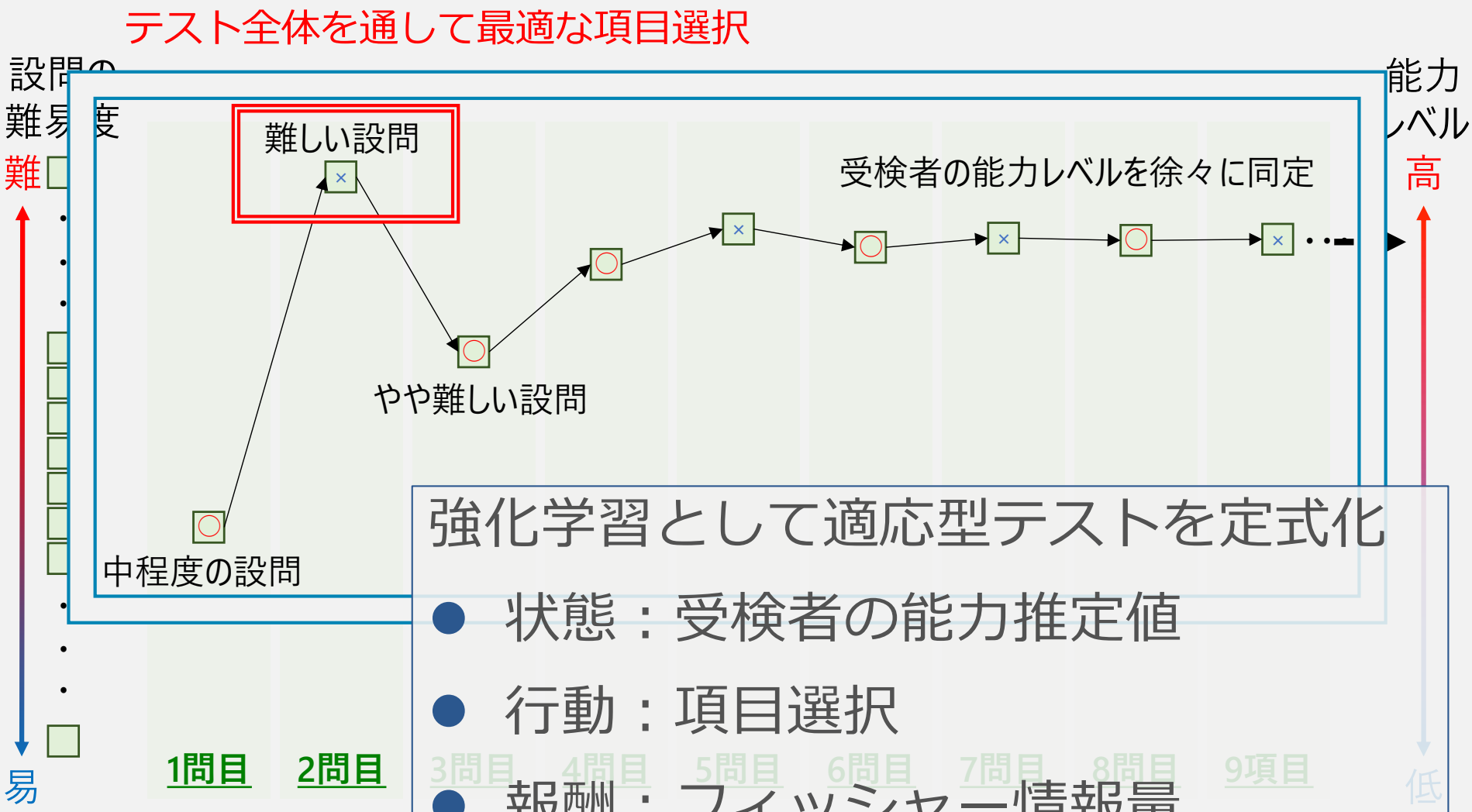
状態	行動	Q
(1,1)	右	3
(1,1)	左	1
(1,1)	下	2

深層強化学習

- Q学習の課題
  - 状態や行動の数が多いとQテーブルの規模が非常に大きくなる
- 深層強化学習
  - 深層学習を用いて状態と行動から報酬を求める関数を近似



深層強化学習を用いた適応型テスト



強化学習として適応型テストを定式化

- 状態：受検者の能力推定値
- 行動：項目選択
- 報酬：フィッシャー情報量

結果

Test length	Method	RMSE
10	DQN (normal)	0.415
	DQN (uniform)	0.348
	MFI	0.493
	KLP	0.443
	FIWL	0.459
	MPWI	0.456
	MEI	0.455
20	DQN (normal)	0.278
	DQN (uniform)	0.254
	MFI	0.331
	KLP	0.322
	FIWL	0.324
	MPWI	0.320
	MEI	0.324

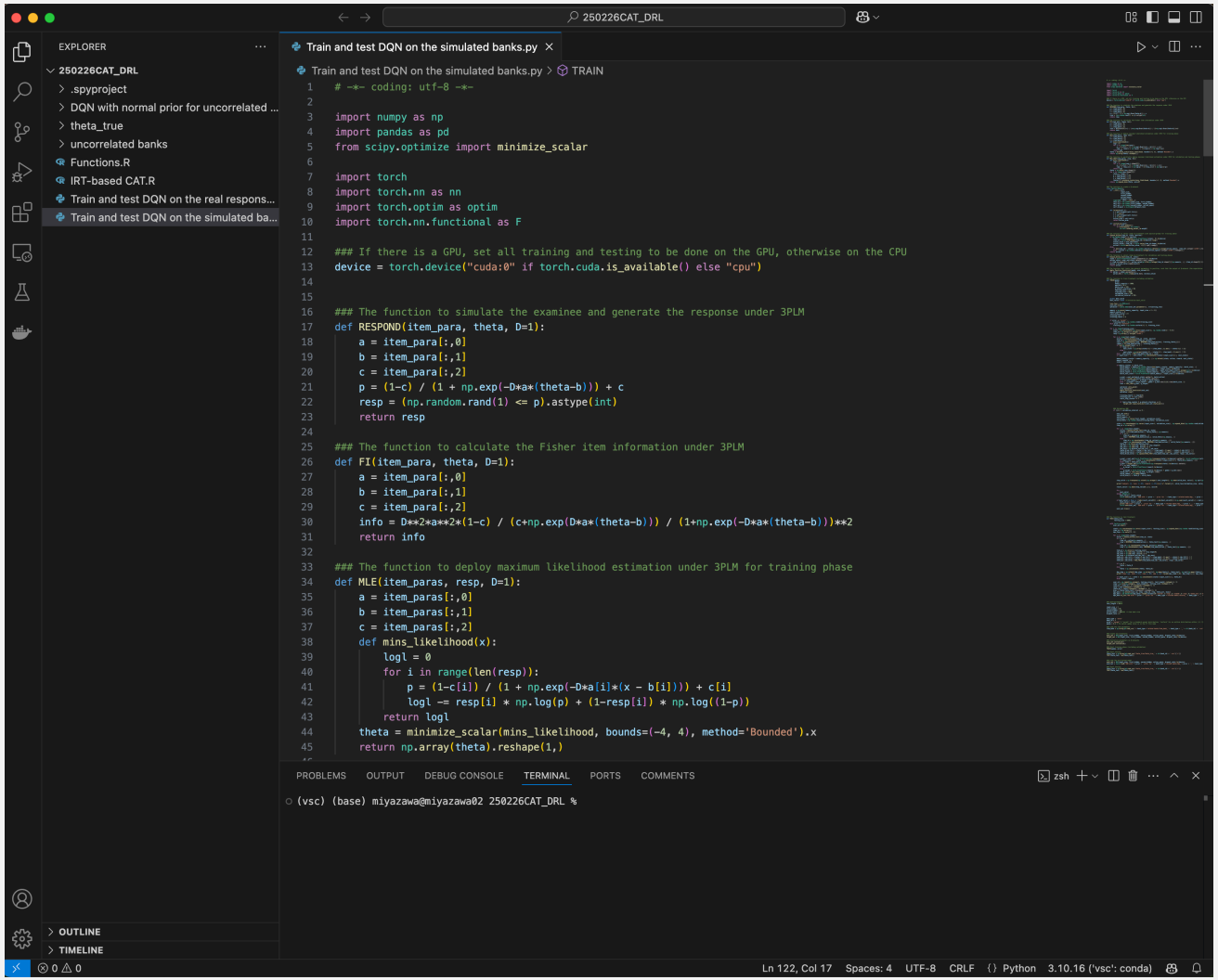
Test length	Method	RMSE
30	DQN (normal)	0.244
	DQN (uniform)	0.207
	MFI	0.277
	KLP	0.272
	FIWL	0.274
	MPWI	0.272
	MEI	0.273
40	DQN (normal)	0.227
	DQN (uniform)	0.207
	MFI	0.245
	KLP	0.243
	FIWL	0.245
	MPWI	0.243
	MEI	0.244

(3) Root mean square error (RMSE)

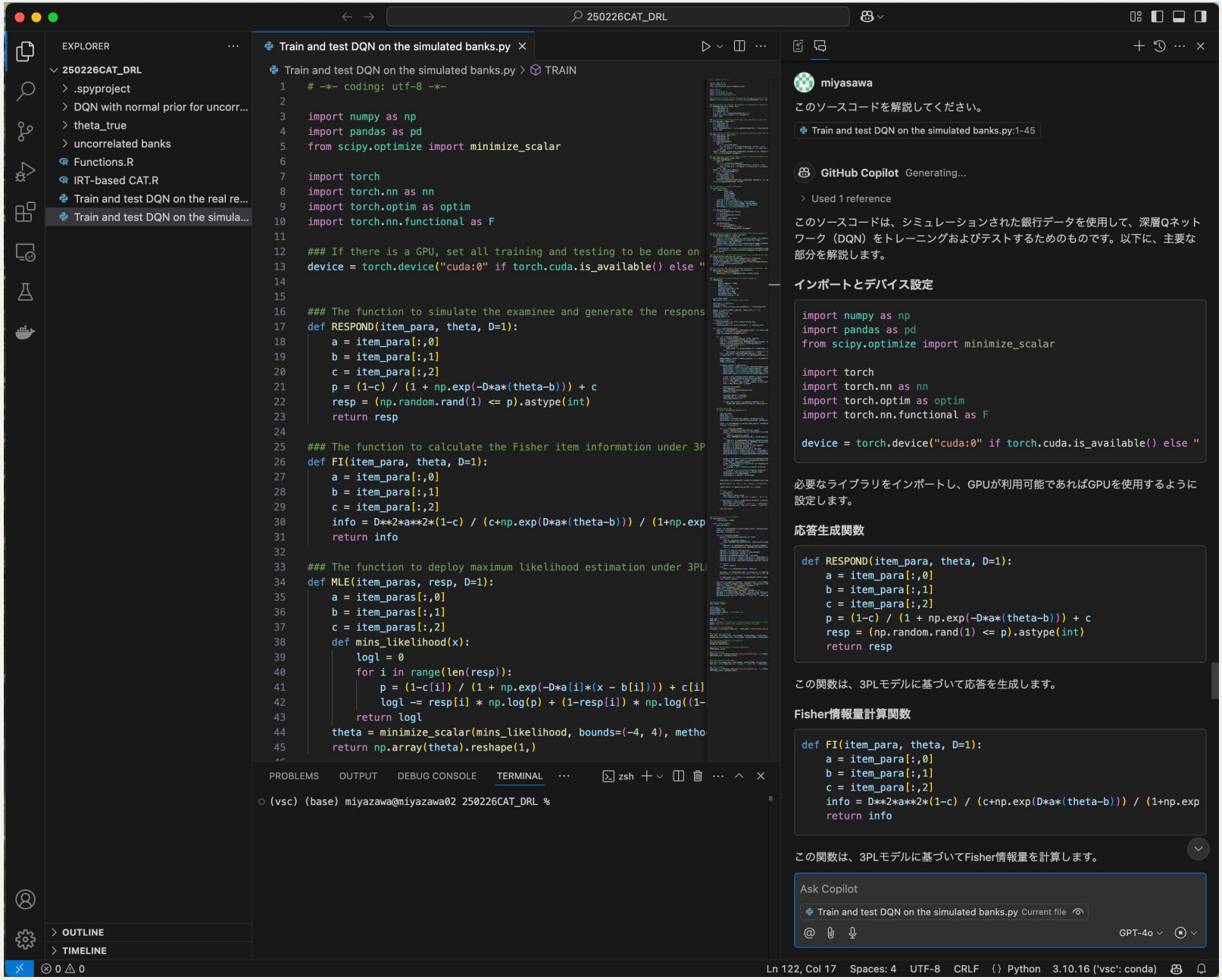
$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)^2}$$

ソースコード

https://osf.io/nv3e4/



番外：GitHub Copilot



まとめ

以下の内容を紹介

- 教育測定における基盤的なコンピテンシー
  - Ackerman, T. A. et al. Foundational Competencies in Educational Measurement. Educ. Meas.: Issues Pr. 43, 7–17 (2024).
- 深層強化学習を用いた適応型テスト
  - Wang, P., Liu, H. & Xu, M. An adaptive testing item selection strategy via a deep reinforcement learning approach. Behav. Res. Methods 1–20 (2024) doi:10.3758/s13428-024-02498-x.