テストによる測定再考

加藤 健太郎
ベネッセ教育総合研究所

日本テスト学会 第19回大会
実行委員会企画録画講演「テスティングの現状と将来展望」#1
2021/9/17-25

概要

教育テストの現在とこれから

1. 測定内容の変化
2. 測定方法の変化
3. 新しいテストの形と課題

※本講演の内容は，発表者による独自の調査・研究にもとづくものであり，ベネッセ教育総合研究所ならびに株式会社ベネッセコーポレーションの見解を代表するものではありません。

Kato, K. (2016). Measurement issues in large-scale educational assessment. *The Annual Review of Educational Psychology in Japan*, 55, 148-164.

教育テストの歴史（主に海外・米国）

Testing movements（19世紀後半～20世紀前半＠米国）
- 知能研究
- 適性検査（軍隊）（Army Alpha & Army Beta, 1917）
- 達成度テスト（Stanford Achievement Test, 1923-; ITBS, 1935-; NAEP, 1969-）
- 大学入試（SAT, 1926-; ACT, 1959-）

教育政策＠米国
- IASA (1994-) → NCLB (2002-) → ESSA (2015-)

国際的大規模調査
- TIMSS (IEA, 1995-)，PISA (OECD, 2000-) etc.

言語テスト
- ケンブリッジ英語検定試験 (1913-)，TOEFL (1963-)，IELTS (1989-) etc.

テストの尺度化・性能評価のための理論・手法として教育測定学が発展
- テスト理論（古典的テスト理論→項目反応理論，等化，DIF etc.）
- 妥当性理論（三位一体論→統一的構成概念妥当性，argument-basedの妥当性検証）

テスト産業の発展
- 大規模教育テストの開発・運用

（参考）Clauser, B. E., & Bunch, M. B. (Eds.) (2021). *The history of educational measurement*. Routledge.

二値採点と合計点

Lord (1952, p. 4)より

Consideration will be restricted to the situation where the examinee attempts every item in the test and the responses are all scored either 0 or 1. Let x_i $(i = 1, 2, \ldots, n)$ represent the score assigned to item i; so x_i is a dichotomous variable that can assume only the values 1 or 0. It will be convenient to use the language of achievement testing and to speak of these alternatives as corresponding to "correct" and "incorrect" item responses, respectively.

Consideration will be restricted further to the case where the test score (s) is the sum of the scores on the n items of which the test is composed:

$$s = \sum_{i}^{n} x_i. \qquad (1)$$

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, 7.

項目反応理論（IRT）モデルとその拡張

正規累積モデル (2PNOM; Lord, 1952)

Raschモデル (Rasch, 1960/1980)

ロジスティックモデル (3PLM; Birnbaum, 1968)

多値型IRTモデル

連続型反応IRTモデル

多次元IRTモデル
　補償型／非補償型，局所依存を許容するモデル（e.g., テストレットIRTモデル）

潜在クラスモデル
　state-masteryモデル，認知診断モデル

階層的IRTモデル

多相IRTモデル

展開型IRTモデル

ノンパラメトリックIRTモデル

(and many, many more)

Birnbaum, A. (1968). Some latent trait models. In Lord, F. M., & Novick, M. R., *Statistical theories of mental test scores*. Addison-Wesley.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). The University of Chicago Press.

測定内容の変化

DeSeCoのキー・コンピテンシー (Rychen & Salganik, 2003)

Future of Education and Skills 2030 (OECD)

21世紀型スキルの指導とアセスメント (ATC21S; Griffin et al., 2012)

文部科学省「資質・能力の三つの柱」 ＝「学びに向かう力・人間性」「知識・技能」「思考力・判断力・表現力」

教科学力からコンピテンシー（competency）の測定へ

“achievement”から“college/career readiness”へ

実生活場面での応用（何を知っているか→どう使うか）＆未知の状況への対応

批判的思考力，協働的課題解決，創造性・・・

非認知スキル，態度，価値観

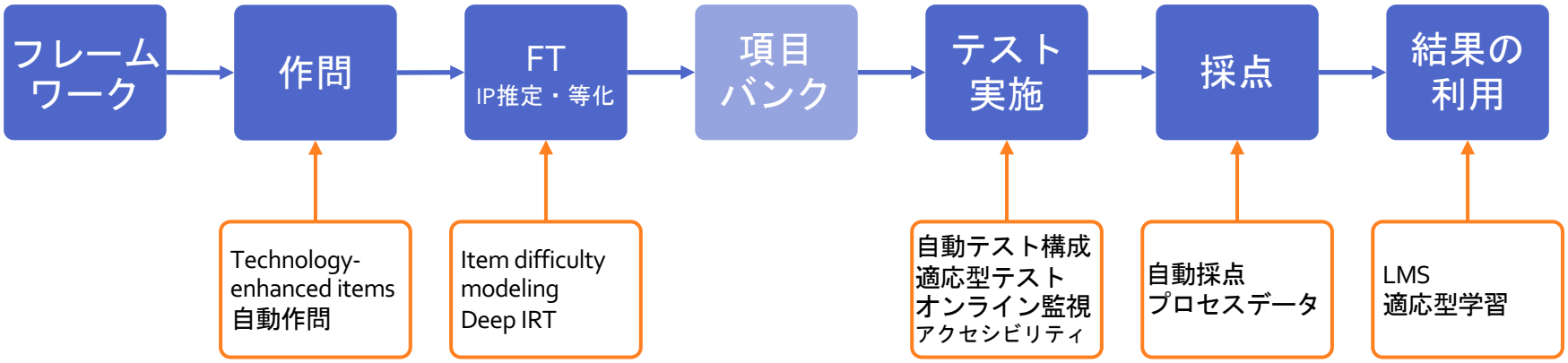
Griffin, P., McGaw, B., & Care, E. (Eds.) (2012). *Assessment and teaching of 21st century skills*. Springer.
文部科学省『学習指導要領「生きる力」』 https://www.mext.go.jp/a_menu/shotou/new-cs/index.htm
OECD “Future of Education and Skills 2030” https://www.oecd.org/education/2030-project/
Rychen, D. S., & Salganik, L. H. (Eds.) (2003). *Key competencies for a successful life and a well-functioning society*. Hogrefe & Huber.

測定方法の変化

Technology-enhanced assessment (TEA)

CBTが前提

テストの開発・実施・運用におけるテクノロジーの役割の拡大



（参考）

NAEP “Technology-Based Assessment Project” https://nces.ed.gov/nationsreportcard/studies/tba/

NAEP “Digitally-Based Assessments ” https://nces.ed.gov/nationsreportcard/dba/

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities. In Veldkamp, B. P., & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Springer Open. https://doi.org/10.1007/978-3-030-18480-3

新しいテストの形

アセスメント進化論（Bennett, 2015）

第1世代
- PBTの焼き直し
- 項目反応理論（IRT）の利用／適応型テスト

第2世代
- 新形式のタスク（technology-enhanced items）
- 開発・運用システムへのITの積極的導入（自動採点，自動作問，オーサリングツール，DB管理・・・）

第3世代 ＝ 次世代アセスメント（next-generation assessment）

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370-407.

次世代アセスメント

「測定 × テクノロジー × 学習科学」

学習への貢献により大きな比重
- institutional useからindividual useへ
- systems approach → 多目的（包括的＋形成的），データの統合＆利用（LMS）
- 学習科学をベースとして設計

	従来型	次世代型
実施のタイミング	学習後	学習中 (embedded)
アクセシビリティ	限定	誰でも (universal design)
学習（テスト）の進行	固定	適応的
結果の返却	時間差あり	即時
項目の形式	解答選択型 (generic)	強化型 (enhanced)

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370-407.
Office of Educational Technology, U.S. Dept. of Education, “Measuring for Learning” https://tech.ed.gov/netp/assessment/

米国での動き

TEA実装に向けた動き
- NAEP → Digitally-Based Assessmentsに移行
- Common Core State Standardsの導入 → 対応するテスト（PARCC，Smarter Balanced）

アカウンタビリティ → 学習者中心の（次世代型）テストへ
- overtestingへの批判 → （州テストの）回数・時間削減
- 一方で，高頻度のモニタリング（単元テスト，パフォーマンス型テスト）＆データ活用への期待
- Innovative Assessment Demonstration Authority (IADA)

大学入学における標準化テスト（SAT，ACT）の受検要件が後退？
- 新型コロナにより，多くの大学で2020～21年度は一時的に免除
- カリフォルニア大学は恒久的な利用廃止を宣言

NAEP “Technology-Based Assessment Project” https://nces.ed.gov/nationsreportcard/studies/tba/
NAEP “Digitally-Based Assessments” https://nces.ed.gov/nationsreportcard/dba/
National Association of State Boards of Education. (September, 2020). Next-generation assessment. *The State Education Standard*, 20(3).
UCA “Exam requirement” https://admission.universityofcalifornia.edu/admission-requirements/freshman-requirements/exam-requirement/

日本国内の動き

新学習指導要領「資質・能力」

教育のデジタル化促進

GIGAスクール構想
https://www.mext.go.jp/a_menu/other/index_0001111.htm

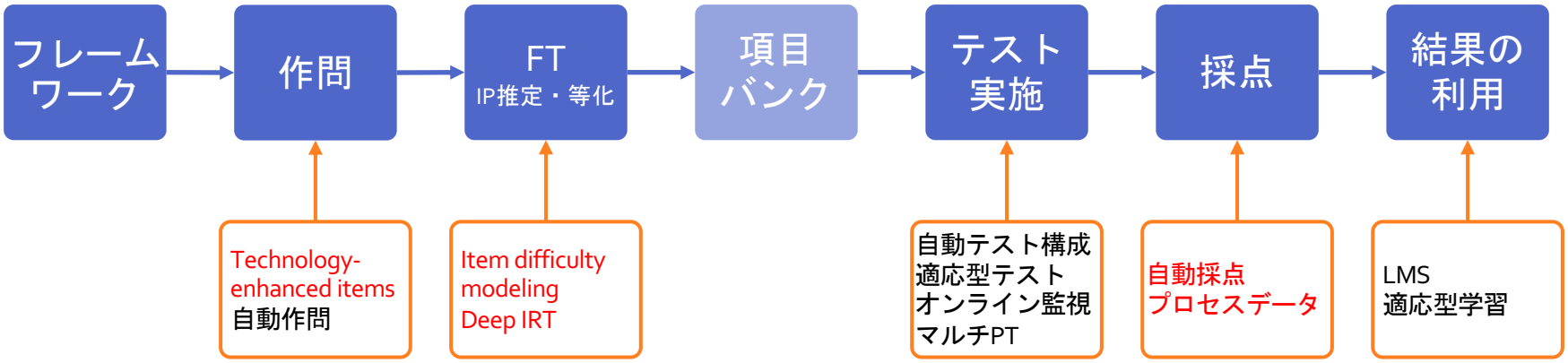
学びの保証オンライン学習システム（MEXCBT）
https://www.mext.go.jp/a_menu/shotou/zyouhou/mext_00001.html

全国的な学力調査のCBT化検討WG
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/1421443_00004.htm

民間企業による，様々なデジタル学習サービスが普及

教育ベンチャーも多数出現
- IT技術の教育分野への応用
- 独自のサービス・プラットフォーム・アプリケーション

測定方法の変化（再掲）


フレームワーク
作問
FT
IP推定・等化
項目バンク
テスト実施
採点
結果の利用
Technology-enhanced items
自動作問
Item difficulty modeling
Deep IRT
自動テスト構成
適応型テスト
オンライン監視
マルチPT
自動採点
プロセスデータ
LMS
適応型学習

Technology-enhanced items (TEI)

より真正 (authentic)かつ複雑なタスク・測定をCBTで実現

- パフォーマンス型タスク： 現実的・具体的な文脈 → 与えられた情報を比較・分析・統合 → 解答
- コンピテンシー測定に適合
- プロダクトデータ＋プロセスデータ

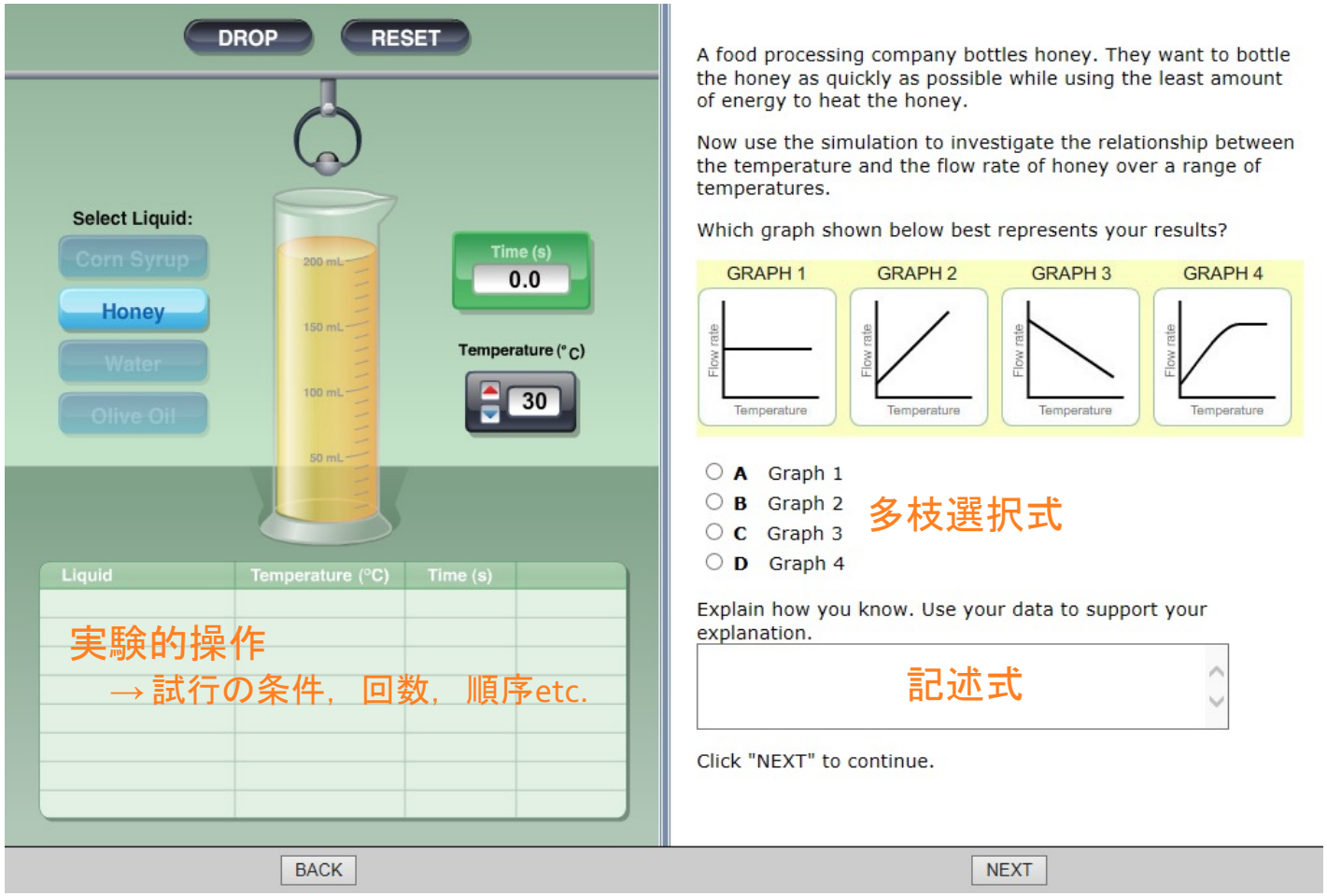
TEIのタイプ分け (Wools et al., 2019)

- シミュレーション型
- マルチメディア型
- ハイブリッド型

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities. In Veldkamp, B. P., & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Springer Open. https://doi.org/10.1007/978-3-030-18480-3

シミュレーション型TEIの例

NAEPのサイト(https://nationsreportcard.gov/science_2009/ict_tasks.asp)より



プロダクトデータ

複合的なデータ
ひとまとまりのタスクから，複数のタイプの異なる解答が得られる

解答生成型／パフォーマンス型タスク
例：記述，音声，動画
従来のテスト理論に乗せるためには，項目レベルでスコア化が必要
観測されたパフォーマンスを分解，符号化，数値化する仕組みが必要

記述解答の自動採点 (e.g., Uto, 2021)
自然言語処理＋機械学習
従来法 → 答案から決まった特徴を抽出し，回帰モデル等でスコアを予測（ETSのe-rater®など）
現在の主流 → 深層学習（DNN）モデル（特徴の自動抽出；形態素解析→ベクトル化→・・・）

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48*, 459-484.

プロセスデータ

反応時間

解答の変更
　頻度，パターン

解答に関わる操作
　キーストロークログ，情報の参照（←ナビゲーション，クリック，スクロールetc.），インタラクション

補助機能／ヘルプの活用

センシング，録画etc.

期待される用途
- スコアの推定精度向上
- 適応型テスト（診断）
- 不正検知（監視）
- 受検態度（スクリーニング）
- 妥当性検証
- など

「プロダクト」は同じでも，学習状態（learning status）に関するよりrichな情報が得られる可能性

（参考）
Jiao, H., Zhou, T., & Ding, Y. (2021). *Analyzing responses, response time, and answer changes for cognitive diagnosis with machine learning algorithms*. Spotlight Talk, IMPS 2021.

キーストロークログ（KSL）

多方面で応用が進んでいる
　例：ライティングの認知・発達プロセスの研究，タイピングスキル，翻訳etc.

ライティングプロセスの研究 (Uto et al., 2020)

Fig. 7 Normalized mean values of emission distributions for each state

Uto, M., Miyazawa, Y., Kato, Y. et al. (2020). Time- and learner-dependent hidden Markov model for writing process analysis using keystroke log data. *International Journal of Artificial Intelligence in Education*, 30, 271–298.

反応時間 (RT)

反応時間を考慮した測定モデリング (Molenaar, 2015)

1. 能力パラメタ（θ）の推定精度改善
2. 反応プロセスや解答方略の推測

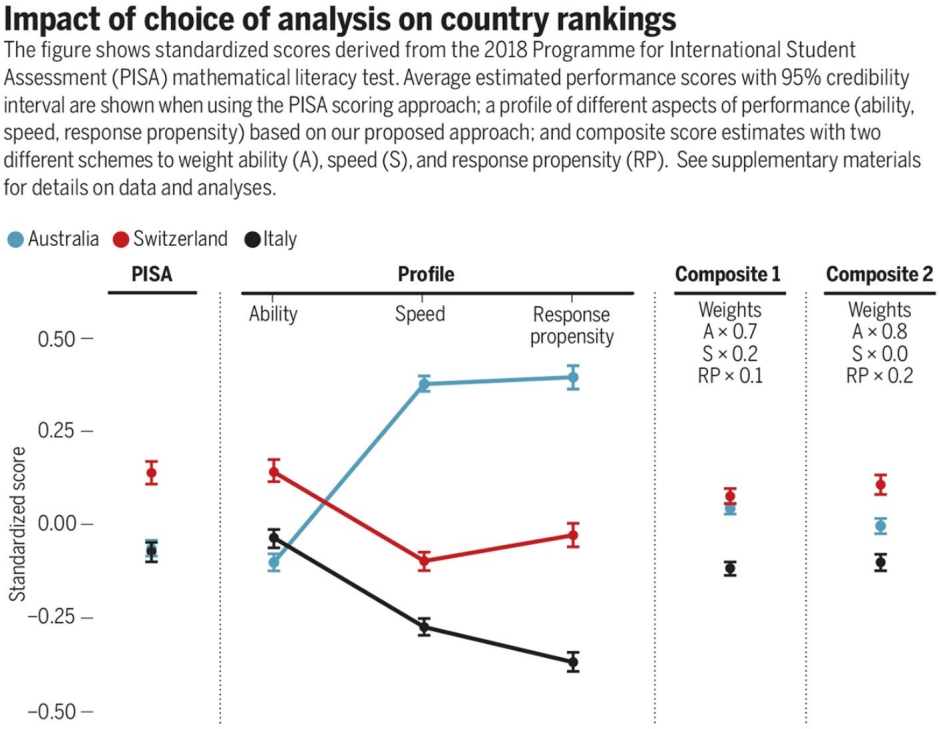
2の研究事例 (Pohl et al., 2021)

1. 解答速度-正確さのトレードオフ
2. 低速解答による時間切れ（未到達）
3. 反応傾向（omitせず解答する傾向）

を考慮して解答行動をモデル化

データ＝正誤，無解答フラグ，未到達率

結果のレポート
→ プロファイル化
→ よりfairな比較＆スコアの具体的な解釈



Molenaar, D. (2015). The value of response times in item response modeling. *Measurement*, 13, 177-181.
Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science*, 372 (6540), 338-340.

IRT

機械学習・深層学習を利用したパラメタ推定

item difficulty modeling (IDM) ← 項目の持つ属性や様々な特徴から困難度を学習・予測 (e.g., Settles et al., 2020)

Deep IRT ← RS仮定や等化デザインの緩和 (Ueno et al., 2021)

on-the-fly calibration

適応型テスト

自動テスト構成

Settles, B., LaFlair, G. T., & Hasegawa, M. (2020). Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics,* 8, 247–263.
Ueno, M., Fuchimoto, K., & Tsutsumi, E. (2021). e-Testing from artificial intelligence approach. *Behaviormetrika, 48*, 409-424.

TEAの課題

データの種類と量の変化

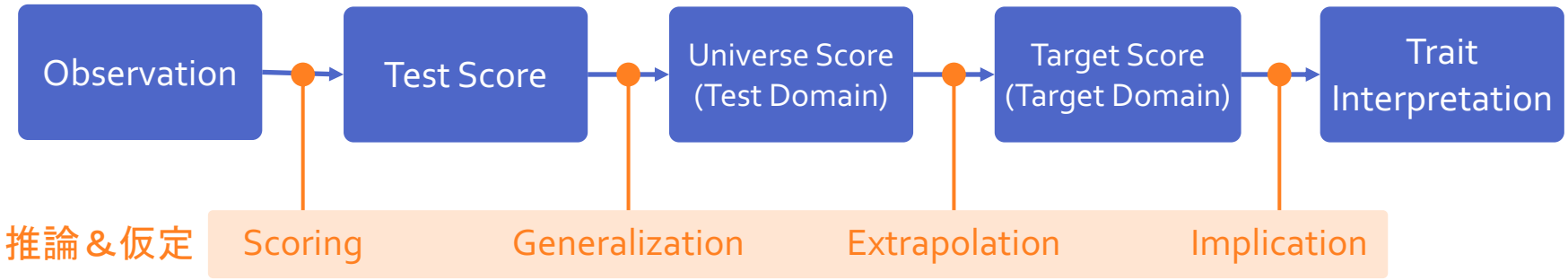
1. 多様なデータに対応する採点＆測定モデリング
   ただし，モデルが複雑になるほど運用にかかる負荷が大きくなる
   → 実用性を考えれば，モデルのアップデート＆検証だけでなく，取り回しのしやすさもカギ（e.g., 等化）
2. 妥当性

妥当性検証の枠組み

「何のために」「何を」「どのように」測るのか？

→ 実現できている程度は？（合理的説明＋エビデンス）

Interpretive arguments for trait interpretations (Kane, 2006)



→ これらの推論の適切さをエビデンスにもとづき評価するのがvalidity arguments

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement* (4th ed., pp. 17-64). ACE/Praeger.

TEAの妥当性：機会と脅威

Wools et al. (2019, p. 16)より

Table 1.1 Opportunities (+) and threats (−) for validity (Implication)

	Scoring	Generalization	Extrapolation	Decision
TEI *Items and tasks*				
Simulations	−	−	+	−
Multi-media enhanced tasks	−	−	+	−
Hybrid tasks	−	−	+	−
Test construction, assembly and delivery				
Automated item generation			+	
Adaptive engines	+	+		
(online) Proctoring	+			
Personal needs and preferences				
Tools for accessibility	+	−	+	
Bring your own device	+	−	+	
Personalized feedback				+

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities. In Veldkamp, B. P., & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Springer Open. https://doi.org/10.1007/978-3-030-18480-3

TEIの妥当性

（−）採点（scoring）
「パフォーマンス」の数量化＋プロセスデータの利用（併用）
→ 採点ロジックが複雑化（高度な測定モデル，機械学習・深層学習etc.）
→ 透明性（了解性）が低下

（−）テストの測定領域への一般化（generalization）
個々のタスク＝深く狭く（特定の文脈，より深い取り組み＆解答生成）
→ 代表性の保証が難しい（construct underrepresentationのリスク）

（+）目標領域への外挿（extrapolation）
真正性を重視したタスク設計（しかし，CUが生じていれば解釈は限定的）
これからの世代の人々 → デジタル環境への親和性

（−）意思決定（decision）
（上記の状況下で）スコアの利用に際して妥当な基準設定ができるかは不透明

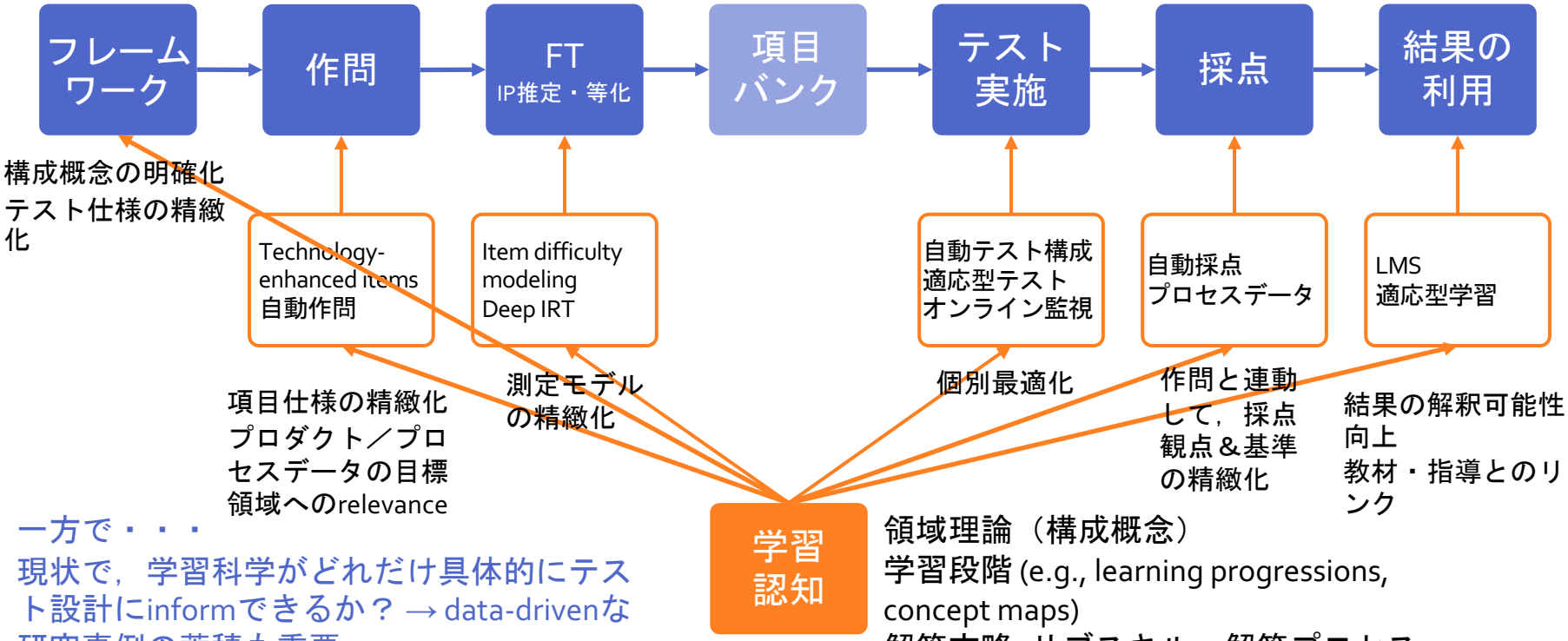
Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities. In Veldkamp, B. P., & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Springer Open. https://doi.org/10.1007/978-3-030-18480-3

妥当性向上のために：scoringとimplicationsの観点に対して

分業から協業へ

現状＝テクノロジーによるパーツの置き換え

文字通り「測定 × テクノロジー × 学習科学」の協業によって妥当性を高めることはできないか



一方で・・・

現状で，学習科学がどれだけ具体的にテスト設計にinformできるか？ → data-drivenな研究事例の蓄積も重要

相当の「作り込み」が必要 → 設計・作問・実装にかかる手間やコストに懸念

妥当性向上のために：generalizationの観点に対して

真正性と領域網羅性・信頼性のトレードオフ

テストの文脈，目的，用途は？
学習指導・形成的評価の文脈
→より深いengagementや具体的なFBが必要
→ 対象とする領域（単元）は測定領域・目標領域の一部で十分である可能性，ローステークス

包括的評価の文脈
→ 波及効果等を考えれば，真正性を高める努力は必要だが・・・
→ depthよりもwidth (coverage)が重要，ハイステークス

cf. systems approach

まとめ

次世代アセスメント

テクノロジーと学習科学の知見を開発・運用の様々な段階に活用し，より学びに貢献できるアセスメント（＞テスト）を追求

TEI → 多種多様なデータ → 多種多様な測定 → 学習状態に関するよりrichな解釈

妥当性には課題も

技術面だけ先行しても結果の解釈や有用性には限界

→ 学習・認知分野の知見活用，研究事例の蓄積 (e.g., Goldhammer et al., 2020)

テストの文脈，目的，用途を踏まえ，実現可能性・継続性のあるテストの形を考えていく必要

Goldhammer, F., Scherer, R., Greiff, S. (Eds.) (2020). *Advancements in technology-based assessment: Emerging item formats, test designs, and data sources.* Frontiers Media SA. doi:10.3389/978-2-88963-506-1

ご清聴ありがとうございました。

EOF