統計数理(2024)
第 72 巻 第 1 号 3–21

特集「心理統計学の新展開」
[総合報告]

技術強化型テスティングにおける測定モデルの考察と展望

加藤 健太郎†

(受付 2023 年 12 月 4 日；改訂 2024 年 3 月 7 日；採択 3 月 11 日)

要　　旨

コンピュータで実施するテスト(CBT)は教育アセスメントにおける主要なテスト実施形態となりつつあり，従来の紙筆式テスト(PBT)と比較して，コンピュータ上で実施することによる様々な利点，特に技術強化型項目(CBT だからこそ実現できるテスト問題)と，学習者の学習状態をよりよく測定する目的でのその利用に注目が集まっている．本稿では技術強化型項目から得られる様々なデータを心理測定学の立場から分析・モデリングする近年の試みや実践のレビューを行い，現在の課題や今後の研究の方向性について検討した．その結果，(a)TEIs で実現できる様々な解答形式の測定論的性質への影響，(b)プロセスデータの利用可能性に関する探索的検討，(c)TEIs を含む尺度と既存の測定尺度との等価性の検討，(d)プロセスデータを利用した解答過程の測定モデリングに関する研究が近年盛んに行われていることがわかった．今後の課題として，特にプロセスデータを用いた解答過程のモデリングについて，学習改善につながる有用な情報を提供していくことができるか，TEIs を含むテストの継続的・安定的な運用を実現できるかという観点で継続的な検討・改善が必要があると考えられた．

キーワード：コンピュータによるテスト, 項目反応理論, 技術強化型項目, プロセスデータ.

1. 問題と目的

近年の教育テストの動向は，測定内容の変化と測定方法の変化によって特徴づけられる(Kato, 2016)．測定内容の変化とは，学校教育において習得すべきものが，従来の教科・科目固有の知識から，現実的な問題解決の場面においてそれらを総合的に応用して問題を考察・解決できるかという，いわゆる「コンピテンシー」にシフトしており，教育テストが測定する対象もこれに対応して変化していることを指す．こうした背景には，DeSeCo のキー・コンピテンシー(Rychen and Salganik, 2003)や 21 世紀型スキルの指導とアセスメント (Care et al., 2018)，経済協力開発機構(OECD)が実施する PISA(Programme for International Student Assessment)といった，欧米における「これからの社会に必要とされるスキル」の定義と測定について再考する動きがあった．2020 年から順次導入されている日本の新しい学習指導要領において，「思考力・判断力・表現力」が資質・能力の要素の一つとして挙げられていることも，こうした世界的な潮流を反映していると言える．

もう一方の測定方法の変化の第一に挙げられるのは，従来の紙筆式(paper-based test [PBT])

† ベネッセ教育総合研究所：〒 206–8686 東京都多摩市落合 1–34

からコンピュータで実施するテスト(computer-based test [CBT])への移行が進んでいることである．例えば，日本国内においては，2019 年に始まった文部科学省の GIGA スクール構想の下で教育のデジタル化が進められ，これに並行して文部科学省 CBT システム(MEXCBT)が整備されつつある．文部科学省が実施する全国学力・学習状況調査や，大学入試センターが実施する大学入学共通テストといった公的な試験においても CBT の導入が検討されており(前者においては 2024 年度より順次導入が始まる)，今後は民間業者が提供するテストや日常の学習ツールも含めて CBT がさらに普及することが予想される．海外に目を向ければ，日本の全国学力・学習状況調査に相当する全米学力調査 NAEP(national assessment of educational progress)は早期に CBT の検討や試行を始めている(現在は digitally-based assessment と呼ばれている；https://nces.ed.gov/nationsreportcard/dba/)．また，先に挙げた OECD の PISA においても 2015 年より本格的に CBT への移行が進められている(こうした世界的な動きを受けて，日本でも CBT 導入の検討が加速したと言える)．

CBT の普及に伴い，単にテストをコンピュータ上で実施することにとどまらず，様々な情報技術がテストの開発・運用の各フェーズにおいて活用されるようになってきている．例えば，分寺 (2023)は，CBT に関する近年の研究トピックを(a)PBT と CBT の得点の比較可能性(モード効果)の検証，(b)適応型テスト(computerized adaptive testing [CAT])，(c)新しい形式のテスト項目，(d)オンライン試験における不正行為とその対策，(e)ログデータの活用，(f)(障がいを持つ受検者らへの)特別な配慮の 6 つに整理している．

Bennett (2015)は，現代における CBT の進化を 3 つの段階に分類・整理している．第 1 世代のテストは，従来の PBT をコンピュータ上で実施・受検できるように移植したものである．PBT 版の項目がそのまま画面に表示され，解答にはマウス操作やキーボードでのタイピングが必要となるものの，項目の形式自体は PBT と同じである．技術面では，項目反応理論(item response theory [IRT])による尺度構成に基づく，適応型の出題(adaptive testing)機能が搭載されたものもある．

第 2 世代のテストを特徴づけるのは，後で述べる技術強化型項目(technology-enhanced items [TEIs])，いわゆる「CBT ならではのテスト問題」の導入である．これにより，先に述べた測定内容の変化と合わせて，従来の形(第 1 世代やそれ以前の PBT)では難しかった能力やスキルの測定が可能になるという期待が高まっている．また，テスト項目の形式に加えて，テストの開発・運用のプロセスの中にも情報技術が積極的に導入される．本稿では扱わないが，近年では特に自然言語処理と機械学習(AI)を活用したテスト問題の自動生成や論述・口述解答の自動採点に注目が集まっており，例えば Jiao and Lissitz (2020)のような研究事例集でもこれらが主要なトピックとして取り上げられている．第 2 世代のテストに見られるような，CBT を前提として情報技術を積極的に取り入れたテストは技術強化型テスト(technology-enhanced tests)と呼ばれる．現在運用されている CBT の多くは第 1 世代であり，一部は第 2 世代に向けて進化しつつあると言われている．

これらに続く第 3 世代のテストを，Bennett (2015)は次世代アセスメント(next-generation assessment)と呼んでいる(本稿では，「テスト」と「アセスメント」という言葉を同義に用いる)．次世代アセスメントは，第 2 世代同様にテクノロジーの積極的な活用を前提としているが，オンラインの学習環境を前提として，学習の評価ではなく学習への貢献により大きな比重を置いて設計される点(assessment of learning から assessment for learning への転換)が強調される．例えば，米国教育省の報告書 (U.S. Department of Education, 2017, Sec. 4) では，学習の改善をゴールとして適切に技術を活用することが謳われており，その中で表 1 のような形で従来型のアセスメントと次世代アセスメントとの対比がなされている．

本稿では，大規模教育テストに見るこうした動向を踏まえて，能力測定の核となる測定モデ

表 1．従来型と次世代型のアセスメントの比較．

	従来型	次世代型
実施のタイミング	学習後	学習中
アクセシビリティ	限定的	誰でも
学習（テスト）の進行	定型	適応的
結果の返却	時間差あり	即時
項目の形式	汎用型	強化型

ルの現状と課題について考えたい．従来の PBT，あるいは第 1 世代の CBT においては，各テスト項目に対する解答を正答 1，誤答 0 の二値で採点し（項目得点），これらをテスト項目全体に渡って合計した正答数（素点）や正誤のパタンから推定される IRT の能力母数を受検者の能力の「測定値」あるいは「推定値」としてきた（cf. Lord, 1952）．しかし，CBT の特性を活かしたテスト項目（i.e., TEIs）の作成やデータ収集が可能となってきたことで，こうした従来の枠組みを超えた測定の可能性が開けてきている．本稿ではまず，TEIs の特徴を概観し，能力測定に関して TEIs がどのようなデータを提供し得るかを考察する．そして，これを踏まえた上で，TEIs の利用を念頭に置いて測定論の文脈で行われてきた分析やその中で応用・提案されてきた測定モデルのレビューを行う．

なお，本レビューでは以下の手順によって文献を収集した．検索には ERIC, PubMed, Google Scholar，および論文検索・閲覧サービスの DeepDyve を主に用いた．基本的な検索ワードとして “technology-[enhanced, based, rich]”，“[measurement, response] model”，“innovative [items, assessment]” の用語の組み合わせを用い，特定のモデルに関する検索結果を適宜追加した．得られた文献リストから，タイトルおよびアブストラクトを参照して，概ね 2020 年以降に出版された，本稿のテーマに合致すると思われる比較的新しい論文を主な検討の対象とした．

2. 技術強化型項目

2.1 技術強化型項目の定義と特徴

技術強化型項目（TEIs）は，広義には「従来の多枝選択式や解答構築式ではない，コンピュータで実施されるテスト項目」とされている（Bryant, 2017, p. 2）．より詳しい定義や呼称については研究者によってバリエーションがあるが（e.g., technology-based items, technology-enabled items, innovative items），主に (a) 解答形式，(b) 解くべき問題（タスク）の特徴，(c) 解答補助ツールの 3 つの観点から特徴づけることができる．

第一の特徴である TEIs で用いられる解答形式に関して，Jiang et al.（2021）は，2017 年の NAEP において採用された例として，タイピングによる記述解答の入力（文字だけでなく，数式エディタも含む），多枝多選択方式（複数選択可能な解答選択式；multiple-selection multiple-choice format），ドラッグ&ドロップ（解答要素を所定の箇所に配置する＝マッチングや順序付けなど），ドロップダウンによる穴埋め，グラフやイラストの領域指定による解答を挙げている．この他にも，言語テストでは音声入力（スピーキングテスト）が従来より用いられており（ただし，TEIs では記録方式がアナログからデジタルになる），さらに今後は，画像や動画などを含む新しい解答形式が出てくる可能性もある．Bryant（2017）は，解答の自由度（制約の強さ）の観点から，より制約が強い順に (a) 多枝選択式（multiple-choice；所与の選択枝から選ぶ），(b) 選択・同定（selection/identification；テキストの一部を指示する，グラフの領域を指定するなど，「選択枝」よりも対象が広いものを指す），(c) 要素の並べ替え（reordering/rearranging），(d) 代入・修正（substitution/correction；既にある要素を置き換えたり変更したりする），(e) 完成（completion；穴埋め形式など），(f) 構築（construction；記述・論述など），(g) プレゼンテー

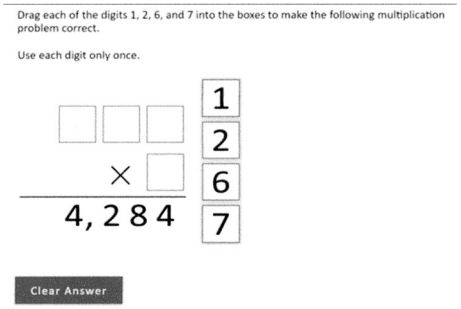

図 1. NAEP 数学テストにおけるドラッグ&ドロップ形式の項目例(U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP], 2017 Mathematics Digitally-Based Assessment; https://www.nationsreportcard.gov/math_2017/sample-questions/?grade=8).

ション(presentation;意図やストーリーを持って行う表現)の 7 つのカテゴリで TEIs の解答形式の分類を試みている.

第二の TEIs のタスクの特徴として,特にコンピテンシーの測定において,より日常生活・社会生活に近い具体的な状況(文脈・シナリオ)を設定し,リアリティ(テストの分野では,特に真正性 [authenticity] と呼ばれる)の高いタスクやタスク素材を提示できるという点が挙げられる. Wools et al. (2019)は,TEIs に特徴的なタスクの形式をシミュレーション型(条件を様々に変えて実験を行い,結果を考察する),マルチメディア型(動画や音声の利用),ハイブリッド型(複数の形式の混合)の 3 種類に分類しているが,これらに共通しているのは,CBT 内で設定された現実を模した環境・状況の中で,受検者がインタラクティブなやりとりをしながら解答に至るタスクを実行していく点であろう. 例えば,CBT で実施された 2022 年の PISA の「数学的リテラシー」アセスメントでは,スプレッドシートを提示して計算式を考えさせたり得られた数値を考察する,地図を表示して所定の箇所をマウスでポイントし,ポイントした各点から目的地までの距離を表示させてその特徴を考察する,期間・利率・元金などを指定して積立貯金のシミュレーションを行わせる,といったタスクの例が示されている (OECD, 2023).

第三の特徴に関して,上述したタスクを実行する際の補助ツールとして計算機,デジタルメモ,デジタル定規,グラフ描画ツール等がデジタル環境内でシームレスに利用できることが挙げられる.

公開されている TEIs の例を図 1 および 2 に示す. 図 1 は NAEP の第 8 学年の数学の項目例で, Jiang et al. (2021)が分析に用いたものである. この項目におけるタスクは,筆算が成立するように 4 つの空欄のそれぞれに「1」「2」「6」「7」のいずれかの数字をドラッグ&ドロップすることである. ドラッグ&ドロップ形式は,大規模テストプログラムにおいて公開されている TEIs の中で最もよく使われている解答形式である (Russell and Moncaleano, 2019).

図 2 は, PISA 2012 で実施された創造的課題解決力(creative problem-solving)アセスメントにおける項目例で, Han et al. (2022)が分析に用いたものである. これは,鉄道駅の自動券売機で指定された条件に合う切符を購入するというタスクであり,画面の右側に示された架空の券売機のボタンをクリックして操作を行う. まず,乗車する鉄道網(地下鉄か広域鉄道か),料金種別(通常か割引か),切符の種類(1 日乗車券か片道か)を決めて,最後に「購入」ボタンを押して購入を完了する.

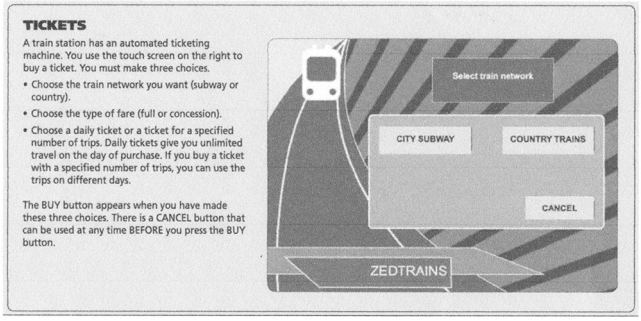

図 2. PISA 2012 における創造的問題解決の項目例（OECD, 2014, p. 39）.

2.2 技術強化型項目から得られるデータ

PBT・CBT のいずれにおいても，従来の能力測定において主に用いられてきたのはプロダクトデータ（product data）である．すなわち，受検者がテスト問題を解き，最終生成物として入力された解答（e.g., どの選択枝を選んだかや記述答案）が記録・採点され，それを能力測定のためのデータとしてきた．プロダクトデータを生成する解答形式は，客観的に正誤が決まる解答選択型（主に多枝選択式）と，基準にもとづく採点を要する解答生成型（穴埋め式や記述式）に大別される．TEIs においてもこの区別が当てはまるが，それぞれのタイプの中で，2.1 節で示したようにより多様な解答形式が実装できる．

一方，CBT の下では，解答の過程で生じる様々なデータ（プロセスデータ）を収集することができる．解答中に発生したイベントがタイムスタンプを伴って記録されたものはログファイル（log files）と呼ばれる（Provasnik, 2021）．プロセスデータ（process data）とは，主にログファイルから抽出・変換される情報から成るが，その中でも特に受検者のテスト項目への解答過程を反映する，測定したい構成概念に関連するデータを指す（Provasnik, 2021）．テスト中に CBT システム以外（e.g., 視線追跡，MRI，ビデオ監視など）から得られるデータもプロセスデータに含まれ得る．

Xiao et al. (2022)は，プロセスデータとして利用される情報は反応時間（response time）と反応行動（response behavior）に大別され，反応時間に関する測定モデリングの研究の蓄積に比して，反応行動の活用に関する研究は依然探索的な段階にあると述べている．反応行動データとして利用され得るものとしては，解答を入力・変更した順序（パタン）や回数，シミュレーション型タスクで設定した条件の値や実験順序・実験結果，補助ツールの使用や回数といったものが考えられる．例えば，図 1 のドラッグ&ドロップ形式であれば，どの数字（ソース）をどの空欄（ターゲット）にどのような順番でドロップしたかという反応列（response sequence）が得られる．図 2 の例では，鉄道網の選択 → 料金種別の選択 → 切符種別の選択 → 購入決定という一連の操作に関して，前に戻ってやり直すといった試行錯誤の行動も分析対象となり得る．こうした主としてマウス操作による反応の記録に加え，文字あるいは単語の入力を記録したキーボードの打鍵記録（keystroke log）も反応行動と考えることができる．

従来プロダクトデータのみにもとづいて算出していたテスト得点の推定精度や妥当性を高めることや，プロダクトデータだけでは把握できない，受検者の解答方略やより詳細な学習状態に関する情報を得るという目的において，プロセスデータの活用が注目・研究されている（e.g., 北條, 2023）．

個々の TEI から得られる多種多様な解答データ（ローデータ）は，次節で述べる測定モデルへの入力とするために，必要に応じて適当な「得点」に変換される．Luecht (2017)はローデータを

項目得点に変換する関数を採点評価器(scoring evaluator)と呼んだ．最も単純な採点評価器はパタンマッチングを行うものである．例えば，正誤の二値採点を行う場合には，解答と正答が一致していれば 1，そうでなければ 0 を返す関数として採点評価器が定義される．ここでの解答および正答は，多枝選択式であればそれぞれの選択枝番号，穴埋め式であれば記入された単語と正解となる単語といったものとなるだろう．後述する 3.3 節に，Betts et al. (2022)が研究対象とした，多枝多選択方式(複数選択可の多枝選択式)の解答に適用される採点方法の例を挙げているが，こうした場合には比較対象がベクトルあるいは集合で表されることもあり得る．

Betts et al. (2022)の例では多値採点も行っている．多値採点は，例えば解答と正答のパタンが一致する程度(個数など)に対して段階的に(一致度が高いほど高得点になるように)数値を割り当てるものである．記述式を中心とする構築型の解答に対しては多値採点が行われることが多いが，これは従来は人間の採点者が予め設定された採点基準を用いて行っていた．しかし，第 1 節で述べたように，近年では深層学習による自動採点が実用化され，これがそのまま採点評価器の役割を果たすようになってきている．

採点ルールの決定には，様々な条件を考慮する必要がある．まず，従来のテストと同様に，TEIs の設計や作問における指針となるのはそのテストで何を測定しようとしているかという意図である．したがって，どのような解答データを収集し(そのような仕組みを作り)，それらをどう採点するかは作問時にある程度想定されていることが望ましい(こうした点に関する知見を積み上げる目的で 3.3，3.4 節で紹介する探索的研究が行われていると言える)．また，測定モデリングを考慮するとき，基本的なモデル仮定である局所独立性(3.1，3.2 節参照)を担保するために複数の解答データをまとめて多値採点するなどの方策が取られることがある (Luecht, 2017)．

3. 技術強化型項目と測定モデル

3.1 項目反応理論モデル

従来の PBT(や第一世代の CBT)に用いられる標準的な測定理論は項目反応理論 (IRT; Lord, 1952)である．IRT における測定モデルは，テスト項目への反応(ここでは，採点評価器によって変換された項目得点を指すものとする)が生成されるメカニズムを表現する確率モデルであり，受検者 $i(i = 1, \ldots, N)$が持つ測定したい特性 θ_i(能力母数)と，その特性を測定する一連のテスト項目$(j = 1, 2, \ldots, J)$の特徴を表す母数 $\boldsymbol{\eta}_j$(項目母数)，そしてこれらの母数によって条件づけたときの受検者 i のテスト項目 j への反応 x_{ij} が生起する確率

$$P(X_{ij} = x_{ij}|\theta_i, \boldsymbol{\eta}_j) \tag{3.1}$$

を表現するものである．この確率を能力母数の関数と見なしたとき，特に項目反応関数(item response function)と呼ぶ．

実際に運用されているテストにおいて最もよく使用されているのは，基本モデルの 1 つとされる 3 母数ロジスティックモデル (Birnbaum, 1968)

$$P(X_{ij} = 1|\theta_i, \boldsymbol{\eta}_j) = c_j + \frac{1 - c_j}{1 + \exp(-a_j(\theta_i - b_j))} \tag{3.2}$$

や，上式で $c_j = 0$ と置いた 2 母数ロジスティックモデルである．ここで $X_{ij} \in \{0, 1\}$ は受検者 i が項目 j に正答すれば 1，誤答すれば 0 の値を取る二値変数である(従って，上式は項目 j への正答確率を表す)．また，$\boldsymbol{\eta}_j = \{a_j, b_j, c_j\}$ であり，それぞれ識別力，困難度，当て推量パラメタと呼ばれる．上記のモデルでは項目反応確率を左右する能力母数は単一の変数であることを仮定しており，これを一次元性(unidimensionality)の仮定と呼ぶ．

一揃いの J 個の項目に対して特定の反応パタン $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{iJ}\}$ が得られる確率は，能力母数と項目母数が所与のときに各項目に対する反応 X_{ij} は互いに独立であると仮定して，

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i | \theta_i, \boldsymbol{\eta}) = \prod_{j=1}^{J} P(X_{ij} = x_{ij} | \theta_i, \boldsymbol{\eta}_j) \tag{3.3}$$

とされる．ただし，$\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_J\}$ である．この，項目反応の条件付き独立性を局所独立性(local independence)の仮定と呼ぶ．

IRT の原点と言えるモデルとして，正規累積モデル（Lord, 1952)や上に挙げたロジスティックモデル（Birnbaum, 1968)，Rasch モデル（Rasch, 1980)が挙げられるだろう．これらの基本モデルは，いずれも正誤の二値反応を対象とし，能力母数に関する一次元性と項目反応の局所独立性を仮定するものである．受検者の特性と項目の特徴を分けて表現しておき，その上で，多数の項目を集めて予め共通の能力母数の尺度上にそれらのパラメタを推定・変換(等化)した項目プールを構築することで，難易度や測定精度を揃えた等質なテスト版を複数構成したり，受検者の能力レベルに応じて最適な項目を出題する(e.g., 適応型テストなど)といった状況でも，受検者の能力を同じ尺度上で比較可能な形で推定できるということが IRT の最大の利点である（加藤 他, 2014)．

こうした基本モデルに対しては，これまでに様々な方向で拡張がなされている（cf. van der Linden, 2016)．項目反応に関する拡張として，多値(3 つ以上の段階値やカテゴリ)の項目反応をモデル化した多値型 IRT モデル(polytomous IRT models)，連続型変数を扱うモデル(continuous IRT model; Samejima, 1973 など)がある．多値型 IRT モデルの中には，名義反応モデル（Bock., 1972)，段階反応モデル（Samejima, 1969)，部分得点モデル（Masters, 1982)，一般化部分得点モデル（Muraki, 1992)などがある．項目 j への解答が $X_{ij} \in \{0, 1, \ldots, K_j\}$ の $(K_j + 1)$ 段階で採点されるとすれば，一般化部分得点モデルの項目反応関数は，

$$P(X_{ij} = k | \theta_i, \boldsymbol{\eta}_j) = \frac{\exp(a_j \sum_{l=0}^{k} (\theta_i - b_{jl}))}{\sum_{m=0}^{K_j} \exp(a_j \sum_{l=0}^{m} (\theta_i - b_{jl}))} \tag{3.4}$$

と表される．a_j は(3.2)と同様に項目 j の識別力を表す母数である．b_{jk} は項目 j において $X_{ij} = k - 1$ の項目反応関数と $X_{ij} = k$ の項目反応関数の大小が入れ替わる θ 尺度上の点を表し，境界母数などと呼ばれる．ただし，(3.4)において便宜的に $\sum_{l=0}^{0} (\theta_i - b_{jl}) = 0$ とする．(3.4)で $a_j = 1$ と置けば部分得点モデルの項目反応関数が得られる．

能力母数を多次元に拡張したモデルとして，多次元 IRT モデル（multidimensional IRT models; Reckase, 2009)がある．多次元 IRT モデルは，正答確率のロジットが多次元の能力母数の線形結合で表される補償型(compensatory)と，正答確率が各能力母数のロジスティック関数の積で表される非補償型(non-conpensatory)の 2 種類に大別される．また，カテゴリカルな能力母数を扱うモデルとして，潜在クラスモデル（Lazarsfeld and Henry, 1968)がある．教育テストへの適用を念頭に置いたモデルとして，測定対象の概念やスキルの習得の有無を二値の潜在変数で表す状態習得モデル（state-mastery model; Macready and Dayton, 1977)や，その発展形として複数のスキルの習得パタンを推定する認知診断モデル（coginitive diagnostic models [CDMs]; Leighton and Gierl, 2007)が挙げられる．

その他にも，項目母数や能力母数に階層構造を仮定するモデル（hierarchical/multilevel IRT models; Fox, 2010)，項目特性曲線に単調増加性を仮定しない展開型 IRT モデル(unfolding IRT models; Andrich, 1988)など，項目反応や能力母数のタイプ以外の側面に関しても様々な拡張がなされている．

3.2 技術強化型項目の測定論的課題

Jiao et al. (2018)は，TEIs の特徴に起因する(IRT の適用を念頭に置いた)測定論的課題として，測定する特性(能力母数)の多次元性と，項目反応の局所依存性(local dependence)を挙げている．多次元性は，TEIs によって測定される能力・特性が多様な要素や側面を含み得ることに起因する．例えば，プロセスデータを考慮することによって，プロダクトデータだけでは捉えきれなかった側面が明らかになる可能性がある．ただし，これを元々意図していた測定の一部の要素として捉える(明示的に能力母数として扱う)か，別物(主たる測定に関する補助的な情報)と捉えるかは判断が分かれるところである．局所依存性は，測定論的には多次元性の問題と表裏一体の性質であるが，TEIs の文脈では主としてプロセスデータや，先に示した PISA の項目のように文脈を設定して一まとまりの素材に対して複数のタスクを要求するようなテストレット(大問)形式にまとめられた TEIs から生じるデータを扱う際に考慮すべき性質である．

また，現時点では TEIs から得られるデータ，特にプロセスデータからどのような情報が引き出せるのかについても十分に明らかになっているとは言えず，データを測定モデルの俎上に乗せる前の問題として，探索的な分析検討も行われている．以下，こうした観点から，TEIs によって生成されるデータを分析した研究や，測定モデルの適用を試みた研究を概観する．

3.3 探索的検討：解答形式

TEIs に関する探索的検討の研究の 1 つの分野として，TEIs ならではの解答形式を採用することによる，従来型のテストとの，あるいは TEIs 特有の解答形式(および採点方法)間の測定論的な性能比較(一種のモード効果の検討)がある．

解答形式の影響に焦点を当てた研究には，単純な項目・テスト統計量を比較するものと，IRT やその他のモデルを適用して項目母数や適合度を比較するものがある．前者の例として，Ponce et al. (2020)や Ponce et al. (2021)は，読解テストにおける「文の並べ替え」「図の整理」「cloze(穴埋め)テスト」といった解答形式の項目について，PBT 版と CBT 版(ドラッグ&ドロップ方式による解答)を比較し，テスト得点やその精度には差はないが，解答に要する時間はドラッグ&ドロップ方式の方が短いことを報告している．

IRT モデルを用いた比較検討の例をいくつか挙げると，Moon et al. (2020)は数学テストにおける多枝多選択方式(複数選択可で，当てはまる選択枝のボックスにチェックを入れる)とグリッド方式(各選択枝について「当てはまる」「当てはまらない」「わからない」などから 1 つの解答にチェックを入れる；チェックを入れるラジオボタンが選択枝を表側，解答を表頭としてグリッドの形に配置されたもの)の解答形式について，2 母数ロジスティックモデルおよび一般化部分得点モデルの母数推定値に生じた違いから，解答形式による項目反応への影響が無視できないことを報告している．また，Betts et al. (2022)は，「当てはまるものをすべて選びなさい」「当てはまるものを N 個選びなさい」という指示の下での多枝多選択方式の解答を多値採点する際の，個別の解答(選択)パタンの採点ルールの影響を検討している．具体的には，(a)二値法(選ばれた選択枝が正しい [選ばれるべき] 選択枝と完全一致するなら 1 点，そうでなければ 0 点)，(b)部分一致法(正しい選択枝のうち実際に選ばれたものの個数を得点とするが，正しくない選択枝が 1 つでも選ばれていたら 0 点)，(c)Ripkey 法(N 個選ぶという指示の場合に，N 個選んだうちの正しい選択枝の個数を得点とするが，N 個より多く選んだ場合には 0 点)，(d)加減法(選ばれた選択枝のうち，正しい選択枝の個数から正しくない選択枝の個数を引いたものを得点とする)，(e)多重真偽法(正しい選択枝を選んだ個数と，正しくない選択枝を選ばなかった個数の合計を得点とする)の 5 つの採点法によって得られる項目得点に部分得点モデルを適用し，困難度，モデル適合度，テスト情報量(測定精度)等の観点から項目およびテストの性能を比較した．その結果，多値採点，特に加減法と多重正誤法の二値採点に対する優位性

(e.g., 情報量の多さ)が示された.

Kim et al. (2022)は,第 1–12 学年対象の英語テストにおいて内容が等価な TEIs と多枝選択式項目を IRT モデルの特徴(識別力・困難度・情報量)を用いて比較した.その結果,TEIs の方が困難度が高くなる傾向があるが識別力には違いはないこと(その結果高学年・高能力層では TEIs の方が情報量が多くなる),TEIs の方がテスト時間が長くなり効率は下がることなどが示された.Ersan and Berry (2023)は,CBT で実施される数学の多枝選択式項目と,ホットスポット(画像の一部を指示する)・画像マッチ(図中の特定の場所に要素をドラッグ&ドロップする)・ドロップダウンリストによる文の穴埋め・テキスト記入による穴埋めの 4 種類の TEIs の性能を 2 あるいは 3 母数ロジスティックモデルを適用して比較し,テスト時間 1 分当たりの情報量は TEIs の方が高くなったことを報告している.解答時間や測定の効率性といった点に関しては,従来のテスト形式と比較して TEIs の解答時間が短い/長い,測定効率がよい/悪いといった一貫した結果は得られていない.インタラクティブで複雑なタスクを要求する TEIs では解答時間が長くなるのは当然であるし (e.g., Kim et al., 2022),PBT でのマーク式を CBT によるクリック式に改めれば解答時間はおそらく短縮されると考えられる (e.g., Ponce et al., 2021).

Eckerly et al. (2022)は,最初の項目への解答によって次に出題する項目を系統的に変える枝分かれ形式(branching format)の TEIs について,枝分かれを含む複数項目への解答パタンを多値採点して一般化部分得点モデルを適用したが,適合が悪かったことを報告している.多値採点に含まれた項目間で識別力が異なることが原因であるという考察がなされている.

こうした研究結果の蓄積は,従来型のテストで測定されている能力との等価性や,TEIs を用いて効率や精度に関してよりよい測定を追求するためのテストデザインを検討する際の参考になることが期待される.

3.4 探索的検討:プロセスデータ

プロセスデータが能力測定にどの程度・どのように寄与できるかということに関してはまだ十分に明らかになっているとは言えず (北條, 2023),従来のプロダクトデータによる測定を補足する,あるいはそれとは異なった側面を測定するものとして,プロセスデータから解答過程に関してどのような情報が得られるかを探索的に検討する研究が行われている.

Jiang et al. (2021)は,図 1 に示した NAEP 数学のドラッグ&ドロップ形式の項目のプロセスデータの分析を行っている.プロセスデータとして解答時のアクション(ソースをターゲットまでドラッグする/元の位置や別のターゲットに移す,解答をクリアするボタンを押す,などのマウス操作)とその順序,解答時間(総時間,最初に「止まった」タイミング,ドラッグ&ドロップ操作にかけた時間,最後に「止まった」タイミング)を収集し,これらのデータの分析から受検者の認知的・メタ認知的な思考プロセスや解答方略を推測し,解答データにもとづくテスト得点と比較することで,高得点者の解答方略がより効率的なものであることを示した.

Zhang and Andersson (2023)は,問題解決タスク(図 2 と同じく PISA 2012 で実施されたもの)のプロセスデータ(受検者の反応列と反応時間)を用いて,ネットワーク分析によって状態遷移(後述)を可視化・数値化して特徴を抽出した.用いられた特徴は(a)操作の多様性(可能な操作のうち実際に行った操作の割合),(b)辺の密度(既に行った操作の間を行き来する程度),(c)相互性(直前に行った操作を再度行う程度),(d)遷移性(ある操作に対して,その 1 つ先の操作を行ってからまたその直前の操作を行う程度),(e)外的-内的指標(一貫して正しい操作を行う程度),(f)平均時間(状態遷移にかかった平均時間),(g)時間の標準偏差(状態遷移にかかった時間のばらつき)の 7 つであり,問題解決の認知プロセス(問題の表象,計画・モニタリング,実行の 3 つ)を表現するものとしてネットワーク(特定の条件に当てはまる頂点や辺の数など)

および反応時間からそれぞれ定義された．これらの特徴に正規混合モデルを当てはめることで受検者のクラスタリングを行い，タスクへの取り組み方を反映する 6 つのグループ(低能力型，低努力型，適応型，行ったり来たり型，慎重型，試行錯誤型)を見出した．

キーストローク・ログに焦点を当てた研究として，Uto et al. (2020)がある．彼らは論述式の答案を入力する過程を記録し，そこから抽出した記入文字数，停止中，カーソル位置といった特徴量から時間区分ごとの状態(e.g., 思考中，入力中，推敲中など)を推測し，隠れマルコフモデルを用いて受検者ごとの状態遷移を分析した．状態遷移のパタンと，プロセスデータとは独立に採点した論述答案の採点結果と突き合わせることにより，答案の採点結果からだけではわからない「よい書き手」の筆記プロセスの特徴を明らかにする可能性が示唆された．

反応時間についてはそれ自体をテーマとして多数の先行研究が存在するが，プロセスデータ全般に対する期待と同様に，Molenaar (2015)は反応時間を考慮した測定モデリングの意図には，主に能力測定の精度改善と，解答過程や解答方略の推測の 2 つがあると述べている．前者の意図での測定モデリングについては，例えば van der Linden et al. (2010)がある．正誤採点されたプロダクトデータに対しては(3.2)の 3 母数ロジスティックモデルを当て，合わせて受検者 i の項目 j に対する反応時間 t_{ij} について，対数正規分布

$$p(t_{ij}|\tau_i,\alpha_j,\beta_j)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))\right]^2\right\} \tag{3.5}$$

を仮定した．すなわち，受検者 i の項目 j に対する対数反応時間の平均が項目 j の強度(intensity)母数 β_j(正に大きいほど項目 j への平均的な反応時間は長くなる)と，受検者 i のスピード母数 τ_i(正に大きいほど受検者 i の平均的な反応時間は短くなる)の差で決まり，識別力母数 $\alpha_j>0$ が大きいほど実際の反応時間が平均値まわりに集中するという形のモデルとなっている． さらに θ と τ の間に二変量正規分布を仮定することで項目反応と反応時間に関わる能力母数どうしをリンクさせ，モデルの母数推定を改善しようとするものである．近年，解答過程に関心を持って行われた研究としては，Pohl et al. (2021)が挙げられる．彼女らは，PISA 2018 の数学的リテラシーアセスメントの項目反応と反応時間に加えて無解答フラグのデータを利用し，上記の能力母数とスピード母数を含む反応時間モデルに反応傾向母数(飛ばさずに解答を行う傾向)を追加したモデル (Ulitzsch et al., 2020)を適用した．そして，これら 3 種類の母数の推定値のプロファイルを国別に比較し，能力レベルが同程度でも，解答スピードや反応傾向(テストの受検態度や受検方略)に国ごとに大きな違いがあること，能力のみによらない質の違いを見ることの重要性を見出した．

視線追跡(eye tracking)に関しても様々な検討が行われている．Man and Harring (2019)は，テスト解答中の個人の関与度(engagement)を反映する指標として固視頻度(fixation counts)のモデリング(負の二項回帰モデル)を行った．Yaneva et al. (2022)は，視線追跡データから多枝選択式項目の解答過程の検討を行っている．多数の特徴量を用いて，機械学習によって正誤解答を予測(分類)する目の動きを検討し，問題文(幹)よりも選択枝を先に見るパタンが誤答に関連し(妥当性を脅かす望ましくない解答過程)，逆に幹 → 選択枝の順に見ていく(かつ，時間をかけて読む，自信を持って選択すると考えられる)パタンは正答に関連することが明らかになっている．Maddox et al. (2018)は，OECD 国際成人力調査 PIAAC(Programme for the International Assessment of Adult Competences)の解答中の注視とサッカード(眼の素早い動き)のデータを用いて，項目内容のどこに注目してどんな順番で読み取っていくのか，テスト解答中の個人の関与度，解答方略(テキストの読み方や検算の有無など)といった解答過程に関する詳細な情報を与えると主張した．また，Mayer et al. (2023)は視線追跡による問題解決過程の分析のレビューを行っている．実際のテスティング場面において装置を使って視線追跡を

行うことは，それが認知的・心理的な面で解答行動への負担になる可能性もあることから現実的であるとは言えないかもしれない．しかし，上記の研究からは解答過程に関する様々な示唆が得られており，視線追跡を用いた実験的研究がテストの妥当性を検証する，あるいは高めるために有用な情報(e.g., 受検者が，テスト開発者が意図した解答過程を実際に経て解答しているかどうか)をもたらすことが期待される．

その他，プロセスデータの利用に関する探索的研究に関しては，北條 (2023)が機械学習を用いた試みなどいくつかの例を挙げている．

3.5 既存の測定尺度との等価性・関係性の検討

CBT の過渡期にあって，既存のテストに TEIs を追加導入する際には，既存の測定尺度との等価性が問題となる (Luecht, 2017)．これは，既存のテストの運用や設計変更，妥当性(i.e., 本来そのテストで測定したい能力が測定できているか)にも関わる重要な論点である．

TEIs の導入そのものが，それまでになかった(それまでのテストで捉えられていなかった)次元を新たに反映するものである可能性もある．真正性を高めるという目的で TEIs が導入されることで，それまで測定できていなかった能力の側面が測定できるようになるのは，それが統計的に別の次元として見なされるかどうかは別として望ましいことである．一方で，TEIs のタスクの遂行に，本来測定したい能力とは関係のない機器の操作スキルなどが過度に求められるようであれば，これは測定の妥当性に対する脅威となる (Bryant, 2017; Luecht, 2017)．

また，TEIs でよく用いられるテストレット形式の項目群をどう扱うかという問題がある．この問題は PBT における課題としても存在したが，特に現実的な状況・文脈設定をする TEIs においてより顕在化しやすいものであると考えられている．テストレット形式では，1 つの状況・文脈・素材等に対して複数の項目がぶら下がる形になるため，設定された文脈に対する親和性や事前知識などが共通してそのテストレット中の項目への反応に影響する可能性がある．また，プロセスデータ(ある操作や解答自体が次の操作や解答に影響するような一連の反応行動)を従来のプロダクトデータの反応のように扱うとしたら，おそらく個人内でも(i.e., 能力母数で条件付けても)相互依存する可能性がある．これらは IRT における局所独立性からの逸脱につながる．

Kang et al. (2022)は，TEIs においてテストレット形式および多値採点が多く用いられることを念頭に，多値型データに対応したテストレットに適用可能な IRT モデルのパフォーマンスをシミュレーションデータによって比較した．比較対象とされたモデルは，テストレットを無視して各反応を一次元・局所独立として扱う一般化部分得点モデル(GPCM)，テストレット内の項目得点の合計点を算出し，各テストレットを 1 つの多値型項目と見なして分析するテストレット多値化モデル (TPIM; Wainer and Kiely, 1987)，能力母数に加えて受検者とテストレットの変量効果交互作用を組み込んだ変量効果テストレットモデル (RTM; Bradlow et al., 1999; 実質的には多次元 IRT モデルの特殊ケースと見なせる)，RTM の交互作用項を受検者に依存しないテストレット固有の固定効果とした固定効果テストレットモデル(FTM)の 4 つであり，GPCM，TPIM，FTM のそれぞれを真のモデルとしてサンプルサイズやテストレット効果の大きさ(i.e., テストレット内の局所依存性の強さ)を操作して項目反応データを生成し，これらの 4 つのモデルを当てはめて適合度等を比較することにより，現実場面における TEIs への適用可能性が検討された．また，Jiao et al. (2018)も，TPIM と同様に受検者とテストレットの変量効果交互作用を組み込んだ多次元非補償型テストレット IRT モデルを提案している．

Luecht (2017)は，TEIs によってもたらされ得る局所従属や新規の能力軸に起因する多次元性に関して，(a)既存の尺度をそのまま使ってよい(等価な尺度と見なしてよい)，(b)基本的に多次元であるが既存の尺度と TEIs による新規尺度の混合尺度(線形結合)とする，あるいは

(c)既存尺度と TEIs による尺度を別物として扱う(多次元)かどうかを，データにもとづく統計的な検証(次元性やモデル適合度のチェック)を行い，中長期的な運用を見据えたときの尺度の一般性・継続性，そして何よりも意図された測定が実現できているか(テスト得点にもとづいて測定したい能力に関する適切な推論ができるか)という妥当性の観点から慎重に判断すべきであると述べている．

3.6 解答過程のモデリング

近年，問題解決(problem-solving)のタスクを含む TEIs におけるプロセスデータを用いた測定モデリングの研究が盛んに行われている．問題解決タスクでは，初期状態(initial state)と目標状態(goal state；目指すべき最終的な状態の指示)が明確に与えられており，受検者は目標状態を達成するために種々のアクションを実行する．図 2 の例で言えば，初期状態 S_1 において取るアクションは乗車する鉄道網の選択であり，地下鉄か広域かを選ぶことによって次の状態 S_2 に遷移する．次は料金種別の選択であり，通常料金か割引料金かによって鉄道網と料金種別の特定の組み合わせとして状態 S_3 が決まるといったように，状態 S_t でアクションを起こすと次の状態 S_{t+1} に移る．これを繰り返して，最終状態 S_T(この例では「購入」を押した状態；ただし，最終状態が目標状態と一致するとは限らず，その場合にはタスクは失敗したということになる)に至るまでのアクションの列 $A_1, \ldots, A_T$ とその結果である状態の列 $S_1, \ldots, S_T$ が得られる(これらがプロセスデータを構成する)．初期状態と最終状態の間に取りうる様々な状態を中間状態(intermediate states)と呼ぶ．明確に定義された(well-defined な)問題解決タスクでは，あり得る状態の数や遷移を経て最終状態に至るパスの数は，無用なループなどを除けば有限であり，最小のアクションで目標状態に至る最良のパスが決まっている．例えば，図 2 の例ではあり得る中間状態の数は 22 個である (Han et al., 2022)．

こうしたプロセスデータに対して，Shu et al. (2017)はマルコフ IRT モデル(Markov IRT model)を適用している．同様のアイデアは Han et al. (2022)の逐次反応モデル(sequential response model [SRM])にも採用されている．Han et al. (2022)は，能力母数 θ_i を持つ受検者 i が，状態 $S_{i,t} = s_k$ のときに(何らかのアクションを起こして)状態 $S_{i,t+1} = s_{k'}$ に遷移する確率を

$$P(S_{i,t+1} = s_{k'} | S_{i,t} = s_k, \theta_i, \boldsymbol{\lambda}) = \frac{\exp(\lambda_{kk'} + I^+_{kk'}\theta_i)}{\sum_{h \in \boldsymbol{M}_j} \exp(\lambda_{kh} + I^+_{kh}\theta_i)} \tag{3.6}$$

のような多項ロジスティック関数とするモデルを立てた．ここで，$\lambda_{kk'}$ は状態 s_k から $s_{k'}$ への「遷移しやすさ」を表す母数，$\boldsymbol{M}_k$ は状態 s_k から遷移可能な状態の集合を表す．$I^+_{kk'}$ は，状態 s_k から $s_{k'}$ への遷移がより目標状態に近づく(正しい)ものであれば 1，そうでなければ -1 を取る，予め定められた指示関数である．従って，能力母数 θ_i が正に大きければ「正しい」状態に遷移する確率が高くなる．状態遷移の確率は直前の状態にしか依存しない，すなわち一次のマルコフ性が仮定されている．この仮定により，局所独立とは行かないまでも各受検者の状態列 $\boldsymbol{S}_i$ の同時確率分布(あるいは尤度関数)を直前の状態にのみ依拠する条件付き確率の積として効率的に表現することができる：

$$L(\boldsymbol{S}_i | \theta_i, \boldsymbol{\lambda}) = \prod_{t=1}^{T_i - 1} P(S_{i,t+1} | S_{i,t}, \theta_i, \boldsymbol{\lambda}) \tag{3.7}$$

なお，このモデルではどんなアクションを取るかではなく，アクションの内容に関わらずどの状態からどの状態に遷移したのかに着目している点が特徴である．

Han et al. (2022)は SRM を図 2 のタスクのデータに適用し，良好なモデルフィットが得られたとしている．また，SRM で推定した能力母数と，創造的思考力テスト全体から正誤デー

タを元に IRT で推定した能力母数を比較したところ，0.581 の相関が得られたとしている．前者は図 2 の単独のタスクに適用したモデルから得られた得点である．仮に両者が同一の能力を測定していると仮定し，やや乱暴ではあるが古典的テスト理論に基づく項目識別力（項目得点と，全項目得点を合計したテスト得点の間の相関係数）を模した値であると見なせば，それなりに高い値が出ていると言える（IRT 得点の分散の約 34% [$\approx 0.581^2$] を単独タスクの SRM 得点で説明できる）．ただし，同一の能力を測定しているという保証はなく，先述した等価性のさらなる検討が必要であると思われる．

Han and Wilson (2022) も協働的問題解決力を測定する類似の構造のタスクに対して，同様の考え方にもとづいてモデリングを行っている．遷移確率には Han et al. (2022) と同様に多項ロジスティック関数を採用しているが，彼らが扱ったタスクの性質上，同一の状態遷移が何回でも起こり得ることを考慮し，同じ遷移の繰り返し回数を反応データとして扱ったモデルとなっている（繰り返し数が多いほど，与えられる得点が低くなるように部分得点モデルの反応確率のロジットにおける θ_i の係数を設定している）点と，この遷移モデルを反応傾向が異なると考えられる 2 つの潜在クラスにネストした（i.e., 遷移確率が受検者が属するクラスによって異なる）混合部分得点モデル（mixture partial credit model）としている点が異なる．分析の結果得られた 2 つの潜在クラスは，ほぼタスク成功群と失敗群に対応することが確かめられた．

状態遷移のマルコフ性を仮定して，プロセスデータを利用して能力推定を行うモデルとしては，他にも Lamar (2018) のマルコフ決定過程測定モデル（Markov decision process measurement model）がある．遷移カーネルには名義反応モデルを用いており，状態遷移ではなくアクションの生起確率をモデリングしている点が異なる．また，Chen (2020) は，Han et al. (2022) が扱ったのと同じ図 2 のタスクのアクション列を点過程データと見なして，連続時間動的選択測定モデル（continuous-time dynamic choice measurement model）を提案している．

これらの一連の研究は，状態遷移の確率に能力母数を組み込むことで，より適切な状態遷移を繰り返す受検者の方が高い問題解決能力を示すというメカニズムを表現するものであり，今後の問題解決タスクのプロセスデータのモデル化に対して一定の方向性を示していると言える．なお，マルコフ性の仮定とは異なった方法でプロセスデータの個人内相互依存性を表現しようとする試みも行われている．Liu et al. (2018) はプロセスレベルと受検者レベルに分けて能力母数を設定し，プロセスレベルにおいては異なる問題解決ステップを選択する受検者群を区別するために，群ごとに異なるアクション実行確率を表現する混合 IRT モデルを提案した（マルチレベル混合 IRT モデル；multilevel mixture IRT model）．

上記の一連の手法とは（目的も含めて）異なるアプローチとして，Xiao et al. (2022) は，既存尺度上での能力推定の精度向上の目的で，プロセスデータそのものを用いず，縮約した情報を能力推定に利用する手法を提案している．Xiao et al. (2022) は，PIAAC の問題解決タスクのプロセスデータ（反応行動）から多次元尺度構成法（MDS）によって特徴抽出（feature extraction）を行い，さらにランダムフォレストを適用して特徴を選択した．MDS の適用に当たっては，受検者間の反応行動列の非類似度にもとづいて次元抽出を行った．また，特徴抽出は 7 つの課題解決タスクのプロセスデータを用いて同時並行で行われ，各タスク 3 つの特徴，すなわち合計 21 の特徴量が選ばれた．こうして選ばれた各受検者の特徴量 $f_{i1}, \ldots, f_{i21}$ を用いて，能力母数 θ_i の事前分布を

$$\theta_i \sim N(b_0 + b_1 f_{i1} + \cdots + b_{21} f_{i21}, \sigma^2) \tag{3.8}$$

と置き，これを各タスクへの解答データ（一般化部分得点モデルでモデル化）と合わせて能力母数（および回帰係数）をベイズ推定することで，θ_i の推定精度（事後分散）が大幅に改善した．このアイデアは，PISA においてアセスメントの得点を算出する際に用いられる潜在回帰（latent

regression)と同様のものである(https://www.oecd.org/pisa/data/pisa2018technicalreport/;ただし,PISA ではプロセスデータの特徴量ではなく,アセスメントと同時に行われるアンケート調査の情報を縮約した変数を用いている).解答データに対する IRT モデルの部分では,解答データのみから推定した項目母数が用いられており,既存の能力尺度の意味付けを変えることなくプロセスデータの情報を利用して能力推定の精度を改善することが可能となっている.なお,認知診断モデルの文脈でも,プロセスデータ(マウスクリック&ドラッグ,反応時間)を利用することによって習得パタンの診断精度や項目パラメタの推定精度が向上したという報告がある(Liang et al., 2023).

4. 現状と今後の課題

本稿では,新しい時代のアセスメントの特徴のひとつとなるであろう TEIs の利用について,決して網羅的とは言えないながら,能力測定・測定論の立場から実践・モデリング研究のレビューを行った.TEIs による CBT のアップデートは,その利用価値の探索とモデリングの試行錯誤が並行して行われている段階にあり,近年は以前にも増して関心が高まっているテーマであると言える.TEIs の利用価値に関する探索的検討は,様々な測定領域,様々な解答形式,(プロセス・プロダクトに関わらず)TEIs が生成する様々なデータの種類に関して,今後も継続的に行われることが望ましい.測定モデルに関しては,基本的(古典的)な二値型・多値型のロジスティックモデルや多次元 IRT モデルを基本として,それらが様々な形で応用されていることがわかった.

TEIs が生成し得るデータの中で,プロセスデータには特に注目が集まっていることがうかがわれる.マルコフ IRT モデルは,状態や遷移パスの数が有限(しかもかなり少ない数)である well-defined な問題解決タスクへの適用にとどまっているが,問題解決のみならず他の内容領域のタスクへの適用可能性の検討が望まれる.

その一方で,さらに掘り下げて検討していくべき課題もいくつかある.1 つ目はモデルの効用(利用可能性)についてである.まず,問題解決タスクにおけるプロセスデータのモデリングは解答過程のモデリングであって,それがどのような能力を測定しているのか(能力母数の解釈)についてはより理解を深めていく必要があるだろう.また,次世代の「学習のためのアセスメント」の実現のためには,特定のタスクの解答過程が明らかになること(類型化できること)だけでは不十分である.個々の受検者=学習者にとって,学習の改善に資する有用なフィードバック(の元となる情報)を提供できるかという視点が必要である.

さらに挙げるとすれば,現状ではモデル設定上の制限により,モデルの適用を優先すると作成できるタスクの幅が狭くなってしまう可能性がある.これは,妥当性を制限することにもつながる.また,現状のほとんどのマルコフ IRT モデルは単一のタスクにおける解答過程のモデル化にとどまっている.通常のアセスメントでは,内容領域をなるべく網羅するために(かつ測定精度を上げるために)なるべく多くのタスクからなるべく多くの反応データを集められるようにテスト版が構成される.現状のマルコフ IRT モデルからも能力母数は推定できるが,単一のタスクから推定されたその数値が,想定している能力の推定値として十分な一般性を持つのかは疑問である.IRT による既存の能力推定とは別系統の測定として行うことも可能であるが,これはテストの利用目的によるであろう.

2 つ目の課題は TEIs を含むテストの運用についてである.測定モデリングやそれに付随する推定法の進展はめざましいものがある一方で,これらのモデルが実際のテスト運用に適うかどうかの検討は追いついていない.測定モデルは,受検者の状態を表す変数と項目の特徴を表す変数が与えられたときに解答が生成されるメカニズムを模すこと(i.e., 現象の記述)を基本と

しながらも，現実のテスト運用においてはテストの得点(能力母数の推定値)を算出するという機能を安定的・継続的に果たせることが重要である．そのためには，項目個別にモデルを立てるのではなく，多数の項目の挙動をなるべく統一的かつ単純に記述できるモデルを用いることが望ましい(一次元・局所独立の二値型 IRT モデルを安定的に運用していくだけでも並大抵のことではなく，既存尺度との関係性を重視する Luecht (2017)の視点には大いに共感できる)．マルコフ IRT モデルに限らず本稿で取り上げた他の多くのモデルは，かなり特定のタスクやデータに依存する側面が強く，上に述べた意味での実用に耐えうるのかは未知である．同じ文脈で，尺度の等化(equating)に関する研究が極めて少ない点も実用性を限定する大きな要素である．プロセスデータを加味して既存の能力測定を改善するという目的では，プロセスデータから抽出した特徴量を，既存尺度上の能力推定の際の事前分布の設定に利用するという Xiao et al. (2022)が提案したような手法は実用性が高いように思える．

「学習のためのアセスメント」の実現を念頭に置いたとき，TEIs がそのポテンシャルを十分に発揮するレベルで実用化されるまでには，解決すべき課題が数多く残されている．

参 考 文 献

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude, *Applied Psychological Measurement*, **12**, 33–51, https://doi.org/10.1177/014662168801200105.

Bennett, R. E. (2015). The changing nature of educational assessment, *Review of Research in Education*, **39**, 370–407, http://www.jstor.org/stable/44668662.

Betts, J., Muntean, W., Kim, D. and Kao, S.-C. (2022). Evaluating different scoring methods for multiple response items providing partial credit, *Educational and Psychological Measurement*, **82**(1), 151–176, https://doi.org/10.1177/0013164421994636.

Birnbaum, A. (1968). Some latent trait models, *Statistical Theories of Mental Test Scores* (eds. F. M. Lord and M. R. Novick), 397–424, Addison Wesley, Reading.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika*, **37**, 29–51, https://doi.org/10.1007/BF02291411.

Bradlow, E. T., Wainer, H. and Wang, X. (1999). A Bayesian random effects model for testlets, *Psychometrika*, **64**, 153–168, https://doi.org/10.1007/BF02294533.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests, *Practical Assessment, Research & Evaluation*, **22**(1), 1–10, https://doi.org/10.7275/70yb-dj34.

分寺杏介 (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**(1), 191–225, https://doi.org/10.24690/jart.19.1_191.

Care, E., Griffin, P. and Wilson, M. (eds.) (2018). *Assessment and Teaching of 21st Century Skills*: *Research and Applications*, Springer, Dordecht, https://doi.org/10.1007/978-3-319-65368-6.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data, *Psychometrika*, **85**, 1052–1075, https://doi.org/10.1007/s11336-020-09734-1.

Eckerly, C., Jia, Y. and Jewsbury, P. (2022). Technology-enhanced items and model–data misfit, Technical Report, RR-22-11, Educational Testing Service, Princeton, https://doi.org/10.1002/ets2.12353.

Ersan, O. and Berry, Y. (2023). Measurement efficiency for technology-enhanced and multiple-choice items in a K-12 mathematics accountability assessment, *Educational Measurement*: *Issues and Practice*, **42**(4), 19–32, https://doi.org/10.1111/emip.12580.

Fox, J.-P. (2010). Multilevel item response theory models, *Bayesian Item Response Modeling*: *Theory and Applications* (ed. J.-P. Fox), 141–191, Springer, New York, https://doi.org/10.1007/978-1-4419-0742-4_6.

Han, Y. and Wilson, M. (2022). Analyzing student response processes to evaluate success on a technology-

based problem-solving task, *Applied Measurement in Education*, **35**(1), 33–45, https://doi.org/10.1080/08957347.2022.2034821.

Han, Y., Liu, H. and Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks, *Multivariate Behavioral Research*, **57**(6), 960–977, https://doi.org/10.1080/00273171.2021.1932403.

北條大樹 (2023). CBT 領域におけるプロセスデータ利活用研究の動向, 日本テスト学会誌, **19**(1), 177–190, https://doi.org/10.24690/jart.19.1_177.

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G. and Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment, *Large-scale Assessments in Education*, **9**(2), 1–31, https://doi.org/10.1186/s40536-021-00095-4.

Jiao, H. and Lissitz, R. W. (eds.) (2020). *Application of Artificial Intelligence to Assessment*, Information Age Publishing, Charlotte, https://www.infoagepub.com/products/Application-of-Artificial-Intelligence-to-Assessment.

Jiao, H., Lissitz, R. W. and Zhan, P. (2018). A noncompensatory testlet model for calibrating innovative items embedded in multiple contexts, *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective* (eds. H. Jiao and R. W. Lissitz), 117–137, Information Age Publishing, Charlotte, https://content.infoagepub.com/files/fm/p593e6be924373/9781681239316.pdf.

Kang, H.-A., Han, S., Kim, D. and Kao, S.-C. (2022). Polytomous testlet response models for technology-enhanced innovative items: Implications on model fit and trait inference, *Educational and Psychological Measurement*, **82**(4), 811–838, http://dx.doi.org/10.1177/00131644211032261.

Kato, K. (2016). Measurement issues in large-scale educational assessment, *Annual Review of Educational Psychology in Japan*, **55**, 146–164, https://doi.org/10.5926/arepj.55.148.

加藤健太郎, 山田剛史, 川端一光 (2014).『R による項目反応理論』, オーム社, 東京.

Kim, A. A., Tywoniw, R. L. and Chapman, M. (2022). Technology-enhanced items in grades 1–12 English language proficiency assessments, *Language Assessment Quarterly*, **19**(4), 343–367, https://doi.org/10.1080/15434303.2022.2039659.

Lamar, M. M. (2018). Markov decision process measurement model, *Psychometrika*, **83**, 67–88, https://doi.org/10.1007/s11336-017-9570-0.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*, Houghton-Miffin, Boston.

Leighton, J. and Gierl, M. (eds.) (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*, Cambridge University Press, New York, https://doi.org/10.1017/CBO9780511611186.

Liang, K., Tu, D. and Cai, Y. (2023). Using process data to improve classification accuracy of cognitive diagnosis model, *Multivariate Behavioral Research*, **58**(5), 969–987, https://doi.org/10.1080/00273171.2022.2157788.

Liu, H., Liu, Y. and Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model, *Frontiers in Psychology*, **9**, 1–12, https://doi.org/10.3389/fpsyg.2018.01372.

Lord, F. M. (1952). A theory of test scores, *Psychometric Monograph*, No. 7, Psychometric Corporation, Richmond, http://www.psychometrika.org/journal/online/MN07.pdf.

Luecht, R. M. (2017). Calibrating technology-enhanced items, *Handbook of Item Response Theory, Volume 3: Applications* (ed. W. J. van der Linden), 87–103, Chapman and Hall/CRC, New York, https://doi.org/10.1201/9781315119144.

Macready, G. B. and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery, *Journal of Educational Statistics*, **2**, 99–120, https://doi.org/10.2307/1164802.

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G. and Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment, *European Journal of Psychology of Education*, **33**(3), 543–558,

https://doi.org/10.1007/s10212-018-0380-2.

Man, K. and Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items, *Educational and Psychological Measurement*, **79**(4), 617–635, https://doi.org/10.1177/0013164418824148.

Masters, G. N. (1982). A Rasch model for partial credit scoring, *Psychometrika*, **47**, 149–174, https://doi.org/10.1007/BF02296272.

Mayer, C. W., Rausch, A. and Seifried, J. (2023). Analysing domain-specific problem-solving processes within authentic computer-based learning and training environments by using eye-tracking: A scoping review, *Empirical Research in Vocational Education and Training*, **15**(1), 1–27, https://doi.org/10.1186/s40461-023-00140-2.

Molenaar, D. (2015). The value of response times in item response modeling, *Measurement*: *Interdisciplinary Research and Perspectives*, **13**, 177–181, http://dx.doi.org/10.1080/15366367.2015.1105073.

Moon, J. A., Sinharay, S., Keehner, M. and Katz, I. R. (2020). Investigating technology-enhanced item formats using cognitive and item response theory approaches, *International Journal of Testing*, **20**(2), 122–145, https://doi.org/10.1080/15305058.2019.1648270.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm, *Applied Psychological Measurement*, **16**(1), 59–71, https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

OECD (2014). *PISA 2012 Results*: *Creative Problem Solving*: *Students' Skills in Tackling Real-Life Problems* (*Volume V*), OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264208070-en.

OECD (2023). *PISA 2022 Assessment and Analytical Framework*, OECD Publishing, Paris, https://doi.org/10.1787/dfe0bf9c-en.

Pohl, S., Ulitzsch, E. and von Davier, M. (2021). Reframing rankings in educational assessments, *Measurement*, **372**(6540), 338–340, https://doi.org/10.1126/science.abd3300.

Ponce, H. R., Mayer, R. E., Sitthiworachart, J. and López, M. J. (2020). Effects on response time and accuracy of technology-enhanced cloze tests: An eye-tracking study, *Educational Technology Research and Development*, **68**(5), 2033–2053, https://doi.org/10.1007/s11423-020-09740-1.

Ponce, H. R., Mayer, R. E. and Loyola, M. S. (2021). Effects on test performance and efficiency of technology-enhanced items: An analysis of drag-and-drop response interactions, *Journal of Educational Computing Research*, **59**(4), 713–739, https://doi.org/10.1177/0735633120969666.

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest?, *Large-Scale Assessment in Education*, **9**, 1–17, https://doi.org/10.1186/s40536-020-00092-z.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, expanded edition, The University of Chicago Press, Chicago.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*, Springer, New York, https://doi.org/10.1007/978-0-387-89976-3.

Russell, M. and Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs, *Educational Assessment*, **24**(4), 286–304, https://doi.org/10.1080/10627197.2019.1670055.

Rychen, D. S. and Salganik, L. H. (eds.) (2003). *Key Competencies for a Successful Life and a Well-Functioning Society*, Hogrefe, Cambridge.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika*, **17**, 1–100, https://doi.org/10.1007/BF03372160.

Samejima, F. (1973). Homogeneous case of the continuous response model, *Psychometrika*, **38**, 203–219, https://doi.org/10.1007/BF02291114.

Shu, Z., Bergner, Y., Zhu, M., Hao, J. and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks, *Psychological Test and Assessment Modeling*, **59**, 109–131, https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/07_Shu.pdf.

Ulitzsch, E., von Davier, M. and Pohl, S. (2020). Using response times for joint modeling of response and omission behavior, *Multivariate Behavioral Research*, **55**(3), 425–453, https://doi.org/10.1080/00273171.2019.1643699.

U.S. Department of Education (2017). Reimagining the role of technology in education: 2017 national education technology plan update, Technical Report, U.S. Department of Education, Office of Educational Technology, Washington, D.C., https://tech.ed.gov/files/2017/01/NETP17.pdf.

Uto, M., Miyazawa, Y., Kato, Y., Nakajima, K. and Kuwata, H. (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, https://doi.org/10.1007/s40593-019-00189-9.

van der Linden, W. J. (ed.) (2016). *Handbook of Item Response Theory, Volume 1*: *Models*, Chapman and Hall/CRC, New York, https://doi.org/10.1201/9781315374512.

van der Linden, W. J., Entink, R. H. K. and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information, *Applied Psychological Measurement*, **34**(5), 327–347, https://doi.org/10.1177/0146621609349800.

Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement*, **24**, 185–201, http://www.jstor.org/stable/1434630.

Wools, S., Molenaar, M. and Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—Threats and opportunities, *Theoretical and Practical Advances in Computer-based Educational Measurement* (eds. B. P. Veldkamp and C. Sluijter), 3–19, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-18480-3_1.

Xiao, Y., Veldkamp, B. and Liu, H. (2022). Combining process information and item response modeling to estimate problem-solving ability, *Educational Measurement*: *Issues and Practice*, **41**(2), 36–54, https://doi.org/10.1111/emip.12474.

Yaneva, V., Clauser, B. E., Morales, A. and Paniagua, M. (2022). Assessing the validity of test scores using response process data from an eye-tracking study: A new approach, *Advances in Health Sciences Education*, **27**(5), 1401–1422, https://doi.org/10.1007/s10459-022-10107-9.

Zhang, M. and Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times, *Educational Assessment*, **28**(3), 172–189, https://doi.org/10.1080/10627197.2023.2222585.

Review of Measurement Models for Technology-enhanced Testing

Kentaro Kato

Benesse Educational Research & Development Institute

As computer-based testing (CBT) has become a major mode in administrating educational assessments, its potential advantages over traditional paper-based testing (PBT) draw attention. Among them are the technology-enhanced items and their use for the purpose of improved measurement of learning status. Given this situation, this paper reviews recent attempts and practices in the context of exploratory analysis and psychometric modeling of various types of data generated from TEIs. The review identified the following research topics that are actively pursued: (a) effects of innovative response formats on the psychometric properties, (b) exploratory analysis of the utility of process data, (c) equivalence between the traditional measurement scale and the new scale involving TEIs, and (d) psychometric modeling of process data for revealing response processes. In terms of the last point, it was pointed out that further considerations will be necessary to produce measurement results more useful for learning and for sustainable and stable operation of tests that involve TEIs.

Key words: Computer-based testing, item response theory, technology-enhanced items, process data.