統計数理(2024)
第 72 巻 第 1 号 79–92

特集「心理統計学の新展開」
［総合報告］

項目反応理論を用いた症状評価項目バンクの現状と今後の課題

国里 愛彦[1]・竹林 由武[2]

(受付 2023 年 6 月 30 日；改訂 10 月 20 日；採択 12 月 6 日)

要　　旨

患者報告アウトカムは，質問紙などを用いて患者から直接得られる健康状態に関する報告であり，臨床試験のアウトカムとしての利用も多い．患者報告アウトカム尺度は多数開発されているが，PROMIS®(Patient-Reported Outcomes Measurement Information System)では，項目反応理論によるコンピュータ適応型テストを想定した項目バンクの開発が行われている．コンピュータ適応型テストにより，測定精度を保ったまま質問数を減らすことができ，回答者の負担を減らすことができる．PROMIS®では，項目プールの開発，項目の心理測定学的検討，妥当性の検討の順番で尺度開発をすすめ，その開発にあたって科学的に妥当な基準を設定している．具体的には，(1)対象の構成概念の定義，(2)項目の構成，(3)項目プールの構築，(4)項目バンクの性質の特定，(5)検査のフォーマット，(6)妥当性，(7)信頼性，(8)解釈可能性，(9)翻訳と文化適応の 9 つの基準である．本論文では，PROMIS®での尺度開発プロセスについて解説をするとともに，日本国内において患者報告アウトカムの項目バンクを開発する上で必要となることについて議論した．

キーワード：患者報告アウトカム，項目反応理論，項目バンク，PROMIS®，COSMIN．

1. はじめに

心理統計学の臨床領域への応用として，患者報告アウトカム(patient-reported outcome [PRO])の開発と運用を挙げることができる．アメリカ食品医薬品局(Food and Drug Administration [FDA])によると，患者報告アウトカムは「患者から直接得られる患者の健康状態に関する全ての報告であり，患者の回答に対して臨床医や他の誰の解釈も介さないもの」とされる (Food and Drug Administration, 2009)．患者報告アウトカムの測定方法には，面接による測定方法もあるが，多くの場合は自己記入式質問紙による測定方法が用いられる．FDA が指針を出してから，臨床試験における患者報告アウトカムの活用が増え，その開発と運用も活発化している．日本国内でも患者報告アウトカムに対する関心は高くなってきており，「患者報告アウトカム(Patient-Reported Outcome: PRO)使用についてのガイダンス集」も作成されている (下妻 他, 2023)．

患者報告アウトカムの運用方法として，項目反応理論(item response theory [IRT])を用いたコンピュータ適応型テスト(computerized adaptive test [CAT])がある．紙とペンを用いて実施

[1] 専修大学 人間科学部：〒 214–8580 神奈川県川崎市多摩区東三田 2–1–1
[2] 福島県立医科大学 医学部：〒 960–1295 福島県福島市光が丘 1 番地

される質問紙とは異なり，コンピュータ適応型テストでは，コンピュータを用いることで参加者の反応に応じて呈示する項目を調整する．コンピュータ適応型テストを用いると，測定精度を保ったまま質問数を減らすことができ，回答者の負担を減らすことができる．コンピュータ適応型テストを用いて患者報告アウトカムを測定する場合，多数の尺度項目から構成される項目バンクを作成する必要がある．コンピュータ適応型テストを想定した患者報告アウトカムの項目バンクの開発においては，PROMIS®(Patient-Reported Outcomes Measurement Information System)の取り組みが参考になる．そこで，本論文では，PROMIS®で提案されている患者報告アウトカム尺度の項目バンク開発について整理し，国内における患者報告アウトカム尺度の項目バンクを開発する際の課題について議論する．

2. PROMIS®とは

PROMIS®は，米国国立衛生研究所(National Institutes of Health [NIH])が 2002 年に打ち出した医学研究のためのロードマップに従って，2004 年に NIH 主導の多施設共同研究グループとして立ち上げられたプロジェクトである（Cella et al., 2007）．Duke 大学，Stanford 大学，Stony Brook 大学，North Carolina 大学，Pittsburgh 大学，Washington 大学などの 6 つの研究拠点と統計調整センター(statistical coordinating center [SCC])を中心に構成されており，国家的なプロジェクトになる．PROMIS®は，様々な慢性疾患の症状や健康状態を測定するための項目バンクを構築・検証し，臨床現場で効率的に利用できるようにすることを目的としている（Cella et al., 2007）．その項目バンクの活用では，項目反応理論を用いたコンピュータ適応型テストを用いる．PROMIS®のウェブサイト(https://www.healthmeasures.net/explore-measurement-systems/promis)では，実際のコンピュータ適応型テストのデモも用意されており，ブラウザ上で回答に対する推定結果をその場で確認することができる．

PROMIS®では，プロジェクトの最初の段階で，慢性疾患にかかわる様々な領域(がん，リウマチ，精神保健など)の専門家，臨床試験の専門家，製薬業界などのステークホルダーを集めた専門家パネルを構成した（Cella et al., 2007）．そして，専門家パネルは PROMIS®がターゲットとするドメインを定め，その概念枠組みについても検討した（Cella et al., 2007）．初期の PROMIS®のドメインは，身体機能，疲労，疼痛，感情的苦痛，社会役割参加の 5 つであったが，現在は，図 1 にあるように身体的健康，精神的健康，社会的健康に大きく分けた上で，それぞれにプロフィールドメインと追加ドメインが置かれている[1)]．これらのドメインを測定する項目バンクの開発と検証については，PROMIS®が作成している検査開発と妥当化のための科学的基準（PROMIS, 2013)に従って説明する．

3. PROMIS®における患者報告式アウトカム尺度開発

PROMIS®では，項目プールの開発，項目の心理測定学的検討，妥当性の検討の順番で尺度開発を進める(図 2)．PROMIS®では，検査の開発と妥当化のための科学的基準（PROMIS, 2013）を定めており，(1)対象の構成概念の定義，(2)項目の構成，(3)項目プールの構築，(4)項目バンクの性質の特定，(5)検査のフォーマット，(6)妥当性，(7)信頼性，(8)解釈可能性，(9)翻訳と文化適応の 9 つの基準がある．以下ではそれぞれの基準について解説する．

3.1 (1)対象の構成概念の定義

PROMIS®の各ドメインで測定される構成概念は，明確に定義される必要がある（PROMIS, 2013）．構成概念の定義にあたり，まずは既存の文献レビューを行って，理論的・実証的観点から検討を行う．Pittsburgh 大学の PROMIS®グループは，包括的な文献検索方法の開発のた

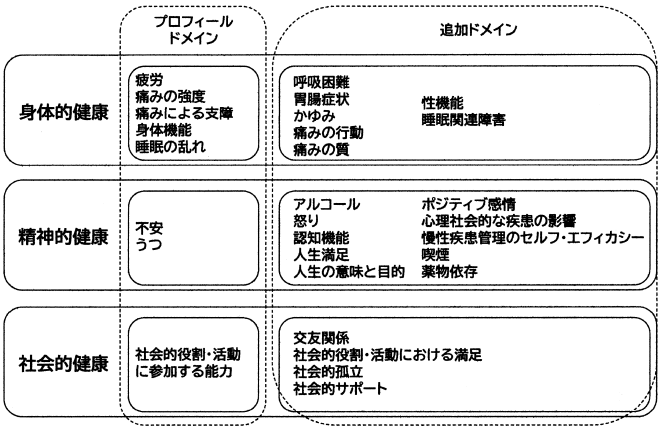

図 1. PROMIS®の成人用ドメインの構成.

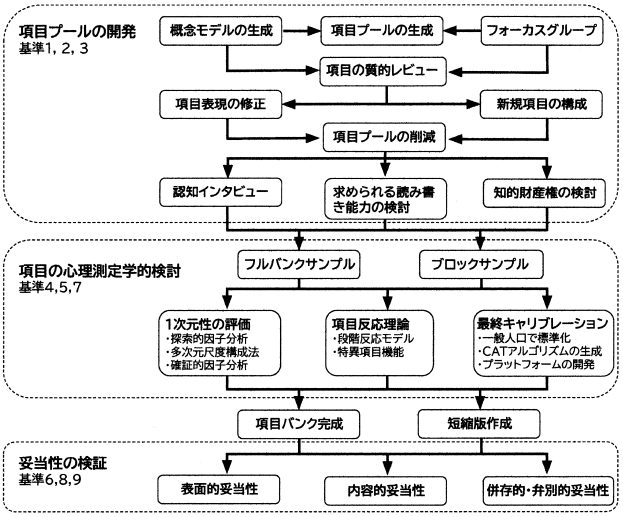

図 2. PROMIS®開発フロー．Pilkonis et al. (2011)の Figure 1 を参考に一部改変.

めに，図書館司書と協働している（Klem et al., 2009）．図書館司書が研究チームに加わることで，より網羅的に既存の文献を見つけることができ，その概念の研究領域での位置づけなどについての情報を得ることができる[2)]．

対象の構成概念は，その構成概念や測定の専門家，臨床家，患者，ユーザー，関連するステークホルダーによって，適切な質的研究法によるレビューが行われる必要がある（PROMIS, 2013）．まず，専門家パネルが暫定的な構成概念の定義について精査する．その具体的な実施方法としては，修正型デルファイ法が推奨されている（PROMIS, 2013）．修正型デルファイ法では，匿名調査を用いて専門家間のコンセンサスを得ることができる．以降の項目作成時のフォーカスグループでのフィードバック，得られたデータの統計解析(因子分析やモデル適合度)，項目作成後の専門家パネルの精査などを通して，構成概念の改訂が行われる．

3.2 (2)項目の構成

PROMIS®の項目を作成にするにあたり，測定する概念に関する既存の尺度と項目を包括的に文献検索して収集する（Klem et al., 2009）．収集した項目は，構成概念ごとにグループ化する「分配(binning)」を行う（DeWalt et al., 2007）．これにより，構成概念を捉えるのに必要な項目と冗長な項目を特定することができる．さらに，ドメインの定義にあわない項目や冗長な項目などを除外する「より分け(winnowing)」を行う（DeWalt et al., 2007）．「分配」も「より分け」も 1 名ではなく 2 名以上の研究者で取り組む．「分配」や「より分け」によって整理された項目に対して，文法的な正しさ，読みやすさ，翻訳可能性の観点から改訂する（DeWalt et al., 2007）．読みやすさに関しては，12 歳程度の読み書き能力があれば理解できる項目にすること，曖昧な表現は避けること，スラングは避けてシンプルな表現にすること，1 つの項目に複数の質問が含まれるダブルバーレルは避けることなどに留意する（DeWalt et al., 2007）．また，米国での使用を想定しているため，多様な文化的背景をもった回答者が回答しやすいように，他言語への翻訳可能な項目になるように調整される．具体的には，特定の国，文化圏，社会階層のみで通用する表現などは避ける．

項目の改訂は，専門家によってなされるだけでなく，認知インタビューを通しても行われる（PROMIS, 2013）．認知インタビューは非専門家や患者を対象に，項目が意図したような意味で理解されているのか，項目は理解しやすいか，どのように回答するのかを検討するために行われる（DeWalt et al., 2007）．PROMIS®では，各項目に対して，最低 5 名の参加者による認知インタビューが行われ，途中で大きな改訂がなされた場合はさらに 3 名から 5 名を追加して精査する（DeWalt et al., 2007）．実施方法には，参加者に項目を読んで回答するように依頼し，その後で，面接によって各項目の理解度や分かりにくい表現などを確認する回顧的プロービング(retrospective probing)法を用いる．認知インタビューでは参加者の多様性を保証するために，高校を卒業していない者，読み書きの水準が 15 歳以下の者，認知障害のある者などを含めたり，地域や人種的な多様性にも配慮する（DeWalt et al., 2007）．

3.3 (3)項目プールの構築

項目の収集・改訂が進み項目プール候補ができてきたら，項目プールが対象の構成概念を漏れなく捉えられているか検討する（PROMIS, 2013）．心理学や精神医学では，これまで様々な患者報告アウトカム尺度を開発しているが，同じ構成概念であっても，項目内容が異なることがある．例えば，主要な抑うつ症状尺度に対する調査では 52 個の抑うつ症状がリストアップされたが，1 つの尺度で測定しているのはそれらの内の最大 40% しかないという指摘もある（Fried et al., 2022）．既存の 1 つの尺度で測定できる範囲は限定的になる．項目プールが対象の構成概念を漏れなく捉えられているか検討する方法には，(1)構成概念の各側面をリスト化して対応する項目があるか検討する，(2)患者を対象としたフォーカスグループによって検討するの 2 種類がある．PROMIS®でのフォーカスグループは，3 つ以上のグループで，各グループが 6 から 12 名の患者，ファシリテーター，記録係で構成される（PROMIS, 2013）．フォーカスをあてたグループにするために，ある一定の基準を満たした患者に参加を依頼する．フォーカスグループでは，議論方法は構造化されておらず，自由な発言を促される．フォーカスグループを通して，患者の目線から，提示された項目が構成概念をまんべんなく測定できているか検討され，必要に応じて新規項目の追加や修正がなされる．

3.4 (4)項目バンクの性質の特定

項目バンクが作成できたら，項目の心理測定学的性質を回答者の代表的サンプルに基づいて決定する（PROMIS, 2013）．2006 年から 2007 年にかけて実施された PROMIS®の第 1 波調査

では，米国の一般人口と複数の患者集団からデータを収集している（Cella et al., 2010）．この調査では，世論調査会社を使って 21133 名からデータを収集した．また，米国国勢調査を参考にしてサンプルマッチング法を用いて代表性を担保している．次に，収集したデータから得られる各項目の心理測定学的性質を検討する．この検討を通して，項目バンクには優れた心理測定学的性質をもつ項目が残される．まず，項目反応理論の適用の前に，各項目に対して項目分析を実施する．項目分析では，(1)反応頻度，平均値，標準偏差，得点範囲，尖度，歪度などの記述統計量の検討，(2)項目間相関行列，項目-尺度相関，項目減による Cronbach の α 係数の変化などの内的整合性に関する検討が行われる（PROMIS, 2013）．

PROMIS®では項目反応理論を用いるが，項目反応理論適用の前提条件(一次元性，局所独立性，単調性)を満たすかどうかを確認する．一次元性の確認では，最初に 1 因子モデルの確証的因子分析による検討を行い，適合度の評価を行う（PROMIS, 2013）．なお，患者報告アウトカム尺度の回答は順序データと考えられるので，因子分析にあたってポリコリック相関係数を用いる．もし，確証的因子分析の適合度が低い場合は，探索的因子分析によって 1 因子構造が得られるか確認する（PROMIS, 2013）．探索的因子分析で 1 次元性を確認する目安の一つとして，第一因子の分散説明率が 20% 以上でかつ第二因子との分散説明率の比が 4 以上であることが挙げられる．下位因子を想定した上で全項目を説明する一般因子を想定する双因子(bifactor)モデルを検証することが適切な場合も存在する．双因子モデルで一般因子を想定することの合理性を判断するために共通分散説明率(explained common variance [ECV])が利用できる（Reise et al., 2013）．局所独立性は，個人の特性で条件づけられた際の項目間の反応が独立していることを指す．具体的には，ある項目が他の項目の反応に依存するものになってないか検討したり，1 因子モデルの確証的因子分析の残差相関行列から検討したりすることができる（PROMIS, 2013）．単調性は，健康に関する特性がより高いほど，健康に関する各項目への反応はより高くなることを指す．具体的には，合計得点から項目得点を引いた得点に条件づけられた項目平均得点のプロットや，ノンパラメトリック項目反応理論モデル(例えば，Mokken 尺度分析)によって検討することができる．Mokken 尺度分析で単調性を検討する場合には，各項目と尺度項目全体について推定されるスケーラビリティ係数(scalability coefficients)H を基準とすることができる（Mokken, 1971）．一次元性，局所独立性，単調性は，項目反応理論の適用の条件となるため，項目バンク内の項目がこれらを満たすか確認してから，項目反応理論モデルを適用する必要がある．また，一次元性の検討などを通して，項目反応理論モデルの適用の前に項目プールから項目を減らし，整理することもできる（Reeve et al., 2007）．

PROMIS®では，一次元性が仮定できる多肢選択式項目が基本になるので，段階反応モデル(Samejima, 1969)が用いられる．段階反応モデルで推定されるパラメータは，個人の特性値 θ，各項目の識別力 a と困難度 b になる．特性値 θ，識別力 a，困難度 b の下で，反応がカテゴリー k(多肢選択の 1 つ)である確率が，以下の式を用いて計算される（PROMIS, 2013）．

$$P\left(X_i = k \mid \theta, b_i, a_i\right) = \frac{1}{1+\exp\left[-a_i\left(\theta - b_{i,k-1}\right)\right]} - \frac{1}{1+\exp\left[-a_i\left(\theta - b_{i,k}\right)\right]}$$

項目反応理論モデルに対してもモデル適合度の評価を行う．モデル適合度の評価には，観測反応頻度と期待反応頻度の比較や適合度指標(χ^2，尤度比 G^2)を用いる．項目反応理論モデルの推定結果は，カテゴリー反応曲線や項目情報曲線から解釈をする．項目反応理論の観点から項目を検討し，性能が低い項目については，専門家パネルが内容のレビューをした上で項目バンクに残すかどうか判断する（Hansen et al., 2014）．性能は低いが，臨床的な関連性が高い項目は，保持されたり，修正されたりする．また，PROMIS®で用いる基本的な項目反応理論モデルでは 1 因子構造を想定しているが，実際の項目プールの構造は 1 因子構造とは限らない．そ

のため，多次元項目反応理論モデルを用いることも適宜検討される（PROMIS, 2013）．

PROMIS®では，多様な背景をもった患者の患者報告アウトカム尺度を開発するため，特定の集団において他の集団とは異なる機能をもった項目を特定する必要がある．このような特定の集団（ジェンダー，年齢，疾患，教育，人種など）に対して項目が特異的な機能を有する性質を特異項目機能（differential item functioning [DIF]）と呼ぶ．特異項目機能のある項目があると，特定の集団に関して過小もしくは過大評価をしてしまうため，現状では PROMIS®では特異項目機能のある項目は項目プールから除外する．特異項目機能については，事前に仮説を設定した上で検討する．特異項目機能の特定は，項目反応理論に基づいて尤度比検定や項目反応理論には基づかない順序ロジスティック回帰，あるいは構造方程式モデリングの多重指標多重原因（Multiple indicators multiple causes [MIMIC]）モデルや確証的因子分析モデルの多母集団同時分析を用いた方法が代表的である（Teresi et al., 2021）．例えば，MIMIC モデルを用いた方法では，項目と測定概念を反映する潜在変数からなる確証的因子分析モデルに集団の属性を反映する項目を加え，属性項目から潜在変数とすべての尺度項目にパスを追加する．そこから修正指数等を参照し，属性項目が尺度項目に有意な影響性を示す項目を特異項目機能を示す項目と判断する．そして，特異項目機能の大きさと影響を考慮して，特異項目機能のある項目を項目プールから除外する．なお，項目プールから除外せずに，特異項目機能も考慮した上で推定することも考えられるが，現状では採用されていない．

PROMIS®の項目プール全体には多数の項目が含まれているので，参加者に全ての項目に回答を求めるのは難しい．例えば，PROMIS®の第 1 波調査では，項目プール全体で 1000 項目以上が含まれており，1 名が全ての項目に回答するのは難しかった（Cella et al., 2010）．そこで，フルバンクデザインとブロックデザインの 2 種類を実施し，1 人の回答者が回答するのは 150 項目程度に限定して調査を実施している．フルバンクデザインは，参加者に PROMIS®の特定のドメイン内の全ての項目に回答を求める調査デザインである．これによって項目反応理論の適用にかかわる一次元性の評価とドメイン内の項目のキャリブレーションが可能となる．また，既存の尺度も一緒にデータ収集することで，既存の尺度との等化も可能となる（Thompson et al., 2017）．例えば，既存の抑うつ尺度である Patient Health Questionnaire(PHQ)-9 と PROMIS®のうつ尺度でキャリブレーションを実施した研究から，9 項目の PHQ-9 よりも PROMIS®の適応型テストの 3 項目の方が測定誤差が小さいことが示されている（Gibbons et al., 2011）．ブロックデザインでは，複数のドメインの一部の項目に回答をすることで，ドメイン間の関係の評価が可能となる．ブロックデザインでは，項目のブロックごとに一般集団から 900 名以上，500 名の慢性疾患患者を含むように調査が設計された（Cella et al., 2010）．項目バンクには項目に関するパラメータが明らかになった項目が収載され，コンピュータ適応型テストで利用される．

3.5 (5)検査のフォーマット

PROMIS®の項目バンクの活用法としては，参加者の反応に応じて質問を提示するコンピュータ適応型テストや，予め定められた項目から構成される項目バンクの短縮版尺度などがある．それぞれの検査形式は，使用目的や項目バンクの性質に合わせて設定される（PROMIS, 2013）．また，検査として使う上では，各尺度の検査としての適切な性質を示し，回答者の負担についても検討を行う．回答者の負担は回答にかかる時間や回答数などが含まれる．パソコン上での回答や紙とペンによる回答などの異なる実施方法での比較可能性についても検討する．

PROMIS®等で用いられるコンピュータ適応型テストの基本的なアルゴリズムを図 3 に示す．コンピュータ適応型テストは，項目反応理論を通じてキャリブレーションされた項目バンクから特定の能力値（θ）を任意に選択し，その能力値に適した項目の選択と提示が行われる．回答者

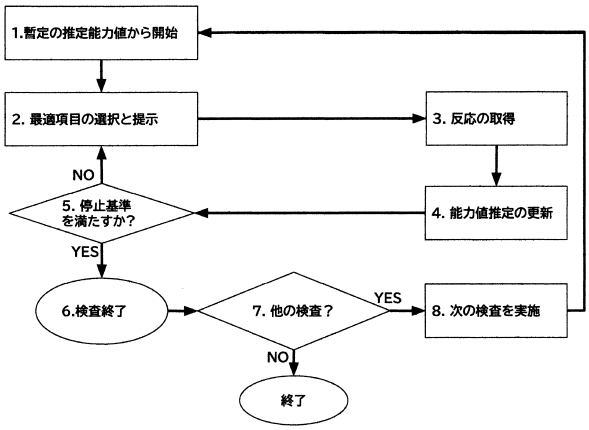

図 3. 適応型テストの基本アルゴリズム．İnce Araci and Tan (2022)を参考に一部改変.

の θ に関する事前情報がない場合には，母集団の θ の平均値を割り当て，それに適した項目の提示から開始する，あるいは項目バンクに含まれる最初の項目を提示するといった方法が一般的である．最初に提示した項目への反応に基づいて θ の推定を更新する．θ の推定は，最尤法またはベイズ推定による期待事後推定量(Expected a Posteriori Estimator [EAP])や最大事後確率推定量(Maximum a Posteriori Estimator [MAP])がよく用いられる（Lord and Novick, 2008; Segall, 1996）．能力値 θ の更新後，次の質問項目を提示する必要性を停止基準に従って判断する．停止基準には標準誤差がよく用いられる．標準誤差が一定の水準まで小さくなった時点で項目の追加提示を止め，その時点の能力値を求めることで精度の高い推定値が得られる．能力値の更新後に次に提示する項目を選択する方法には，A ルール，E ルール，D ルール，T ルール，W ルール，Kullback-Leibler 情報量規準(KL)，連続エントロピー法などがある（İnce Araci and Tan, 2022）．これらの規準は最終的な能力値推定における規準値を最大化または最小化するという点で共通しているが，規準値の定義が異なる．例えば，KL 法と連続エントロピー法は，それぞれ事後期待 KL 情報を最大化し，期待連続エントロピーを最小化する（Mulder and van der Linden, 2009）．

3.6 (6)妥当性

PROMIS®の妥当性検討では，基準関連妥当性，構成概念妥当性，内容的妥当性，反応性を検討する（PROMIS, 2013）．基準関連妥当性では，基準となる尺度を定めて，その尺度との関連の強さについて事前に仮説を立てて検証する．PROMIS®の場合，基準としては既存の尺度や診断などのゴールドスタンダードとなる基準が挙げられる．構成概念妥当性では，測定する構成概念とその他の構成概念との間の関係について事前に仮説を立てて検証する（Cella et al., 2010）．具体的には，測定する構成概念と関連する構成概念との関係について検証する収束的妥当性，測定する構成概念と関連しない構成概念との関係について検証する弁別的妥当性について検討する．内容的妥当性の検討は項目の作成の段階でも行われてきているが，項目反応理論によるキャリブレーション後にも検討する（Riley et al., 2010）．特に，項目反応理論の適用などに際して項目の絞り込みがなされている場合に，当初測定する予定だったドメインと項目バンクの内容に乖離が生じる可能性がある．内容的妥当性では，最初に定義したドメインと作成された検査との間の一貫性について検討する．

反応性は構成概念の時間的な変化に関する妥当性であり，患者の変化を尺度が捉えることが

できているのかを検討する（PROMIS, 2013）．そのため，ある程度の変化が期待できる時間間隔でデータを測定し，変化が期待できるグループと安定していると期待できるグループで比較したり，医師や患者による変化に関する評定のような外部アンカーとの関連を検討したりする．例えば，PROMIS®のうつ尺度に関する反応性を調べた研究では（Pilkonis et al., 2014），うつ病患者の治療初期から 3 ヶ月後までのフォローアップ調査を行って，治療に伴う症状変化を測定した．その結果，PROMIS®のうつ尺度は，各測定時点の既存の尺度の PHQ-9 と CES-D（Center for Epidemiological Studies Depression）と相関しており，変化に関する収束的妥当性を示した．また，患者による全般的改善の評価（「非常に改善」から「悪化」で評価）をアンカーとして，そのアンカーと PROMIS®のうつ尺度の変化が対応することも確認している．

3.7 （7）信頼性

PROMIS®では，対象構成概念の信頼性と再検査信頼性を検討する（PROMIS, 2013）．対象構成概念の信頼性については，古典的テスト理論の場合は尺度全体の Cronbach の α 係数や測定誤差から検討できる．項目反応理論の場合は特性値を横軸においてテスト情報量をプロットしたテスト情報曲線を描くことができ，構成概念の特性値の全範囲における信頼性を検討することができる．対象構成概念の再検査信頼性については，2 時点の間での測定値の相関を検討する．その際には，2 時点の間で対象集団に大きな変化はなく安定していることが想定できるようにデータ収集する必要がある．

3.8 （8）解釈可能性

PROMIS®では回答から連続的な得点が得られるが，その得点を臨床的な意味として解釈しやすい質的なカテゴリーに分けることで解釈可能性を高めることができる．PROMIS®では，最小限の重要な差（minimally important difference [MID]）を示すことが推奨されている．最小限の重要な差の計算方法としては，尺度の測定誤差などの分布に基づく方法と患者や医療者による変化に関する外的アンカーに基づく方法がある（宮崎, 2015）．なお，時間的な変化に関しては，最小限の重要な差ではなく，最小限の重要な変化（minimally important change [MIC]）と呼ぶことがある（Mokkink et al., 2010）．

妥当性，信頼性，解釈可能性については，PROMIS®の基準やその付録も参考になるが，それらの検討方法は COSMIN（COnsensus-based Standards for the selection of health Measurement INstruments）が詳しい．COSMIN では，患者報告アウトカムの選択・評価に関するガイドラインを作成しており（Mokkink et al., 2010; Prinsen et al., 2018; 佐藤・土屋, 2022），それらは尺度開発においても有用である．COSMIN のガイドラインは PROMIS®の基準と内容的に重複する部分もあるが，COSMIN は格付けのための明確な規格を持ったチェックリストを提供するため，患者報告アウトカム尺度を開発する場合にも活用することができる．

3.9 （9）翻訳と文化適応

PROMIS®は，英語で項目バンクを作成しているが，PROMIS®の項目バンクを他言語に翻訳し，活用することもできる．PROMIS®では，教示，項目，選択肢について厳密な翻訳過程を通して他言語への翻訳が行われる（PROMIS, 2013）．具体的には，英語から他言語への順翻訳と翻訳した他言語から英語への逆翻訳をおこなって，他言語への翻訳によって項目の意味が異なってないか確認する．順翻訳と逆翻訳に関しては，繰り返し実施して，適切な翻訳になるように調整し，そのプロセスにおいてはバイリンガルの専門家による検討も推奨される（PROMIS, 2013）．また，逆翻訳を想定して翻訳された項目は，読みにくかったり，分かりにくかったりすることもあるため，認知インタビューによる検討を通してその文化にあった項目

を開発することが重要になる．なお，PROMIS®では，尺度翻訳にかかわる最小限の要件を設定している（PROMIS, 2014）．この最小限の要件としては，リリース前の要件（適切な翻訳，特異項目機能を含む予備検討）とリリース後の要件（項目バンクのキャリブレーション，その言語における参照得点，信頼性，妥当性，反応性，解釈可能性）が設定されている．

4. PROMIS®の検査成熟モデル

上記の 1 から 9 の基準にしたがって PROMIS®の項目は作成されるが，PROMIS®では検査としての成熟モデルを設定し，各検査がどの開発段階か分かるようにしている（PROMIS, 2012）．具体的には，ステージ 1（開発：概念化と項目プール），ステージ 2（開発：キャリブレーション），ステージ 3（一般公開：キャリブレーションと予備的な妥当性検討），ステージ 4（成熟化：反応性と拡張），ステージ 5（完全に成熟したユーザーサポート）の 5 ステージからなる．ここまで説明をした基準にもとづいて，ステージ 1 からステージ 3 まで開発し，一般公開する．その後，ステージ 4 において，異なる臨床集団に対して実施したり，反応性や解釈可能性について検討し，より検査を使いやすくする．最後のステージ 5 は，様々な集団で検討され解釈も可能となった状態であり，ユーザーが利用しやすいように採点と解釈マニュアルなどを作成する．PROMIS®の検査成熟モデルを用いることで，患者報告アウトカム尺度の開発を，プロダクトの開発プロセスのように実施することができ，プロジェクトの進行確認などにも有用と思われる．

5. PROMIS®の現状と今後の課題

PROMIS®は，項目反応理論を用いたコンピュータ適応型テストをベースにした患者報告アウトカム尺度の開発を行っている．患者報告アウトカム尺度や精神症状を測定する尺度は，紙とペンを用いたものが多く，臨床現場で実施する場合は採点などを行う必要があった．また，毎回同じ項目を用いることによって患者があまり深く考えずに回答することもあるかもしれず，項目数などの負担も大きい．一方で PROMIS®は，コンピュータ適応型テストによって，患者の反応に合わせて項目を提示し反応を取得するため，患者の負担は小さくなり採点も自動化されるため実施者の負担も小さくなる．また，実施や採点に対する負担だけでなく，項目反応理論に基づく推定値は古典的テスト理論に基づく得点よりも精度が高くなる．例えば，慢性的な腰痛や首の痛みに関する治療上の変化について，古典的テスト理論に基づくものよりも，PROMIS®の項目反応理論に基づく変化の推定のほうが正確になるとされる（Hays et al., 2021）．PROMIS®のような患者報告アウトカムの項目バンクの作成は，臨床的にも有用であり，今後も利用の拡大が予想できる．

当初，PROMIS®の項目は英語で作成されたが，米国内の英語以外の話者（スペイン語と中国語）も使えるように他言語への翻訳も進められている．また，米国内だけでなく，米国以外の国で活用するための翻訳も行われている．ノルウェー語版 PROMIS®（Rimehaug et al., 2022）のように翻訳と検証プロセスを論文化しているものもあるが，それ以外にも PROMIS®のウェブサイトから多くの言語に翻訳されていることが確認できる．痛み，うつ，不安，疲労，身体活動，睡眠，イライラなどの PROMIS®尺度については，既に日本語訳されていることも分かる．

PROMIS®を日本でも活用することができれば，臨床研究と臨床実践の両面への多大な貢献が期待できる．しかし，英語とスペイン語については無料で利用できるが，その他の言語に関しては手数料が必要とされる（PROMIS, 2023）．そのため，日本語版は項目などがオープンにはなっていない．なお，研究利用の場合は，手数料が免除される可能性もあるとされている．患者報告アウトカムの開発や翻訳には多くの研究資金が投入されていることを考えると手数料や利用料がかかることは尤もであるかもしれない．しかし，手数料や利用料によって

PROMIS®の活用に制限がかかる可能性もある．

手数料の問題だけでなく，患者報告アウトカムの測定は使われる言語に依存することを考慮すると，日本でも PROMIS®のような患者報告アウトカムの項目バンクを開発する必要があるかもしれない．日本語の特性も考慮した項目バンクの作成は患者にとって切実なアウトカムの測定につながることが期待できる．もし，日本において患者報告アウトカムの項目バンクを開発する場合，(1)方法論的な厳密性を担保した開発フレームの設定，(2)教示，項目，選択肢，推定された項目パラメータ，解釈のための情報のオープン化，(3)図書館情報学研究者，患者，臨床実践家，統計学者，臨床研究者からなるプロジェクトチームが必要と考える．

まず，日本において患者報告アウトカムの項目バンクを開発するために，方法論的な厳密性を担保した開発フレームを設定する必要がある．海外で開発された項目バンクの翻訳ではなく日本で項目バンクを開発する場合は，方法論的な厳密性に基づく等価性を担保することが必要になる．既に PROMIS®や COSMIN において，患者報告アウトカム尺度開発の指針は示されている．開発にあたっては，PROMIS®が行ったようなドメインの枠組みの検討のように，大枠の検討から初めて，それぞれのドメインの開発をすることが重要になる．その点において PROMIS®は参考になるが，文化特異的なドメインが含まれる可能性もある．PROMIS®の検査成熟モデルにあるような，開発の進捗が分かりやすいような枠組みも採用し，プロジェクトが順調に進むような工夫も必要となる．

次に，教示，項目，選択肢，推定された項目パラメータ，解釈のための情報はオープン化し，研究利用から実践まで幅広く利用できるようにする必要がある．そのためにも PROMIS®において行われている尺度の項目のライセンスのチェックなども必要になる．また，オープン化することで，利用者の項目への曝露が増えてしまう可能性もある．項目バンクは随時項目を追加しながら持続的に発展するようなシステムにできると良いだろう．項目バンクのオープン化に関しては，参加者に日に複数回質問を行う経験サンプリングに関する項目リポジトリが参考になるかもしれない（Kirtley et al., 2019）．ESM Item Repository（https://www.esmitemrepository.com/）では，経験サンプリングで利用できる項目が各種メタ情報とともに複数の言語で利用可能になっている．

最後に，項目バンクの作成にあたっては，図書館情報学研究者，患者，臨床実践家，統計学者，臨床研究者からなるプロジェクトチームが必要となる．そのためには，異なる専門性や立場の者が集まってプロジェクトに持続的に取り組むための仕組みが必要となる．2〜3 年で終わるプロジェクトではなく長期間にわたり，メンバーが入れ替わりつつ取り組むための枠組みや組織が必要となるだろう．

上記のように，日本での患者報告アウトカムの項目バンク開発には，多くの困難があると予想される．困難はあるものの，項目バンク開発は患者のケア向上につながり，さらなる研究利用にもつながることが期待できる．本論文が，日本における患者報告アウトカムの項目バンク開発の一助になれば幸いである．

謝　辞

本研究は JSPS 科研費 JP20K20870，JP20H00625 の助成を受けたものです．

注．

[1] ドメイン構成，ドメインに含まれる構成概念，項目数などの詳細は，PROMIS®のウェブサイトに掲載されている．例えば，2023 年の段階では，成人用 PROMIS®のうつ尺度の項目プールには 28 個の項目が含まれる．また，英語版 PROMIS®については，具体的な項

目もウェブサイトから入手することができる．

[2] 欧米の研究機関などでは，図書館情報学などの学位のある者が図書館司書を担っている．彼らは文献検索などについて専門的知識を有しており，系統的レビューなどをサポートすることができる．日本国内においては，図書館情報学の専門家を研究チームに加えることが対応すると思われる．

参 考 文 献

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Rose, M. and PROMIS Cooperative Group (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years, *Medical Care*, **45**(5 Suppl 1), S3–S11.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., Hays, R. and PROMIS Cooperative Group (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008, *Journal of Clinical Epidemiology*, **63**(11), 1179–1194.

DeWalt, D. A., Rothrock, N., Yount, S., Stone, A. A. and PROMIS Cooperative Group (2007). Evaluation of item candidates: The PROMIS qualitative item review, *Medical Care*, **45**(5 Suppl 1), S12–S21.

Food and Drug Administration (2009). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims（最終アクセス日 2023 年 6 月 26 日）.

Fried, E. I., Flake, J. K. and Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement, *Nature Reviews Psychology*, **1**(6), 358–368.

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D., Schumacher, J., Saag, M., Kitahata, M. M. and Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: Calibrating the PHQ-9 to the PROMIS depression measures, *Quality of Life Research*, **20**(9), 1349–1357.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G. and Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks, *Nicotine and Tobacco Research*, **16**(Suppl 3), S175–S189.

Hays, R. D., Spritzer, K. L. and Reise, S. P. (2021). Using item response theory to identify responders to treatment: Examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) physical function scale and emotional distress composite, *Psychometrika*, **86**(3), 781–792.

İnce Araci, F. G. and Tan, Ş. (2022). Multidimensional computerized adaptive testing simulations in R, *International Journal of Assessment Tools in Education*, **9**(1), 118–137.

Kirtley, O. J., Hiekkaranta, A. P., Kunkels, Y. K., Verhoeven, D., Van Nierop, M. and Myin-Germeys, I. (2019). The Experience Sampling Method (ESM) item repository, http://dx.doi.org/10.17605/OSF.IO/KG376.

Klem, M., Saghafi, E., Abromitis, R., Stover, A., Dew, M. A. and Pilkonis, P. (2009). Building PROMIS item banks: Librarians as co-investigators, *Quality of Life Research*, **18**(7), 881–888.

Lord, F. M. and Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*, Information Age Publishing, Charlotte, North Carolina.

宮崎貴久子 (2015). QOL 評価の臨床的意味：Minimally Important Difference（臨床における最小重要差：MID）, 行動医学研究, **21**(1), 8–11.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis with Applications in Political Research*, De Gruyter Mouton, Hague, Netherlands.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M. and de Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content, *BMC Medical Research Methodology*, **10**, 22.

Mulder, J. and van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection, *Psychometrika*, **74**(2), 273–296.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D. and PROMIS Cooperative Group (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger, *Assessment*, **18**(3), 263–283.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C. and Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study, *Journal of Psychiatric Research*, **56**, 112–119.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W. and Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures, *Quality of Life Research*, **27**(5), 1147–1157.

PROMIS (2012). PROMIS® Instrument Maturity Model, https://staging.healthmeasures.net/images/PROMIS/PROMISStandards_Vers_2_0_MaturityModelOnly_508.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2013). PROMIS® Instrument Development and Validation Scientific Standards Version 2.0, https://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2014). Minimum Requirements for the Release of PROMIS Instruments after Translation and Recommendations for Further Psychometric Evaluation, https://staging.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf（最終アクセス日 2023 年 6 月 26 日）.

PROMIS (2023). Available Translations, https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations（最終アクセス日 2023 年 6 月 20 日）.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-S., Cella, D. and PROMIS Cooperative Group (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS), *Medical Care*, **45**(5 Suppl 1), S22–S31.

Reise, S. P., Bonifay, W. E. and Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality, *Journal of Personality Assessment*, **95**(2), 129–140.

Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A. and Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks, *Quality of Life Research*, **19**(9), 1311–1321.

Rimehaug, S. A., Kaat, A. J., Nordvik, J. E., Klokkerud, M. and Robinson, H. S. (2022). Psychometric properties of the PROMIS-57 questionnaire, Norwegian version, *Quality of Life Research*, **31**(1), 269–280.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika*, **34**(Suppl 1), 1–97.

佐藤秀樹, 土屋政雄 (2022). 尺度研究における COSMIN ガイドラインの動向, 認知行動療法研究, **48**(2), 123–134.

Segall, D. O. (1996). Multidimensional adaptive testing, *Psychometrika*, **61**(2), 331–354.

下妻晃二郎, 鈴鴨よしみ, 宮崎貴久子, 内藤真理子, 中島貴子, 川口崇, 山口拓洋, 齋藤信也, 兼安貴子, 星野絵里, 小嶋智美, 堺琴美, 白岩健, 宮路天平, 森脇健介 (2023). 厚生労働省科学研究班開発患者報告アウトカム(Patient-Reported Outcome:PRO)使用についてのガイダンス集, https://www.lifescience.co.jp/pro/index.html (最終アクセス日 2023 年 6 月 20 日).

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N. and Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions, *Psychometrika*, **86**(3), 674–711.

Thompson, N. R., Lapin, B. R. and Katzan, I. L. (2017). Mapping PROMIS global health items to EuroQol (EQ-5D) utility scores using linear and equipercentile equating, *PharmacoEconomics*, **35**(11), 1167–1176.

Current Status and Future Directions of Patient-Reported Outcome Item Banks Using Item Response Theory

Yoshihiko Kunisato[1] and Yoshitake Takebayashi[2]

[1]School of Human Sciences, Senshu University
[2]School of Medicine, Fukushima Medical University

Patient-reported outcomes are health status reports obtained directly from patients through methods such as questionnaires and are often used as outcomes in clinical trials. Although numerous patient-reported outcome measures have been developed based on classical test theory, the Patient-Reported Outcomes Measurement Information System (PROMIS®) has been developed an item bank for computerized adaptive tests based on item response theory. The computerized adaptive test enables the number of questions to be reduced while maintaining measurement precision, thereby lessening respondent burden. PROMIS® advances scale development in the order of item pool development, psychometric testing of items, and check of validity, setting scientific standards for scale development. The scientific standards involves (1) definition of target construct, (2) composition of individual items, (3) item pool construction, (4) determination of item bank properties, (5) testing and instrument formats, (6) validity, (7) reliability, (8) interpretability, and (9) language translation and cultural adaptation. In this paper, we discuss the scale development process in PROMIS® and deliberate the requirements when developing patient-reported outcome item banks in Japan.

Key words: Patient-reported outcome, item response theory, item bank, PROMIS®, COSMIN.