行動療法研究，41(2), 107–116, 2015

特集：「行動療法研究」における研究報告に関するガイドライン
〈展　望〉

尺度研究の必須事項

土屋　政雄

要　約

質問票による測定指標についての尺度研究は、医療や心理学の分野をはじめ、認知・行動療法の研究においても広く行われてきた。しかし類似概念尺度の乱立や、尺度は開発されたものの科学的に明らかにすべき尺度特性が十分に検討されていないなどの問題がある。こうした現状を変えるため健康における測定の分野を中心にCOSMIN（COnsensus-based Standards for the selection of health Measurement INstruments）チェックリストがさまざまな領域の研究者らの合意に基づき作成された。本稿では、COSMINチェックリストの概要を紹介するとともに、これに準拠し尺度研究において失敗しない研究計画を立てるため、特に重要な四つの留意事項（①例数設計、②再検査信頼性・測定誤差の評価、③仮説の設定、④反応性・解釈可能性の評価）の解説と、その具体的な記載事例を紹介することを目的とする。

キーワード：尺度特性　心理測定　質問票　患者報告アウトカム　COSMINチェックリスト

はじめに

認知・行動療法の実践や研究において、自己および他者評定による尺度は重要な測定方法の一つである。一方で、実証的に違いが見られない類似概念尺度の乱立の問題や（Schmidt, 2010）、測定結果を解釈する際に重要な尺度特性（psychometric properties）について、既存の尺度での再評価があまり行われていない（南風原，2011）という問題がある。すでに広く使われている尺度であっても、科学的に明らかにすべき尺度特性が十分に検討されていないことが、系統的レビューにより明らかになってきている。例えば仕事の機能尺度（Abma et al., 2012）、民族的にマイノリティの若者の外在化メンタルヘルス問題を測定する尺度（Paalman et al., 2013）などの系統的レビューによると、信頼性などの尺度特性が十分に検討されていないことが明らかにされている。こうした現状を改善するために、多数の研究者らの合意に基づき、COSMIN（COnsensus-based Standards for the selection of health Measurement INstruments）チェックリストが作成された。本稿では、COSMINチェックリストに準拠し、特に尺度研究の研究計画の段階で特に重要な留意事項に焦点化して解説し、記載事例を紹介する。

独立行政法人労働安全衛生総合研究所作業条件適応研究グループ
（2015（平成27）年3月13日受理）

COSMINとは

1.　概要

さまざまな学問分野において尺度による健康の測定が行われてきたが、信頼性や妥当性等の尺度特性について、用語（どのように呼ぶか？）と定義（何を意味するか？）の合意がなされていない状況があった（Mokkink et al., 2010b）。COSMINとは、専門家の集団の合意に基づき、尺度特性の用語と定義、および標準的な基準を設定するために作成された、健康関連尺度の選択に関する合意に基づく指針である。COSMINに関連して作成されたマニュアルおよびチェックリストがCOSMINのWebサイト（http://www.

cosmin.nl/cosmin.html）にて公開されている。COSMINの作成にあたっては、2006年から2007年にかけて、心理学、疫学、統計学、臨床医学の分野の専門家43名により4回の文書のやり取りにわたる国際デルファイ研究が実施され、合意形成が行われた（Mokkink et al., 2010a, 2010b）。従来の尺度研究で検討されてきた尺度特性は三つの領域（domain）に分類され、それは信頼性、妥当性と反応性に分けられている（Fig. 1）。各領域には、一つ以上の尺度特性あるいは尺度特性の下位分類が含まれる。例えば、信頼性は、内的一貫性、信頼性、測定誤差の三つの尺度特性が含まれる。なお、尺度特性の信頼性については、再検査信頼性／評定者間信頼性／評定者内信頼性が含まれている。また、妥当性には、内容的妥当性、基準関連妥当性、構成概念妥当性の三つの尺度特性が含まれ、さらに構成概念妥当性には、構造的妥当性、異文化間妥当性、仮説検証の三つの下位分類が内包される。反応性は一つだけであり、名前も同じ反応性となっている（Mokkink et al., 2010b）。解釈可能性は、尺度特性とは区別されるが、尺度の重要な特徴であるため、含められている。

COSMINは、それぞれの尺度特性などを検討するうえで、研究者が留意すべき点を明確にしている。例えば、基準関連妥当性を検討する場合、確定基準（gold standard）の合理性を説明することなど、7点に留意するよう規定されている（Fig. 2）。これらの留意事項は、すでに広く使われている尺度の科学的特性の適切性を評価することなどに活用できる。

2. 派生物

作成されてからまだ年月が経っていないため、COSMINから派生した資料は筆者の知る限りまだ見られていない。ただし、COSMIN公式Webサイトにおいて2014年9月末に更新された情報によれば、三つの異なるバージョンが現在作成中とある。それぞれ、尺度特性の研究計画のためのCOSMIN-P（Protocol）、尺度特性の研究報告のためのCOSMIN-R（Reporting）、

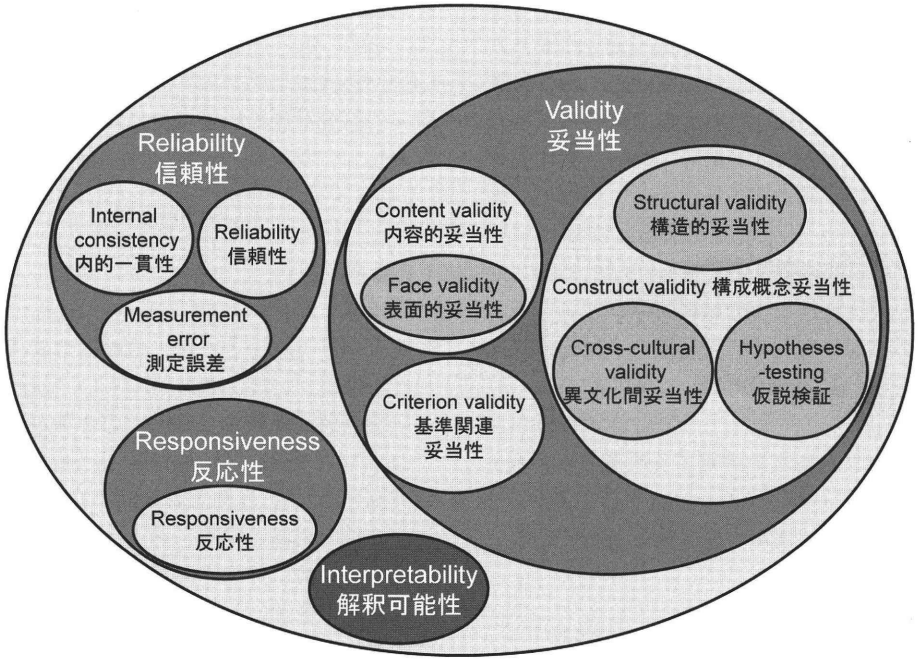

Fig. 1 COSMINの分類（COSMINチェックリストマニュアルの図を改変）

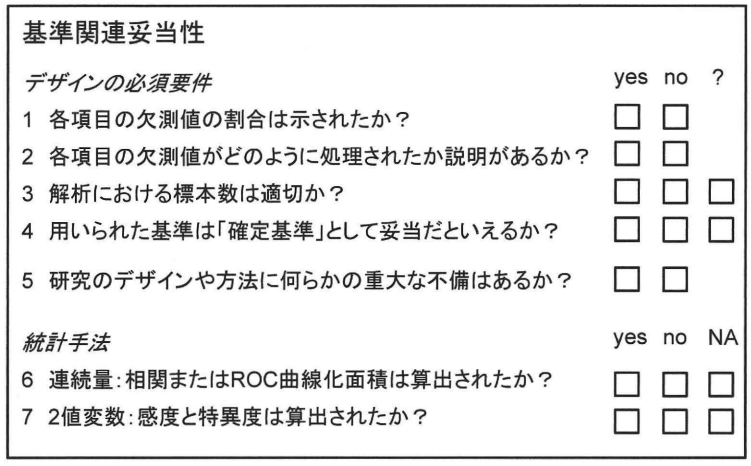

Fig. 2　基準関連妥当性の留意事項

尺度特性の系統的レビューのためのCOSMIN-S（Systematic review）の作成が進行しているとのことである。

3.　普及状況

特に医療系の生活の質（QoL）や理学療法の分野を中心に広く認知されており、産業領域など普及範囲は徐々に広がっている。認知・行動療法に関係した分野についても、はじめに挙げた文献に加え（Abma et al., 2012; Paalman et al., 2013）、例えば自閉スペクトラム症の子どもの不安への介入研究に用いられるアウトカム（Wigham & McConachie, 2014）やマインドフルネスに関する尺度（Park et al., 2013）の系統的レビューなどが出てきており、COSMINチェックリストに沿った評価が行われ始めている。Web of ScienceでCOSMIN作成論文（Mokkink et al., 2010b）の被引用件数は本稿執筆時点（2015年1月31日）で218件にのぼっている。一方で、個々の尺度研究論文におけるCOSMINの使用状況については、認知・行動療法と関わりの深い心理学および精神医学等の分野ではCOSMINに沿った構成のものはほとんど見られず、英文および和文誌のデータベース検索でもまだ数編しか該当しない。したがって、現段階では認知・行動療法を扱う分野の学術雑誌に投稿する際は、COSMINについての丁寧な説明が求められる可能性もあることに留意しておく必要がある。

尺度研究におけるCOSMINの活用

1.　特徴

尺度研究とは、多項目による尺度の新規開発を行う「尺度の開発に関する研究」と、尺度特性の検討による既存尺度の再評価等を主な目的とした「尺度の評価に関する研究」の二つに大きく分かれると考えられる（南風原，2011）。尺度研究のプロセスを広い視点から見ると、研究者が自身で研究計画を立てデータをとり論文を書く行為をはじめ、その論文の査読をする者がおり、出版後に研究が系統的レビューに組み入れられたり、現場で活用されたりなど多くの側面がある。COSMINのマニュアルには、チェックリストの活用場面が提案されているため、これらを参考に尺度研究自体の解説に代えて紹介することで尺度研究のイメージをつかみたい。

2.　チェックリストの活用場面

(1) 研究計画を立てる

尺度研究の研究計画の段階で、チェックリストを参照することで、「研究法の質」を高めることができる。新しい尺度を開発する際はもち

ろん、既存の測定指標の尺度特性を評価する際の研究計画立案にも活用できる。すでに行われた研究において、不十分なまたは評価されていない尺度特性を特定できるからである。無駄な研究を産み出さず、価値の高い研究を行うためには、既存の測定指標の評価研究を積極的に行っていくことが重要である。

(2) 研究を報告する・論文を書く

尺度特性について研究報告を行う際にチェックリストを使うことで、研究の質を適切に評価するための情報を、すべて報告できているかどうか確認できる。その際は、COSMINによる尺度特性の用語や定義にそろえ、ほかの用語や定義と混同させないことが推奨される。

(3) 尺度特性の系統的レビューを行う

尺度特性についての系統的レビューを行う際に、取り上げた研究の方法論的な質の評価を行う際にチェックリストを使用できる。質の低い研究を含めると系統的レビューの結論が真の結果に対するものからずれてしまう可能性が高くなるため、方法論的な質の評価は対象の研究デザインにかかわらず系統的レビューでは重要な手順となっている。系統的レビューで明らかになった尺度特性の検討状況を基に、既存の測定指標についてさらなる研究の必要性を見つけることも可能である。

(4) 測定指標を選択する

自身の研究や臨床で用いる測定指標を選択する際に、使用を考えている尺度やほかに候補となる尺度について、質の評価を行うために使用できる。

(5) 査読をする際の参考基準として使う

編集者や査読者が尺度特性についての研究を査読する際に、その研究が出版に値するだけの十分な質の高さを備えているかどうか評価する際にチェックリストが使用できる。加えて、まだ適切に報告されていない点を見つけるためにも使える。

留意事項

1. 例数設計をする

解説 研究開始前の段階で、解析に必要な標本数（sample size）を決めておくことは、心理学の研究はもちろん、どのようなタイプの研究をする際も必須になってきている（Wilkinson, 1999）。COSMINでは、内容的妥当性を除くすべての尺度特性において、実際解析に用いられた標本数が十分かどうか確認することが共通項目となっている。どのくらいの標本数が適切なのかについて確定した基準はなく、合意に至っているものではないがCOSMINの4件法得点化システム（Terwee et al., 2012）に示されている数が当面の目安になると考えられる。ほとんどの場合には、十分な標本数（100名以上）、良好な標本数（50名以上99名以下）、ほどほどの標本数（30名以上49名以下）、少ない標本数（30名未満）が適用されているが、一部因子分析や項目反応理論が関わる内的一貫性などのいくつかの尺度特性については、個別のより標本数の多い基準も設定されている。上述のとおり、合意に至っている基準ではないが、excellent（項目数×7かつ100名以上）、good（項目数×5かつ100名以上、または項目数×6〜7だが100名未満）、fair（項目数×5だが100名未満）、poor（項目数×5未満）といった内容になっている。ここで紹介した標本数の基準はあくまで目安であり、正式には用いる統計手法を基に例数設計を行うことが望ましい。

記載事例 「腰痛のコアアウトカム測定指数（COMI）ノルウェー語版の妥当性と異文化間修正」と題した研究では（Storheim et al., 2012）、方法の節に『研究の標本数はTerwee et al. の推奨により決定された[文献]…（中略）…構成概念妥当性、再検査信頼性、天井／床効果は少なくとも50名が必要で、内的一貫性の分析には約100名が必要であった』と記載されている。そして結果の節で『全部で90名の患者が研究に参加した。61名が再検査信頼性の研究に参加し、59名

がCOMI得点の両方の測定を完了した』と記載されている。ここで用いられているTerwee et al.の基準（Terwee et al., 2007）は、COSMIN以前にまとめられたものであるが、被引用数も多くCOSMINのマニュアルでも紹介されていることからCOSMINと同様に参照できる基準だと考えられる。

用いる統計手法を基に例数設計を行った尺度研究の例はまだ少ないが（Anthoine et al., 2014）、例えば感度と特異度を求めるための標本数計算例を示した研究では『感度と特異度を0.75、また有病率を40%（±10%）[文献]と想定した場合、診断のための十分な感度と特異度を推定するためには約180名が必要である[文献]』（Westergren et al., 2011）と記載されている。

2. 再検査信頼性・測定誤差を評価する

解説 従来の尺度研究では、Cronbachのα係数の算出だけをもって「信頼性が確認された」といった記載も多く見られたが、COSMINでは信頼性についても三つの構成要素に分類され、α係数はその中の内的一貫性の部分に対応しているのみである。十分な信頼性の検討には、ほかに再検査信頼性・測定誤差を検討することも必須となる。尺度研究において、再検査信頼性と測定誤差の尺度特性が検討されていないことが多いと指摘されている（Abma et al., 2012; Park et al., 2013）。

再検査信頼性とは、時間の間隔を空けて同じ測定を繰り返した際に、変化していない対象者が同じ測定結果である度合いである。一方、測定誤差とは、測定された構成概念の真の変化に起因しない、得点の系統的およびランダムな誤差のことである。測定誤差を検討すると、ある患者個人の尺度得点の介入前後の変化量が、尺度が検出可能な最小の値（誤差）を超えているかを確認できる。この値を超えた変化は、統計的に意味のある変化であると解釈される。

再検査信頼性・測定誤差については、共通のチェックリスト項目が設定されており、研究計画の段階で以下の4点が特に留意すべき点となっている。すなわち、①1回目の測定結果が2回目の測定結果に影響を与えないこと（実施の独立性）、②測定の方法、環境や教示などが測定時点間で類似していること（実施の類似性）、③測定の間隔は、思い出しバイアスを避ける程度に長く、測定する構成概念の状態が変化しない程度に短いこと（実施間隔の適切性）、④測定する構成概念の状態は、測定時点間で安定していること（状態の安定性）、である。ここでは状態の安定性について特に取り上げる。安定性については尺度特性を検証したい尺度とは別にアンカー（anchor）と呼ばれる直感的に解釈しやすい別の指標を測定することが重要となり、特に注意が必要だからである（de Vet et al., 2011）。アンカーは、客観的な指標を使う場合と主観的な報告を使う場合があり、主観的な報告の指標についてはさまざまな呼ばれ方をしている（例：global perceived effect [GPE]）。具体的には、例えば1問で「あなたの打撲症に関して、発生直後に比べた今の状態について評価してください」とたずね、−5（とても悪くなった）から0（変化なし）を経て5（すっかり回復した）までの11段階で評定を求めるといった指標のことである（Kamper et al., 2009）。現段階では、共通して使えるアンカーが用意されているわけではないので、先行研究を参考にしながら自身の研究内容に合わせたものを設定して実施することになる。

記載事例 「オランダ版下肢機能尺度は変形性股関節／膝関節症の者において高い信頼性、妥当性、反応性を持つ：妥当性研究」と題した研究では（Hoogeboom et al., 2012）、ある病院で整形外科医により変形性股関節／膝関節症の診断を受けた患者に変形性関節症の機能評定（LEFS）を求め、約3週間後に再度のLEFSおよびアンカーである7件法の全体的評定尺度（GPE）への回答を求めている。結果に書かれている『5名が改善（5%）（GPE＝1–2）、3名が悪化（3%）（GPE＝6–7）、ほとんどが安定（92%）（GPE＝3–5）』という箇所を参照する

ことで、状態の安定性を示す患者の割合が定量化できることを示している。

3. 仮説を設定する

解説 従来の尺度研究では、ほかの確立した測定指標との相関などを統計的に検討することで妥当性の検証とする例が多く見られた。COSMINでは、このほかの測定指標が確定基準（gold standard）であると説明可能な場合は基準関連妥当性を検討することになる。しかしそうでない場合、つまり多くのケースでは、構成概念妥当性の一部である仮説検証[注1)]を検討することになる。仮説検証の尺度特性を検討するためには、「同じ構成概念を測る二つの尺度間の相関係数は0.60以上」、「尺度が測る構成概念に違いがあると期待される二つの患者群を比較すると、尺度得点の平均値差は10ポイント」といった仮説を事前に設定する。ここで、仮説を設定する際、予期される相関係数や平均値差の方向や大きさを明確にすることが推奨される。論文では、事前設定したすべての仮説、それぞれの結果と、仮説が支持されたかの評価を表示することが望ましい。

検証に用いられる仮説の数は多いほど好ましいものの、COSMINでは仮説の数については基準を設定することはできないという結論に至っている。これまでの尺度研究における論文では、10個以上にのぼる仮説が設定されていることが多く、このうち75%以上の仮説が確証されることが一定の目安とされている（Terwee et al., 2007）。ただし、これはCOSMINでは明記されていないため、今後どのような扱いになるか注視する必要がある。

記載事例 子どもとその家族に対する中耳炎-6質問票（OM-6）の研究では（Heidemann et al., 2013）、方法の節で『質問票間および質問票内において項目の間（項目間）、項目と総合得点（項目−総合）、総合得点の間（総合−総合）で仮説を設定した[文献]…（中略）…相関の値で＜0.3を弱い、0.3–0.5を中程度、＞0.5を強いと定義した[文献]。』とあり、結果の節で『構成概念妥当性は24個の相関の仮説を検証することで評価した。[表]に仮説の例を示し、[表]に正しく、または間違って予測された相関の数を示した』とある。Table 1に仮説例の表を引用して示した。当該論文の雑誌のWebサイトにあるページから追加ファイルをダウンロード可能であり、すべての仮説が参照できる。

4. 反応性・解釈可能性を評価する

解説 認知・行動療法は、介入に伴う症状や行動の変化に関心のある領域であるため、尺度特性として、反応性や解釈可能性を明らかにすることが必須だと考えられる。

反応性は、これまでさまざまな定義が提案され、文献間で評価法の混乱が見られており（de Vet et al., 2011; Terwee et al., 2003）、COSMINによってやっと一定の合意が図られた尺度特性である。反応性とは、測定される構成概念における、時間経過による変化を検出することについての患者報告アウトカムの能力のことであり、つまり変化量の妥当性のことを指している。COSMINでは反応性は妥当性とは別の尺度特性として考えられているが、反応性を明らかにする手続き自体は、縦断デザインになる点を除けば仮説検証と基準関連妥当性で説明されている方法と同様である。

解釈可能性は、臨床的に理解されうる含意を尺度の変化量に付与できる程度のことであり、いわゆるカットオフ値の設定を指している。上述の測定誤差との違いとして、解釈可能性を検討すると、ある患者個人の尺度得点の介入前後の変化量が、臨床的意味のある最小の値を超えているかを確認できる。解釈可能性は尺度特性ではないとされているが、測定尺度の重要な特

注1) 東京大学教養学部統計学教室編『統計学入門』（東京大学出版会）によれば、hypothesis testingは「仮説検定」とされているが、その説明は“統計的仮説の「有意性の」検定”となっており、限定的である。COSMINではp値を用いずに説明することを推奨しており、求められる作業が異なるため、従来の「検定」の訳語を与えると混乱が生じると考え「検証」と訳した。

Table 1　構成概念妥当性の仮説例（すべての一覧は追加ファイル1から入手可能）

	相関の対象			
	質問紙（下位尺度または項目番号）	相関の仮説	コメント	得られた相関
FHS（身体的苦痛）	CHQ-PF50（体の痛み）	強い、ネガティブ	中耳炎に痛みがあれば、より一般的な質問にも表れてくる	−0.82
FHS（活動制限）	活動水準が低下していた日数	強い、ポジティブ	OM-6の項目はより広いが、どちらの項目も子どもの活動水準に関係している	0.56
NRS-child	CHQ-PF50（全体的健康）	中程度、ネガティブ	全般的健康は疾患特異的QoLに影響されそうである	−0.33
FHS（総合得点）	CHQ-PF50（総合得点）	強い、ポジティブ	子どものFHSは養育者のFHSに強く影響する	0.72

以下の論文のTable 4を本稿著者が翻訳した、Heidemann et al.（2013）Health Qual Life Outcomes. 11: 201. doi: 10.1186/1477-7525-11-201, CC BY 2.0（http://creativecommons.org/licenses/by/2.0/）
訳注：FHS, functional health status, 機能的健康状態；CHQ-PF50, Child Health Questionnaire 50 item version, 子ども健康質問票50項目版；OM-6, Otitis Media-6 questionnaire, 中耳炎-6質問票；QoL, quality of life, 生活の質；NRS, numerical rating scale, 数値評価尺度；相関係数の強さ：＜0.3弱い、0.3–0.5中程度、＞0.5強い

性であり、認知・行動療法の分野でも以前からJacobsonの指標などの臨床的有意性（clinical significance）として用いられてきた概念でなじみ深いものである（Crosby et al., 2003; Jacobson & Truax, 1991）。

反応性と解釈可能性を検討するためには、①介入の変化が生じうる一定期間を設定、②測定指標を介入前後の特定の2時点で測定、③アンカーの測定、④反応性の場合、関連があることが予想される別の尺度についても、介入前後の特定の2時点で測定、することが必要となる。再検査信頼性と測定誤差を検討する際との明瞭な違いは、介入により変化が生じる期間を設定することである（Fig. 3）。

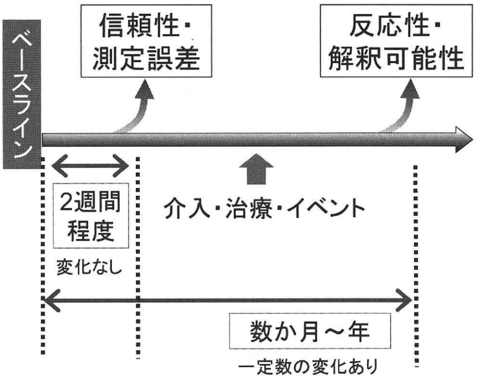

Fig. 3　縦断デザインの必要な点のまとめ

記載事例「慢性腰痛と変性椎間板疾患の患者におけるSF6D、EQ5Dおよびoswestry disability indexの比較」と題した研究では（Johnsen et al., 2013）、多施設無作為割付比較試験において172名の慢性腰痛または変性椎間板疾患の患者に対し、ベースライン時測定の後、治療を挟み2年後に再度尺度の評価を行っている（ベースライン時の欠損のない回答者は133名で、2年後にも回答したのは113名）。この研究では、反応性と解釈可能性が両方とも検証されている。測定に用いた尺度は、健康状態への価値や選好についての量的な値を測るための効用指標である、SF6DとEQ5D、および腰痛による日常の生活活動における身体的障害について評価するためのOswestry Disability Index（ODI）であった。また全体的得点（global score）、すなわちアンカーとして「受けている治療からどれ位の便益を得られていると思いますか」に

ついて「完全に障害されている」から「完全に回復している」までの7件法で尋ねている。

反応性については、方法の節で『ODIと2年後の7段階評定尺度を“確定基準”として反応性が評定された。まず、SF6D、EQ5D、ODIのベースラインから2年後フォローアップ時点での変化量についてスピアマンの順位相関を算出した。次に、SF6D、EQ5D、ODIの変化量と、全体的得点の2区分（1–3：改善、4–7：非改善）でROC曲線下面積を算出した』と記載されている。

解釈可能性については、方法の節で「MIC（minimal important change）は上述のROC解析の結果である感度と特異度に基づき計算された。最適な感度と特異度で患者の改善の有無を区別するカットオフ値がROC解析により決定された」と記載されている。

結 論

本稿では、COSMINの概要を解説するとともに、これに準拠し研究計画の段階で、特に重要な留意事項に焦点化して解説し記載事例を紹介した。失敗しない研究計画を立てるために、①例数設計をする、②再検査信頼性・測定誤差を評価する、③仮説を設定する、④反応性・解釈可能性を評価するといった側面を意識して準備する必要があることを述べた。本特集全体において共通する指摘事項であるが、紹介した四つの留意事項は研究計画における必須事項ではあるものの、COSMINの一部にすぎないことを踏まえておく必要がある。より有用な情報を提供できる尺度研究の報告を行うためには、COSMIN全体に目を通して理解を深めておくことが重要である。というのも、例えCOSMINを無視して尺度研究の論文を出版しても、必要な尺度特性が適切に報告されていなければ、後に系統的レビューに含められてpoorと評価される状況に徐々になりつつあるからである。COSMINの登場により、改めて尺度研究における用語や手法の見直しが必要となっており、新たな学習の必要に迫られている。本稿がその学習の一助になれば幸いである。

謝 辞

本稿の草稿に目を通し、詳細なコメントをくださった奥村泰之先生に深く感謝申し上げます。

文 献

Abma, F. I., van der Klink, J. J. L., Terwee, C. B., Amick, B. C., III, & Bültmann, U. 2012 Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. *Scandinavian Journal of Work, Environment and Health*, **38**, 5–18.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J. B. 2014 Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes*, **12**, 176.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. 2003 Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, **56**, 395–407.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. 2011 *Measurement in medicine: A practical guide*. New York: Cambridge University Press.

南風原朝和 2011 臨床心理学をまなぶ7 量的研究法 東京大学出版会

Heidemann, C. H., Godballe, C., Kjeldsen, A. D., Johansen, E. C., Faber, C. E., & Lauridsen, H. H. 2013 The Otitis Media-6 questionnaire: Psychometric properties with emphasis on factor structure and interpretability. *Health and Quality of Life Outcomes*, **11**, 201.

Hoogeboom, T. J., de Bie, R. A., den Broeder, A. A., & van den Ende, C. H. M. 2012 The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: A validation study. *BMC Musculoskeletal Disorders*, **13**, 117.

Jacobson, N. S. & Truax, P. 1991 Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, **59**, 12–19.

Johnsen, L. G., Hellum, C., Nygaard, O. P., Storheim, K., Brox, J. I., Rossvoll, I., Leivseth, G., & Grotle, M. 2013 Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. *BMC Musculoskeletal Disorders*, **14**, 148.

Kamper, S. J., Maher, C. G., & Mackay, G. 2009 Global rating of change scales: A review of strengths and weaknesses and considerations for design. *Journal of Manual & Manipulative Therapy*, **17**, 163–170.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. 2010a The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, **19**, 539–549.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. 2010b The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, **63**, 737–745.

Paalman, C. H., Terwee, C. B., Jansma, E. P., & Jansen, L. M. 2013 Instruments measuring externalizing mental health problems in immigrant ethnic minority youths: A systematic review of measurement properties. *PLoS ONE*, **8**, e63109.

Park, T., Reilly-Spong, M., & Gross, C. R. 2013 Mindfulness: A systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research*, **22**, 2639–2659.

Schmidt, F. 2010 Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, **5**, 233–242.

Storheim, K., Brox, J. I., Lochting, I., Werner, E. L., & Grotle, M. 2012 Cross-cultural adaptation and validation of the Norwegian version of the Core Outcome Measures Index for low back pain. *European Spine Journal*, **21**, 2539–2549.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. 2007 Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, **60**, 34–42.

Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. 2003 On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*, **12**, 349–362.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. 2012 Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research*, **21**, 651–657.

Westergren, A., Norberg, E., Vallén, C., & Hagell, P. 2011 Cut-off scores for the Minimal Eating Observation and Nutrition Form—Version II (MEONF-II) among hospital inpatients. *Food & Nutrition Research*, **55**.

Wigham, S., & McConachie, H. 2014 Systematic review of the properties of tools used to measure outcomes in anxiety intervention studies for children with autism spectrum disorders. *PLoS ONE*, **9**, e85268.

Wilkinson, L. 1999 Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594.

Improving Study Design for Measurement in Cognitive Behavior Therapy

Masao TSUCHIYA

Health Administration and Psychosocial Factor Research Group,
National Institute of Occupational Safety and Health

Abstract

Many studies of self-report psychological scales have been conducted in the area of cognitive behavior therapy, including in psychological and medical fields. The current situation is that too many duplicative constructs have been created, and many studies do not have adequate methodological quality; this has led to a lack of information necessary for clinical practice and further research. To change this prevailing situation, the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) checklist was developed, based on various fields of research, but focusing on health measurement. In order to plan a successful research plan, the present article first introduces a brief overview of the COSMIN checklist. Next, 4 important issues are explained: (a) planning sample size, (b) evaluating test-retest reliability and measurement error, (c) hypothesis testing, and (d) evaluating responsiveness and interpretability. How these issue comply with the COSMIN checklist is indicated, alongside concrete examples of writing a research report.

Key Words: psychometric properties, psychometrics, questionnaires, patient-reported outcomes (PRO), COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) checklist