「概念」の数量化
尺度開発の基本的な考え方

髙木廣文　東邦大学看護学部長

量的研究の基礎
測定について

ある現象について，量的研究を行なう場合，何はともあれ「数値によるデータ」を得る必要がある。当たり前すぎて，考えたことがない研究者も多いと思うので，まず「測定」ということについて考えてみたい。

身長や体重を測るといった場合，身長ならば身長計で，体重ならば体重計で測ることができるのだが，実は，身長と体重では測定方法に大きな違いがある。身長の測定は，長さを測るための「ものさし」で，身体の最長の部分を測ることでデータを得ることができる。しかし，体重の測定は「重さ」を測らなければならないので，そのような方法で直接測ることはできない。例えば，天秤のように梃子の原理を使ったり，バネを使って重さを長さに変換して測っている。血圧などの生理学的な検査値も同様にして，すべて特定の反応などを利用して，一次元の長さに変換してデータを得ている。

このようにして得られた量的データの多くは，原点が0で，マイナスの値のない「比尺度(ratio scale)」(2つのデータの比が計算できる)のデータであるが，気温のようにマイナスの値をもち，データの比較は「差(間隔)」のみが意味をもつ「間隔尺度(interval scale)」によるデータもある。なお，尺度とは「ものさし」のことであり，ある現象について測定するための「ものさし」をつくることを，「尺度化(scaling)」とか「尺度構成」と呼ぶ。

身長や体重のように，その測定方法が目に見えるやり方で測られるならば，問題になることはほとんどない。また，電圧計や電流計のように，電磁気力などを長さに変換するのであれば，物理学の理論式から導かれるので，その理論がよくわからないとしても納得する以外はないだろう。測定で問題になることは，「不安」「自己効力感」「エンパワーメント」など，直接見ることのできない「概念」を測る場合である。

尺度構成が必要とされるのは，もともと我々が目で見て簡単に共通了解が得られるような現象ではない。次に，そのような現象とはどのような現象であり，測定するときの問題にどのようなことがあるか考えてみよう。

KEY WORDS 尺度化，概念，測定，信頼性，妥当性

0022-8370/11/¥360/論文/JCOPY

現象と概念について

我々を取り巻く世界ではさまざまな現象が起きているが，目に見える現象と目に見えない現象では，研究を行なう上で大きな違いが生じてくる。目に見える現象は，多くの人が確認できるような形式で測定を行なうことが可能である。例えば，交通事故による外傷であれば，傷の大きさを測ったり，生理学的な各種の検査値や骨の状態を調べるためにX線撮影を行なったりすることができるだろう。ところが「痛み」を測ろうとすると，途端に難しくなってしまう。そもそもどのようにして「痛み」を測ればよいのだろうか。同じように，手術前の「不安」や「死への恐れ」など，誰もが抱いていることは確かであると思えるのだが，いざ測ろうとすると，何をどのようにして測ればよいのか途端にわからなくなってしまうのではないだろうか。

結局，現象として存在することは誰にも確かに思えることでも，心のなかに生じる現象，すなわち目に見えないことは，そのままでは誰も測ることはできないということである。我々は感情を表現したりする場合に，言葉以外でも「非言語的(nonverbal)」コミュニケーションにより，身体を使って無意識に感情が現われるそうである。もしもそれが本当ならば，それを利用してチェックリストを作成して，みんなで確認できるような方法で，量的に心的な事象を測定できるはずである。臨床症状などをチェックして，特定の疾患についてある尺度をつくることも可能であるが，ここでは，共通了解が得にくいような心的事象についての測定を考えることにする。

上記のように目に見える現象については，多くの人の共通了解が得られるような測定方法により数値化され，データとして解析が行なわれている。一方，目に見えない現象の測定は，言葉を通じて行なう以外，いまのところ方法はない。そのような心的事象を測るための方法について，以下に考えてみよう。

「ものさし」のつくり方と問題点

実際に，ある心的事象，例えば「共感性」をどのように測ればよいかを考えてみよう。言語論や哲学を持ち出すと話は厄介であるが，尺度化(ものさしづくり)の実際の手順は，それほど面倒なものではない。

不安や共感性のように，測定のための「ものさし」がないものは，身長のようには測ることはできない。しかし，そのような心理的な特性や状態を測定するための手段として，比較的簡単でよく用いられる方法がある。すなわち，測定したい概念に関係すると考えられる項目を多数作成し，それぞれの項目に対する回答に適当な得点を与え，その総合得点で不安や共感性などの概念を測定するという方法である。そのために，あらかじめ測定したい概念がどのように使用されているかを既存の文献で検討し(概念分析)，ある程度の操作的な枠組みを設定することもある。全く新しい概念やほとんど研究されたことがない現象であれば，対象となる患者に対してインタビューなどによる予備調査を行ない，どのような概念が関係しているのかを調べる場合もある。

どのような手順で行なうにしろ，このような方法である概念を測定する方法は，「順序尺度(ordinal scale)」からなる多数の項目を用いて，「間隔尺度を構成する方法」と解釈することができる。

表は，共感性尺度の項目の一例を示したものである。ここでは，6つの質問項目があり，それぞれの質問に，「3. はい」「2. どちらでもない」「1. いいえ」のうちから1つだけ回答を選んでもらうようにする。その後，各回答カテゴリに対して，3，2，1の得点を与え，その合計点で共感性を測定するものである。実際の研究では，回答方法は3段階でも5段階でも構わないが，回答の答えやすさや内容によって適宜変える必要がある。

表　共感性尺度の項目の例

1. 他人の気持ちが人一倍よくわかる
2. たとえ自分は損をしても友人の苦境はみすごせない
3. 他人の苦しみがよくわかる
4. 人のために尽くすことは好きだ
5. 人一倍感受性が豊である
6. すて犬，すて猫を見捨てるのはつらい

このような尺度化の方法は，分野によっては各個人に対して得点(スコア)を与えるので，「スコア化(scoring)」などとも呼ばれることがある。

このように，複数の項目に対する回答から得点化することで尺度を構成する場合，それらの項目の内容がその尺度の概念に適切なものでなければならないことはいうまでもない。最も重要な点は，作成した尺度が，自分が考えた概念を正しく測定しているのかということである。また，「ものさし」が「ものさし」としてその概念の測定に役に立っているのかも問題である。すなわち，朝と夕方でその測定結果が大きく異なるようでは，「ものさし」として役に立つとはいいにくいだろう。

一般に，上記のような手順で尺度構成された測定用具の適切さは，「妥当性(validity)」と「信頼性(reliability)」の2つの側面から評価する必要がある。以下に，これらの考え方について簡単に説明しよう。

尺度の信頼性について

尺度構成における「信頼性」とは，ある概念に関して繰り返し測定をした場合の一致の程度(一貫性)や「安定性(stability)」を示すものである。したがって，2つの反復する測定の結果が一致的であれば，その尺度の信頼性は高いことになる。以下に，よく用いられる信頼性を測るための指標について，その統計理論を簡単に解説しよう。

信頼性係数について

ある事象の測定において，真の得点を t，測定による得点を x，測定の誤差を e とすると，

$$x = t + e$$

と書くことができる。

このとき，信頼性(係数)は ρ，真の得点 t と測定による得点 x の分散の比，

$$\rho = \frac{V(t)}{V(x)}$$

によって定義される。ここで $V(t)$，$V(x)$ はそれぞれ真の得点 t と測定 x による得点の分散を示す。

実際には，真の得点 t を直接測定することは不可能である。そのため，2つの同等な尺度の同時的な測定(平行測定)や，調査の反復測定〔再テスト法(test-retest method)〕により，2つの繰り返し測定された尺度の相関係数により，信頼性の推定を行なうことができる。

1回の測定で信頼性を推定するには，以降に述べるような方法を用いる。これらの方法は測定用具としてすべての項目が同一の特性を測定している程度，すなわち「内的一貫性(internal consistency)」または「均質性(homogeneity)」の程度を測るものである。ここでは，複数の項目への回答から合成得点を尺度得点とする場合について，信頼性係数を求める方法について以下に解説する。

折半法による推定

簡単な計算方法として，「折半法(split-half technique)」と呼ばれる方法がある。この方法では，尺度を構成する項目について2つに分け，例えば奇数番の項目の合計得点と偶数番の項目の合計得点について，その相関係数を計算することで信頼性の程度を推測するという方法である。

信頼性係数を ρ とし，折半法による2つの項目群の得点の相関係数を r としたとき，

$$\rho=\frac{2r}{1+r}$$

により推定することができる。この公式は，「スピアマン-ブラウンの公式(Spearman-Brown prophecy formula)」と呼ばれている。例えば，折半法による相関係数 r が 0.6 ならば，上式から信頼性は $\rho = 0.75$ と推定できる。

折半法の問題点は，項目をどのように 2 分割するかによって，その数値が変わってしまうことである。項目の「奇数番-偶数番」「前半-後半」「ランダムな 2 群間」などで，求めた相関係数と信頼性係数の数値が異なることは明らかだろう。

アルファ係数

内的一貫性による信頼性の推定には，通常「アルファ(信頼性)係数(coefficient alpha)，クロンバックのアルファ(Cronbach's alpha)」が用いられている。

いま p 個の項目 $x_1, x_2, ..., x_p$ の合計点(尺度得点)を y とする。項目 $x_i(i = 1, 2, ..., p)$ と合計点 y の分散を $V(x_i)$，$V(y)$ としよう。

一般によく用いられているアルファ係数 α は，

$$\alpha=\frac{p}{p-1}\times\left[1-\frac{\sum_{i=1}^{p}V(x_i)}{V(y)}\right]$$

によって求められる。

なお，アルファ係数を求める場合，尺度得点の計算に用いる項目間の相関係数は，すべて正である必要がある。他の項目と負の相関である項目は，尺度得点の計算に用いないか，得点のつけ方を逆にしなければならない。

シータ信頼性係数

尺度の信頼性係数は，アルファ係数の他にもいくつか提唱されている。比較的よく用いられる信頼性係数に「シータ信頼性係数」がある。

多くの変数から合成変数を求める多変量解析の方法として，「主成分分析」がよく知られている。理論は省略するが，主成分分析の第 1 主成分は，分析に用いた変数の重みづけの合成得点であり，最適な尺度構成を行なった場合に対応している。

通常の尺度構成では，各項目にかかる重みはすべて同一であり，1 に設定されているのだが，主成分分析では項目間の相関構造を考慮して，最適な重みを求めて合成得点を計算しているのである。

変数間の相関係数行列を用いた主成分分析において，第 1 主成分の固有値を λ とした場合，シータ信頼性係数 θ は，

$$\theta=\frac{p}{p-1}\times\left[1-\frac{1}{\lambda}\right]$$

によって求められる。

シータ信頼性係数 θ は，用いられる項目の最適な重みづけをした場合の信頼性なので，アルファ係数 α とは，$\alpha \leqq \theta$ の関係があることが知られている。

尺度構成における項目の有用性の検討

尺度得点の計算に用いる各項目が，その尺度と関係があることが尺度構成の前提となっている。しかし，当初の項目の設定の考えに反して，実際のデータでは，ある項目は作成したい尺度とそれほど関係していない場合もあり得る。そのような状況を検討し，尺度構成に必要な項目か否かを検討するための方法がいくつかある。

最も簡単な方法は，尺度得点と，各項目の得点の相関係数を求めることである〔IT 相関(item-to-total correlation)〕。この IT 相関が 0.3 未満の項目は，尺度構成に関係のない項目として，尺度構成の項目から除外するのが適当だろう。

計算はやや面倒ではあるが，すべての項目を用いた場合のアルファ係数と，各項目についてその項目を除いた場合のアルファ係数を計算し，その違いを比較する方法もある。すなわち，その項目

を除くとアルファ係数の減少幅が大きな項目は，その尺度に重要な役割をもつと考えてよい。逆に，その項目を除いたほうが，アルファ係数の値が大きくなるようならば，その項目は用いないほうがよいことになる。この手順を繰り返すことで，項目を1つずつ減らしていき，最もアルファ係数が大きくなるように，尺度構成に最適な項目群を得ることができる。

また，主成分分析を利用してシータ信頼性係数を求めることができるのだが，各項目の主成分負荷量を検討することで，その項目が尺度構成に有用かどうかを判断することができる。すなわち，尺度を構成する項目間に相互に高い相関があり，その和を計算できるのであれば，すべての項目の第1主成分の負荷量(主成分との相関係数)が正に大きくなるはずだからである。もしも第1主成分が負ならば，その項目の得点のつけ方は逆順にして尺度得点の計算に用いなければならない。また，主成分負荷量が0.3未満の項目は，その尺度の計算に用いる意味がほとんどないことを示していると考えてよい。

尺度の妥当性について

尺度構成における妥当性とは，その尺度が「測定しようとしている概念内容を，どの程度適切に測定しているか」を示す用語である。作成した尺度の妥当性の評価は，以下に述べるようにいくつかのタイプに細分して考えることができるが，尺度の信頼性に比べて，妥当性の検証は努力目標の面があり，完全に検証するのは実際にはかなり難しい。

表面妥当性(face validity)

作成した尺度が，目的とした概念の適切な「ものさし」として，その構成概念を測定して「いそうかどうか」を意味する用語である。

複数の専門家に測定項目などをチェックさせ，その一致性などで表面妥当性の評価を行なうことは可能であるが，尺度の主たる妥当性の証拠と考えるべきものではない。

内容妥当性(content validity)

小学校卒業時の算数の計算能力を測る場合，どのような問題を作成すればよいだろうか。もしも，すべての問題が足し算ばかりで，引き算，かけ算，わり算などを含んでいなければ，そのような計算問題は内容に問題があると考えられるだろう。このような場合，その問題は「内容的に妥当ではない」といわれる。このような妥当性のタイプは「内容妥当性」と呼ばれている。

内容妥当性の確認のために，作成した尺度項目をその分野の専門家にチェックしてもらうのも1つの方法である。個々の項目がその尺度に適切か，また下位尺度が複数ある場合には，複数の専門家に全項目を各下位尺度に分類してもらい，各項目ごとの専門家による分類の一致度を検討するという方法もある。この場合，一致度が低ければ，尺度項目としては適切なものではないと考えられるだろう。

この方法では，まずその領域の専門家3人以上に尺度項目の評価を依頼する。すなわち，その概念を測定するのに各項目が適切か(ふさわしいか)どうかという点について，また各項目が構成概念の各次元を十分に(妥当に)測定しているかという点について評価してもらう。評価方法は，各項目について，4段階で「1. 関連がない」「2. あまり関連がない」「3. やや関連がある」「4. かなり関連がある」までで評価を行なう。これらの結果から，「内容妥当性指数(content validity index ; CVI)」を計算することができる。尺度全体のCVIは，3または4と評価された項目の割合であり，これが0.8以上であれば，適切な内容妥当性であると考えてもよいだろう。個々の項目についても，同様にしてその項目が尺度構成に適切なものかを判断できるだろう。

ただし，この方法は洗練されてはいるが，専門家とはいえ表面妥当性を検討しているだけなので，あまりその結果に信頼をおいて項目の削除などを行なうと，最終的に各下位尺度の項目不足などを生じることが多々あり，適切な尺度構成にならないこともあるので，注意が必要である。

一般に，上記の算数の問題例のように，すべての尺度でその内容妥当性が簡単に評価できるわけではない。高齢者のQOLの内容は何か，患者の不安を構成するものは何かなど，実際には明確に定義できないことも多いだろう。そのような場合には，内容妥当性は，抽象的な概念の測定に対する，その尺度の内容に関しての努力目標的な意味をもつものと考えられる。

■基準関連妥当性（criterion-related validity）

大学入学試験の妥当性は，何によって評価できるだろうか。1つの考え方として，大学入学後の成績と入学試験の相関が高ければ，入学試験は妥当であったと考えることができるだろう。また生活習慣を測定する場合，そのうちの1つとして，質問紙を用いて塩分の摂取量を推定したとしよう。推定した塩分摂取量と，実際の食事調査での塩分摂取量との相関が高ければ，質問紙での塩分摂取量推定法は妥当性をもつと考えてよいだろう。

このように，新たな尺度の妥当性を調べるために，適当な評価基準（外的基準）との相関で検討される妥当性は「基準関連妥当性」と呼ばれている。

さらに，基準関連妥当性は，問題にしている尺度の測定と外的基準の測定時期によって2つに分類して考えることができる。

質問紙調査による塩分摂取量推定法と栄養調査による食塩摂取量などのように，その測定が2つとも同時に実施されている場合，「同時的妥当性〔concurrent validity（併存妥当性）〕」と呼ばれている。

一方，大学入学試験と大学入学後の成績は，同時に測定することは不可能である。このように，評価基準が時間的に未来に測定される場合，「予測的妥当性（predictive validity）」と呼ばれている。

同時的妥当性でも予測的妥当性でも，どちらの場合も評価方法は，問題となっている尺度と外的基準との相関を検討することにより行なわれるのが普通である。

しかし抽象的な概念の尺度化では，上記の例のように比較的容易に評価基準を見つけられることは稀である。

■構成概念妥当性（construct validity）

作成した尺度が本当は何を測定しているのか，実際に測ろうとしている概念について十分に測定しているのかは，検討する必要がある。しかし現実には，概念が抽象的であればあるほど，その妥当性の検証は難しい。

構成概念妥当性については，その概念を構成する理論的な因子の検討が必要である。例えば，「死の不安」について測定のための尺度を作成する場合，「死の不安」について論理的で説得力のある概念化が重要であろう。

構成概念妥当性の検討方法はいくつかあるが，理論的考察により予測される論理的な分析や検証に基づいている。通常，作成した尺度の測定結果が，他の諸概念・構成概念の尺度の測定結果と，理論的に導かれる仮説に関して，どの程度符合しているかにより評価することができる。

例えばQOL尺度を開発した場合，QOL尺度得点の高い対象では，生活満足度（得点）や社会的適応（得点）が高いはずであり，また情緒不安定（得点）などは低いものと期待されるだろう。

ある概念の測定尺度が複数個ある場合には，「収束妥当性（convergent validity）」と「弁別妥当性（discriminant validity）」の観点から構成概念妥当性の検討を行なうことができる。異なる方法で同一の概念を測定した場合，類似の結果が得られれ

ば，同一概念に収束していることになる。実際には，同一概念を測定するための異なる方法間での相関がかなり大きくなければ，収束があることにはならない。また，異なる概念を測定するための方法とは，相関がないか弱い逆相関であれば，異なる概念を測定しているものと考えられるので，弁別の能力があると考えられるだろう。ただし，新たに尺度開発を行なった場合には，複数の測定方法を開発するのは困難であるため，必ずしもこのような方法が応用できるとは限らない。

尺度の妥当性が疑われた場合

作成した尺度について，その理論的な仮説と実際の測定との間に破綻がある場合，いくつかの理由が考えられるだろう。主な理由として，

1）基礎にある理論の誤り
2）理論を検証する方法などの誤り
3）測定の信頼性の低さ

などが考えられるだろう。

1)は，ある概念を測定するための尺度構成の基礎となった理論上の誤りである。例えば，QOLを「経済的要因と家族関係要因の相互作用で決定できる」という簡単なモデルから作成した場合などがそうである。このような場合には，その概念に関係する他の要因を新たに導入したり，より適切な理論を参考にして尺度構成のための概念モデルを再考する必要がある。

2)は，調査などを簡単に済ませようとした場合などに比較的多くみられるのではなかろうか。例えば，高齢者の日常生活活動度(ADL)を測定するための尺度を開発したとしよう。そのADL尺度の妥当性を調べるために，大学生を対象に調査を実施したとすると，仮にその調査結果から開発したADL尺度は妥当でないという結果になっても，その尺度を否定する結論にはならないだろう。なぜならば，明らかに調査対象の選択が不適切であり，得られた結果からは，ADL尺度が本来目的とする事象の測定が可能か否かを評価することはできないからである。

3)でいう「信頼性の低さ」の問題は，信頼性の低い尺度間の相関係数は，「それらの尺度の信頼性を超えない」ことが理論的に明らかとなっているからである。したがって，基準関連妥当性などの評価を相関係数で行なう場合，信頼性の低い尺度を用いると，見た目の相関があまり大きくならないことがある。

実際には，上記のどれが原因で尺度の妥当性を脅かしているのかを特定することは，困難な場合が少なくない。そのような場合，尺度構成に使用した項目を用いて因子分析などを行なうことで，項目間の構造が理論的に仮定したように構成されているかを検討したり，複数の尺度間の相関分析を行なうのも，妥当性を評価する上では有用なものとなることもある。

尺度構成を行なう過程では，すでに説明したように主成分分析を用いたり，構成概念が仮定したようになっているかを確認するために因子分析を用いることも多い。ここでは紙面の都合もあり，主成分分析と因子分析の詳細に関しては他の成書に譲る。

おわりに

尺度構成は，複数の項目について対象者に回答してもらい，その総合得点によって特定の事象を測定するものである。したがって，できるだけ少ない項目で測定ができれば，それに越したことはない。ただし，通常は項目数が大きければアルファ係数などをある程度大きくすることはできるが，極端に項目数が少なくなると信頼性係数も小さくなるのが普通である。このため，尺度構成の作業開始時には，ある程度多くの項目を用いて検討を行ない，信頼性をある程度維持したまま(0.8以上の信頼性があれば望ましい)で項目数を減らすという作業が必要である。

また，測定したい事象に関して下位の概念が存在する場合，調査データから仮定した概念がうまく抽出できるのかを，因子分析で検討することもよく行なわれる。この手順は現実の尺度構成ではかなり重要な分析となることが多い。

実際の研究では，構成した尺度とともに他の変数も一緒に解析を行なうのが普通である。その場合，尺度得点は間隔尺度による量的データとして解析されるのが普通である。すなわち，尺度構成とは，順序尺度で構成された多数の項目から，間隔尺度の量的データとなるような「ものさし」をつくる方法なのである。したがって，尺度得点の平均値や分散を求めて，量的データに基づく統計学的検定を行なうのに特に問題はない。

しかし，統計学的検定の基礎として，尺度得点の母集団での分布が正規分布に従うか不明であることから，ノンパラメトリック検定を用いる例も多く見受けられる。しかし，実際に比較すると両者の検定結果などに大差がないことも多い。この理由は，よく用いられる母平均値の差の検定のような検定方法が，母集団分布が正規分布から逸脱した分布であっても，十分に統計学的検定を行なうことができるという特性をもつことによる。これは，「頑健性(robustness)」と呼ばれる性質であり，「中心極限定理(データの和の分布が正規分布するという定理)」とともに，多くの統計学的検定方法がもつ特徴でもある。したがって，標本データの分布が正規分布していないからといって，多くの場合にはノンパラメトリック検定をする必要があるとは限らないということを知っておくべきだろう。

尺度構成では，作成している尺度の信頼性や妥当性の検討を行なう場合，多変量解析の手法をよく用いるので，理解するのがそれほど容易ではないかもしれない。しかし，統計解析用のソフトの普及で，計算自体は簡単に行なえるようになった。この点から，分析結果の解釈の誤りなども起こりやすくなっているのだが，まだ誰も測定しようとしていない現象などもあるかもしれない。そのような現象に名前を与えて概念化し，数量化するということは，研究としては大いに意義があることになるので，失敗を恐れずにどんどん研究を進めていただきたいものである。

文献

- Polit, D.F. & Beck, C.T.(2004). *Nursing Research : Principles and Methods*(7th ed.). Lippincott Williams & Wilkins. ／近藤潤子監訳(2010). 看護研究—原理と方法，第2版. 医学書院，pp.427-462.
- Walker, L.O. & Avant, K.C.(2005). *Strategies for Theory Construction in Nursing*(4th ed.). Pearson/Prentice Hall. ／中木高夫，川崎秀一訳(2008). 看護における理論構築の方法. 医学書院.
- 高木廣文，林邦彦(2006). エビデンスのための看護研究の読み方・進め方. 中山書店，pp.145-156.
- 高木廣文(2009). ナースのための統計学，第2版. 医学書院，pp.179-191.
- Carmines, E.G. & Zeller, R.A.(1979). *Reliability and Validity Assessment.* Beverly Hills : Sage Publications. ／水野欣司，野嶋栄一郎訳(1983). テストの信頼性と妥当性. 朝倉書店.
- Rogers, B.L. & Knafl, K.A.(2000). *Concept Development in Nursing : Foundations, Techniques, and Applications*(2nd ed.). Saunders.

たかぎひろふみ●東邦大学看護学部
〒143-0015　東京都大田区大森西4-16-20