# 抽象概念の測定方法について

YOSHIDA, Kenji / 吉田, 健二

(出版者 / Publisher)
法政大学産業情報センター
(雑誌名 / Journal or Publication Title)
グノーシス : 法政大学産業情報センター紀要 = Γνωσις
(巻 / Volume)
4
(開始ページ / Start Page)
27
(終了ページ / End Page)
36
(発行年 / Year)
1995-03-31

# 抽象概念の測定方法について

吉田健二

## 1．はじめに

社会科学における研究には大きく分けて実体（substantive）研究と概念実証（construct validation）研究の2種類がある。実体研究は2つまたはそれ以上の異なる概念間の関係を解明しようとするものであり、独立変数と従属変数の間の関係に焦点が当てられる。例えば、論文のタイトルに何々の影響とか、何々の効果とか、何々との関係とかがついたものは、全てこの実体研究に属している。これと同等に重要なものとして概念実証研究が存在し、これはコンセプトまたは概念を測定しようとする尺度（measures）から得られた結果とその尺度が測定しようとするコンセプトまたは概念の間の関係を解明しようとするものである。

日本の社会科学の分野においては、今まで主に実体研究の方に力が入れられており、概念実証研究は残念ながらほとんど行われていない状態である。したがって、この論文においては概念実証研究の方に焦点を当て、特に抽象概念を測定する方法を提示することを目的としている。その際に、信頼性（reliability）と妥当性（validity）という2つの概念が測定ということにおいて重要であるので、まずそれらの定義と違いについて述べる。それから、Churchill(1979）が示した概念の測定方法に沿って、6つのステップを説明する。概念のドメインを明確にし、そのドメインをとらえるアイテムを生成し、尺度を純化し、新しいデータでもって信頼性を評価し、概念の妥当性を評価し、そして標準（norms）を作成するというステップである。

## 2．測定の理論

測定（measurement）の定義には、「数量を表わすようにもの（objects）に対して数字を割り当てるためのルール」（Nunnally、1978、P. 3）が含まれるべきである。このことは、測定について注意しておくべき2つの重要な側面を示唆している。まず初めに、測定はものの属性を測ることであり、もの自身を測るのではないということである。次に、このことは数量化のためのルールの性質については何も述べていなく、このルールが特定化される程度によって、概念がどの程度よく測定されるかが決まるということである。

測定の理論においては、観察されたスコア(X)は真のスコア(T）とシステム的な測定エラー(S)とランダムな測定エラー(R）という3つの部分から構成されていることを理解することが、まず必要である。それを数式で表わすと、以下のようになる。

$$X = T + S + R$$

観察されたスコアは個人がテストにおいて実際に得るスコアのことであり、誤りやすい（fallible）スコアとも呼ばれている。真のスコアは「個人が実際に所有している測定された特性（trait）の量」（Ghiselli、Campbell and Zedeck、1981、P. 196）と定義されている。しかし、この真のスコアというものを知ることは全く現実的には不可能であり、同一のものに対して同じ尺度を使って実施された結果の平均スコアであると実際にはみなされている。観察されたスコアと真のスコアの差がエラーの部分であり、システム的なものとラン

ダムなものの２つに分けられている。システム的な測定エラーはスコアにおいてシステム的な変化を引き起こすものであり、ランダムな測定エラーはスコアにおいてシステム的でないすなわちランダムな変化を作り出すものである。

次に、先の数式から以下のような数式が導き出される（Zeller and Carmines、1980）のであるが、これを理解する必要がある。

$$\sigma_x^2 = \sigma_t^2 + \sigma_s^2 + \sigma_r^2 + 2\sigma_{ts}$$

ただし、$\sigma_x^2$ ＝ 観察されたスコアの分散

$\sigma_t^2$ ＝ 真のスコアの分散

$\sigma_s^2$ ＝ システム的な測定エラーの分散

$\sigma_r^2$ ＝ ランダムな測定エラーの分散

$\sigma_{ts}$ ＝ 真のスコアとシステム的な測定エラーの分散

測定において大変重要な概念である信頼性は、「測定が繰り返される程度」（Nunnally、1978、P. 191）と定義されている。すなわち、信頼性は測定がエラーからフリーの状態を指し、首尾一貫した測定結果を生み出す程度のことである。言い換えれば、信頼性はランダムでない分散の割合であり、先の数式を使って表わせば、以下のようになる。

$$\text{信頼性} = \frac{\sigma_t^2 + \sigma_s^2 + 2\sigma_{ts}}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_r^2}{\sigma_x^2}$$

測定のための道具である尺度は Nunnally(1978) が主張しているように、それが測定しようとしているものを実際に測定していれば妥当なのであると、一般的な意味においては言うことが可能である。したがって、妥当性は観察されたスコアが真のスコアと共有する分散の割合と定義され、以下のように表わされる。

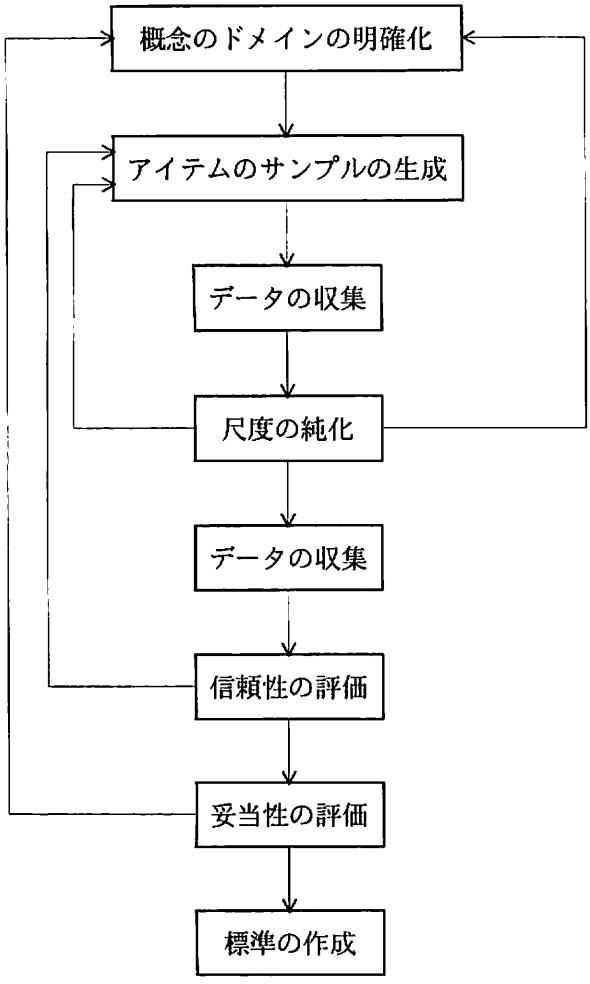
$$\text{妥当性} = \frac{\sigma_t^2}{\sigma_x^2}$$

信頼性と妥当性の違いを説明するために、よくライフル銃が例として使われる。しっかりと固定されたライフル銃からの玉が全てターゲットの同じ位置に当たってはいるが、残念ながらそれは目標とした位置ではない場合を考えてみるとよい。ライフル銃の玉は同じ箇所に命中しており、その意味で首尾一貫していたということでライフル銃は信頼性があると言うことができるが、目標としていた位置に玉が当たっていないという意味で、ライフル銃は妥当性がないと判断されるのである。

Lin(1976) が述べているように、ランダムな測定エラーは妥当性よりも信頼性に影響を与え、システム的な測定エラーは信頼性よりも妥当性に影響を及ぼすと言うことができる。もし測定が妥当であるならば、それは信頼性があると言うことが可能であるが、その逆は必ずしも真ではない。したがって、信頼性は妥当性の必要条件ではあるが、十分条件ではないのである。

抽象的な概念を測定するためには、信頼性並びに妥当性がある尺度を作成する必要がある。そのような尺度を作り出すために開発された Churchill(1979) が示した概念の測定方法を図に表わせば、以下のようになる。

**図1 より良い尺度を開発するための方法**




出所： Churchill、1979、P. 66

### 3．概念のドメインの明確化

抽象的な概念を測定するための尺度を作成する第1のステップは、その概念のドメインを明確にすることである。すなわち、概念のドメインのなかに何が含まれ、何が含まれていないのかを明らかにすることである。 Schwab （1980）が述べているように、概念を測定するようにデザインされた尺度の信じられそうな精神測定の（ psychometric ）属性を見分けることが重要である。例えば、概念の安定性が知られていない場合には、その概念を定義するときに安定性は含まれるべきではないのである。しかしながら、尺度を作成する際に概念の定義というものは仮定されたものであり、それは目的自身というよりも単なる手段にすぎないということも、同時に心得ておくべきことである。

尺度が理論のなかにおいて評価される際には、その概念とその他の概念を測る尺度との仮説的な連結が具体的に述べられることが必要である。Cronbach and Meel(1955）は理論を構成する法則の重なり合うシステムのことを精神法則論的（ nomological ）ネットワークと呼んでいる。考慮されている概念の明確化や概念の妥当性を調べるときの手続きを確立するのに、この法律学的ネットワークは役立つと、 Schwab （1980）は述べている。仮説的な連結を明確にすることは、概念を測定するための実証テストを提供することになるのである。

概念のドメインを明確にするときに、それに関する文献を十分に検討することが必要不可欠である。幸いにも、使用するのに適しているように思われる尺度が、今までに研究者によって開発されているかもしれない。もし自分の研究の目的のために妥当な尺度が存在しない、または現存している尺度が満足のいくものでないと考えられる場合には、自分自身の新たな尺度を作成しなければならない。他の人と異なった尺度を使用することは、その学問の分野における研究成果の蓄積というものを難しくすることになる。したがって、どうして古い尺度よりも自分自身の尺度の方が優れているのか、その理由を示す必要があるように思われる。

### 4．アイテムのサンプルの生成

尺度を作成する第2のステップは、第1のステップにおいて明確化された概念のドメインをとらえるアイテムを生成することである。そのためには、その概念のいくつかの次元（ dimensions ）の各々にふれる一組のアイテムを開発しなければならない。アイテムをプールするときに微妙に異なった意味を持つアイテムを含めて、その後にそれらを改良することによって、より良い尺度を作り出すことができるかもしれない。

アイテムを生成するときに使われるいくつかのテクニックが存在する。まず初めに、文献を調べることによって、どのようにその変数が以前に定義され、またその変数がいくつの次元からできているのかなどについて洞察が得られるように思われる。第2番目に、その分野において特殊な専門知識を持った複数の研究者に相談することによって、その概念についてのアイデアや洞察力を得ることが可能である。第3番目に、Flanagan (1954) が提唱した重大な事件テクニック（critical incidents technique）を使用することができる。これは、特定の状況を描写した多数のシナリオを作成し、経験者に各々のシナリオにおいてどのような行動をとるのであろうかと尋ねることによって、何らかのヒントを得ようとするものである。

このようなテクニックが存在するのではあるが、Ghiselli、Campbell and Zedeck (1981)が述べているように、心理的プロセスについての知識や以前において同様な変数を作成したことがあるかという経験、並びに知識、常識そして本当に全くの直観などによって、アイテムの生成が最終的には大きく左右されることになる。

アイテムを編集するときには、言葉づかいに特別の注意を払う必要がある。特殊な質問をしたり、ある単語や語句を使用することによって、回答にバイアスが生じることがある。例えば、Moser and Kalton (1971) は社会調査における言葉づかいについて気をつけるべき重要な側面が11もあることを指摘している。それらは十分に具体的でない質問、単純な言語、あいまいな文章、あいまいな単語、ある方向へ導くような質問、すでにそうであると仮定されたような質問、仮説的な質問、個人的な問題に関する質問、困惑的な質問、周期的な行動に関する質問、そして記憶に頼るような質問である。

## 5．尺度の純化

尺度を開発する第3のステップは、尺度を純化することである。このステップは、測定モデルによってかなり影響される。最も有名な測定モデルは、Tryon (1957)によって提唱されたドメイン・サンプリング・モデルと呼ばれるものである。これは、前述した「真とエラーのスコアの理論」とは考え方が全く異なっている。

ドメイン・サンプリング・モデルは、特性というものをある属性を共通に持った行動の集まりであると考える。ドメイン内の全ての行動に対して、その各々の尺度を作成したり、また多数の人々から各々の行動に関するスコアを得ることは、原理的には可能である。しかしながら、Ghiselli、Campbell and Zedeck (1981)が主張しているように、有限でかつ一定数の可能なアイテムが容易に見つけられるドメインがいくつか存在するように思われる。さらにまた、現実に有限の母集団からのランダムなサンプリングは、等しく良くそのドメインを代表していると思われる全ての利用可能なアイテムを我々は考慮に入れるということを、あらかじめ仮定している。しかし、等しく良いと考えられる全てのアイテムを我々が考えるとは、決して思えない。実際には、我々はそのドメインからの単なるサンプルを使用するにすぎないのである。ドメインから抽出されたサンプルは、理論的には構成要素の中央値と標準偏差の平均、構成要素間の共分散の平均、構成要素とそのドメイン外の他の変数との間の共分散の平均という3つの統計上の特徴が、社会全体のドメインと同じであるときに、初めてそのドメインを代表していると仮定されるのである。

ドメインに対してどのアイテムを最終的に選択するかは、その特性を測る他の尺度と関連している共通の属性を有する程度によって決定される。したがって、全てのアイテムの合計について各人のスコアを計算し、それから合計のスコアと関連がない、または関連しているがその程度が低いようなアイテムを削除することになる。スプリット・ハーフ法（split-half method）、Cronbach(1951)の$\alpha$係数、Kuder-Richardson の数式20と21などのアイテムの内的一貫性（internal consistency）を判断するいくつかの方法が、今までに開

発されている。例えば、α係数が高いということは、アイテムがその概念とうまく関連していることを示しており、逆に低いα係数はアイテムがその概念をうまくとらえていないことを意味している。よって、いくつかのアイテムは共通の属性を共有していないことになり、それらは削除されることになる。このように、各アイテムとアイテムの合計との相関係数を計算することによって、削除すべきアイテムを見つけることが可能である。なお、内的一貫性を判断する方法については、後に詳しく説明する。

α係数が計算され、不必要なアイテムが削除され、再度α係数が計算されるというように、満足のいくα係数が得られるまで、この過程は繰り返されるべきである。その際に、アイテムが加えられるべき分野や削除されるべきアイテムを見つけるために、因子分析が使用されることがある。また、次元の数が実証的に満足のいくものであるかどうかを調べる手段としても、因子分析は有用である。しかしながら、Schwab(1980) が指摘しているように、因子分析はその分析に含まれているアイテムのみから次元を抽出するのであり、また分析の結果はそのサンプルにのみ当てはまるという限界に注意しておくべきである。

もしこの尺度の純化というステップで満足のいく結果が得られない場合には、先に述べた概念のドメインの明確化という第1ステップそしてアイテムのサンプルの生成という第2ステップに戻って、初めからやり直す必要がある。

## 6．新データでの信頼性の評価

今まで説明してきた尺度を作成するステップは、アイテムを選択することによって起こる測定のエラーを削減するためのものである。もしこれらのステップがうまく行われたならば、外観（face）または内容（content）妥当性のある尺度が作成されることになる。

しかし、尺度において信頼性のないものを作り出すエラーの源は、まだ他にも多数存在している。例えば、Lin (1976)によれば、研究と関連したエラーとしてテストの状態、測定道具の腐敗化、テストと再テストの状態の相違、回答者の集団の形成、個人的、状況的、物理的、または処理上の変化、特殊な質問の選択、質問や物理的要因の不明確さなどが挙げられ、研究と関連していないエラーとして社会的背景、回答者の履歴、成熟化、文化的要因などが指摘されている。

尺度の信頼性を判断するための方法として、テスト・再テスト法（test−retest method）、内的一貫性、類似様式（parallel forms）という3つが開発されている。これら3つの方法は全て、尺度におけるシステム的な分散の割合を調べることによって、その信頼性を測定しようとするものである。

テスト・再テスト法においては、1つの尺度が同じ回答者に対して2回またはそれ以上使用され、それぞれの実施によって得られたスコア間の相関関係が、信頼性の係数として考えられるのである。この方法は時間を越えた尺度の安定性に焦点が当てられているために、この方法によって得られた信頼性係数は、しばしば安定性係数とも呼ばれている。

この方法は、2つの利点を備えている。まず初めに、信頼性を判断する他の方法は2つ以上の様式のテストを必要とするが、テスト・再テスト法は1つの様式で済むということである。次に、アイテムの特定のサンプルまたはテストにおける刺激の状態が一定に保たれるという利点がある。しかしながら、この方法はいくつかの不都合な点も同時に持っている。まず第1番目に、テストを行う間隔を変えることによって、結果が異なってくるということである。Bohrnstedt(1970)が述べているように、時間の間隔が長くなればなるほど、信頼性は一般的に落ちてくる。第2番目に、個人の真のスコアが変化する傾向があるという問題である。第3番目に、テストの間隔を長くすることによって、1回目のテストと同じ回答者を再度集めることが困難であるという問題である。

テスト・再テスト法は尺度の信頼性について有

用な情報を提供してくれるが、多くの問題を含んでいるために、内的一貫性を判断する際の補助的な役割を果たすものとして一般的には考えられている。

尺度の信頼性の内的一貫性を判断する基本的な方法として、スプリット・ハーフ法と呼ばれるものがある。この方法によると、偶数番目のアイテムと奇数番目のアイテムによって、またはランダムにテスト全体が2つの部分に分けられ、その2つの相関関係から尺度の信頼性が以下のようなSpearman−Brownの数式を使うことによって評価されるのである。

$$r_{cc} = \frac{r_{xx}}{1-r_{xx}}$$

ただし、$r_{cc}$ ＝ 信頼性係数

$r_{xx}$ ＝ テストの2分されたスコア間の相関関係

このスプリット・ハーフ法には、2分された各々の部分のスコアが一度に得られること、また異なる状態でテストされることから起きる測定エラーが排除されるという利点がある。しかしながら、テストをどのように2分するかによって得られる結果が異なってくるという基本的な問題が、この方法には存在している。この問題を克服する1つのアプローチは、テストのアイテム間の相関関係を使って信頼性を測定するという方法である。もしテストのアイテムが2分される（dichotomous）変数であるならば、Kuder−Richardson の数式20(KR20)と21(KR21)の2つが、尺度の信頼性を測定するために使用される。KR20とKR21は、それぞれ以下のように表わされる。

$$KR20 = \frac{N}{N-1}\left(1 - \frac{\Sigma p_i q_i}{\sigma_x^2}\right)$$

ただし、N ＝ 2分されたアイテムの数

$p_i$ ＝ i番目のアイテムに肯定的に答えた割合

$q_i = 1 - p_i$

$\sigma_x^2$ ＝ 全スコアの分散

$$KR21 = \frac{N}{N-1}\left(1 - \frac{N\bar{p}\bar{q}}{\sigma_x^2}\right)$$

ただし、$\bar{p} = \Sigma p_i / N$

$\bar{q} = 1 - \bar{p}$

Cronbach（1951）のα係数は、アイテムが連続する変数であるときに最も適したものであり、以下のように表わされる。

$$\alpha = \frac{K}{k-1}\left(1 - \frac{\Sigma V_i}{V_x}\right)$$

ただし、K ＝ 部分の数

$V_i$ ＝ 部分の分散の合計

$V_x$ ＝ 全スコアの分散

このα係数は、アイテムの共分散マトリックスから容易に計算することが可能である。測定における主要なエラーは内容のサンプリングから発生するものであると考えられるので、α係数はほとんどの状況において信頼性の優れた評価を提供するものであると、Nunnally（1978）は主張している。

テストの類似様式においては、同じ回答者が2つの異なる時間に2つのテストを使って測定される。各々のテストは内容的には類似しているが、そのアイテムは異なっている。類似様式の2回の実施によって得られるスコアが、信頼性係数を計算するために相関される。2つの様式が本当に同等または類似している程度によってこの相関係数が決まってくるので、これはしばしば同等様式（equivalent−forms）係数と呼ばれる。

信頼性を判断する手段としての類似様式の使用の主要な問題点の1つは、Gulliksen（1950)が述べているように、各様式におけるアイテムの平均、分散、並びに相関係数が全く等しいような同等の様式を作成することは困難であるということである。類似様式の使用のもう1つの問題点は、それ

らの様式を開発するときに必要とされる労力である。わずか1つのテストを作成するのにも多大な時間と努力を要するのに、2つ以上の様式を作ることは不可能である。さらにまたそれ以上に複雑な問題として、2つの様式が内容的に同等であるということを立証するということがある。例えば、もし2つの様式におけるスコア間の相関関係が低いならば、その尺度が低い信頼性しか持っていないのか、それとも様式の1つが単に内容的に他のものと同等でないのかどうかを決めることは、困難なことである。

類似様式でもって信頼性を判断する必要性は、尺度のタイプによって変わってくるように思われる。内容のドメインが容易に特定化できるような場合やスコアをつけるときに全く主観性が入らないような場合、また短期間で人々がほとんど変化しないように思われる場合には、Nunnally（1978）が主張しているように、α係数は信頼性の優れた評価になりうると考えられる。もし特性が比較的に短期間で変わるような場合には、類似様式の方がその変化を解明するために有用であると考えられる。

今まで述べた信頼性への伝統的アプローチは、残念ながら測定エラーが多くの源から発生するということを指摘することに失敗している。したがって、Cronbach et al.(1963) や Cronbach et al.(1972)は測定の手続きにおける分散の複数の源を一度に分析するという一般化（ generalizability ）理論を提唱している。

一般化理論は時間、道具、観察者などの尺度の側面に焦点を当てたものであり、サンプルされた状態において得られたスコアが、それらの状態に対する全スコアを代表しているかどうかが問題となるのである。全スコアは、信頼性についての伝統的理論における真のスコアと類似したものであると考えられている。

一般化理論の係数は観察されたスコアから決定され、観察されたスコアの期待された分散に対する全スコアの分散の割合として定義されている。しかし、これは一般化研究の主要な目的ではなく、測定の道具における全ての分散をいっせいに測定することが、この主要な目的なのである。分散の複数の源を同時に調べることによって、測定のより効率的な手続きを開発することが可能となると考えられている。

この一般化理論の主要な利点は、研究者が一般化したいと考える多くの世界があるということを明白に認めていることである。興味のある各々の世界から測定の状態をサンプリングすることによって、測定の手続きがそれらの世界を解明するように一般化研究においてはデザインされている。したがって、信頼性の伝統的な方法は一般化研究のただ単に1つの側面にすぎないと、考えられるのである。例えば、テスト・再テスト法を使って得られる信頼性係数は、測定の道具から得られたスコアが測定の全ての時間を超越して全スコアに一般化することができるかどうかに関心があると、考えられることになる。たとえテスト・再テスト法によって得られる係数が高くても、この尺度の他の世界に対する一般化の可能性については、何も言うことができないのである。

Cronbach et al.(1972) は、分散分析の論理を分散の複数の源を含むように拡大することによって、一般化研究における分散の様々な構成要素を評価するための方法を開発している。

要素的（ factorial ） デザインやアイテム特徴曲線（ ICC 理論）などのように、信頼性を判断するためのモデルは他にもいくつか存在するということに、注意しておくべきである。また、測定の回数を増やしたり、測定するときに実験的コントロールを施したり、より優れたアイテムを選択することによって、尺度の信頼性は高まるということにも注意しておくべきであろう。

## 7．概念妥当性の評価

前述した概念のドメインを明確にし、そのドメインをとらえるアイテムを生成し、そして尺度を純化するという3つのステップは、外観または内容妥当性や信頼性のある尺度を作成するためのも

のである。しかし、それらのステップは概念妥当性のある尺度を作り出すかもしれないし、出さないかもしれない。ここで言う概念妥当性とは、「概念（変数の概念的定義）とその概念を測定または操作する運営上の手続き間の一致」（Schwab、1980、PP. 5-6）であると定義される。すなわち、概念妥当性とは尺度が概念を測定する程度のことである。

尺度の概念妥当性を評価するためには、尺度が同じものを測定するようにデザインされた他の尺度とどの程度関連しているかを調べる必要がある。この概念妥当性を評価する最も有名な方法は、Campbell and Fiske（1959）によって開発された複数特性・複数方法（multitrait-multimethod: MTMM）マトリックスと呼ばれるものである。このマトリックスは、少なくとも2つの概念が最低2つの異なる方法によって測定されることが必要である。MTMMマトリックスの最も単純な例であるAとBという2つの特性と1と2という2つの方法が使用された場合を表に示すと、以下のようになる。

**表1 MTMMマトリックスの例**

| | | 特性A | | 特性B | |
|---|---|---|---|---|---|
| | | 方法1 | 方法2 | 方法1 | 方法2 |
| 特性A | 方法1 | $\gamma_{1A,1A}$ | $\gamma_{1A,2A}$ | $\gamma_{1A,1B}$ | $\gamma_{1A,2B}$ |
| | 方法2 | | $\gamma_{2A,2A}$ | $\gamma_{2A,1B}$ | $\gamma_{2A,2B}$ |
| 特性B | 方法1 | | | $\gamma_{1B,1B}$ | $\gamma_{1B,2B}$ |
| | 方法2 | | | | $\gamma_{2B,2B}$ |

出所：Ghiselli、Campbell and Zedeck、1981、P. 286

もし何を測定しているかを理解しているならば、同じものを測定している2つの異なる尺度の相関関係は高くなるはずであり、このことをCampbell and Fiske（1959）は収斂的（convergent）妥当性と呼んでいる。さらにこのことをより厳密にすれば、同じ特性を測定するようにデザインされた2つの異なる方法間の相関係数（表における$\gamma_{1A,2A}$）は、同じ方法で測定された2つの異なる特性間の相関係数（$\gamma_{1A,1B}$）よりも高くなるはずである。この条件のことは、拡散的（divergent）妥当性と呼ばれている。

概念妥当性は、収斂的妥当性と拡散的妥当性の両方の程度によって決まってくると考えられる。しかし、MTMMマトリックスからこれら2つの妥当性を判断することは、いくつかの問題点を含んでいるように思われる。まず第1に、使用される方法と特性はできる限り独立であるべきであるとCampbell and Fiske(1959）は主張しているが、このことは現実には難しい。というのは、方法から生じる分散の源に関しての知識をほとんど持っていない状態で、相互に独立した方法を選択することは困難であると、Kalleberg and Kluegel（1975)は述べている。方法が相互に独立しているかどうかは、後付けで決められるものである。

第2の問題点としては、問題となっている特性または方法と関係していない分散の共通の源を含むことによって、尺度の作成において無意識的なバイアスが収斂的妥当性と拡散的妥当性の立証に

おいて入りこむ可能性があるということである。

第３番目に、MTMMマトリックスを解釈するときに起きる現実的な問題がある。収斂的妥当性を調べるときに使われる基準は、あいまいかつ誤解を与えるものである。例えば、ある研究者はある研究では0.52の相関係数を収斂的妥当性の十分な証拠であると見なしているのに対し、他の研究では0.56の相関係数が必要であると主張していると、Schwab （1980）は述べている。また、どのような２つのスコアであっても必ずゼロとは異なる相関関係を示すものであり、収斂的妥当性の基準は相関関係を解明するには弱いハードルであると、Peter（1981）は主張している。

最後に、MTMMマトリックスの方法は先に述べた一般化理論の一部分であり、使用される方法はただ単に分散の１つの源にすぎないと考えられるのである。

尺度の概念妥当性を評価するためには、尺度が予想されたように作用するかどうか調べる必要がある。異なるがしかし概念的には関連している複数の概念を測定するように意図された尺度間の観察された相関関係を意味する精神法則論的妥当性と、概念妥当性は関係していると考えられる。Nunnally （1978）によると、概念妥当性を決定するためには、尺度が概念に関する理論と適合していなければならない。しかし、このことを証拠として使おうとすると、その理論は真であるということを仮定しなければならなくなる。すなわち、①概念ＡとＢは確かに相関している、②Ｘは概念Ａの尺度である、③Ｙは概念Ｂの尺度である、④ＸとＹは確かに相関している、という４つの仮説にこの循環的な論理は基づいていることになる。実際には、この第４番目の仮説のみが実証されるのであり、残りの３つの仮説についてはこの実証によって十分な妥当性があると仮定されているのである。言うまでもなく、このような仮定は実際には危険である。Nunnally（1978）が述べているように、仮説の真実性が明白であるような状況のみに概念妥当性の解明を絞ることによって、この危険性を軽減することは可能である。

## ８．標準の作成

生のスコア自身は、その特徴に関して個人がどこに位置しているのかについて、何も語ってくれない。スコアは他の人々が得たスコアと比較されることによって初めて意味を持ち、また有用となるのである。例えば、１から５までのスケールでできた10のアイテムによって企業のイメージを測定しようとするテストで、ある企業が40のスコアを得たとしよう。この40のスコアが非常にイメージが高いことを示しているかどうかは、他の企業のスコアと比較されて初めて分かるのである。したがって、尺度を作成する最終ステップは、標準すなわち「スコアに意味を付与する基礎となるある特定の集団によって得られたスコアの分布」（Ghiselli、1964、P.49）を作成することである。

スコアの分布は平均や標準偏差などを含むいわゆる記述統計によって一般的には示される。年齢、性別などによって分布が異なる集団が存在する場合には、それぞれの集団に対して標準を別々に作成することが必要である。

標準が適切に作成されているかどうかは、測定されるケースの数とそれが集団全体を代表しているかどうかによって決まってくる。ケースの数が大きければ大きいほど、その標準は安定したものになるし、選択されたサンプルが集団全体を代表していればいるほど、標準もそれをより良く表わしていることになる。

## ９．おわりに

測定の理論における信頼性と妥当性の定義と相違点を説明し、それから尺度を作成するための６つのステップを提示してきた。特に、新データでの信頼性の評価と概念妥当性の評価に、多くの紙面を割いた。

最初に社会科学における研究には、実体研究と概念実証研究の２種類があると述べたが、もはやこれらの研究の違いはあまり明白ではないように思われる。例えば、Cronbach and Meel(1955）は

テストの概念妥当性の解明は理論を開発し、それを実証する一般的な科学的手続きと本質的には違わないと述べている。また Schwab(1980)も実証研究が妥当性を持っているかどうかは、それにおいて使用される尺度や研究される概念を関連づける仮説の正確さについて研究者が立てる仮定に左右されると主張している。

日本の社会科学の分野においては、純粋な意味での実体研究や概念実証研究は、ある1部の分野を除いてはほとんど行われてこなかったと言っても過言ではないだろう。もし学問における知識の蓄積ということを欲するのであれば、やはり信頼性並びに妥当性のある尺度を使った研究がなされなければならないと考える。その意味で、この論文がその一助になれば幸いである。

## 参考文献


Bohrnstedt, G.W. "Reliability and Validity Assessment in Attitude Measurement," in Gene F. Summers(ed.), Attitude Measurement (Chicago: Rand McNally, 1970).

Campbell, D.T. and D.W. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin, Vol. 56,1959,81-105.

Churchill, G.A. "A Paradigm for Developing Better Measures of Marketing Constructs," Journal of Marketing Research, Vol.16,1979,64-73.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, Vol16, 1951,297-334.

Cronbach, L.J., Gleser, G., Nanda, H., and N. Rajaratnam, The Dependability of Behavioral Measurement: Theory of Generalizability for Scores and Profiles (New York: Wiley, 1972).

Cronbach, L.J. and P.E. Meel. "Construct Validity in Psychological Tests," Psychological Bulletin, Vol.52, 1955,281-302.

Cronbach, L.J. and Rajaratnam, N., and G. Gleser," Theory of Generalizability: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, Vol.16,1963,127-163.

Flanagan, J. "The Critical Incident Technique," Psychological Bulletin, Vol.51,1954,327-358.

Ghiselli, E.E. Theory of Psychological Measurement (New York: McGraw-Hill, 1964).

Ghiselli, E.E., J.P. Campbell, and S. Zedeck. Measurement Theory for the Behavioral Sciences (San Francisco: W.H.Freeman and Company, 1981).

Gulliksen,H. Theory of Mental Tests (New York: Wiley, 1950).

Kalleberg, A.L. and J.R. Kluegel. "Analysis of the Multitrait-Multimethod Matrix: Some Limitations and an Alternative," Journal of Applied Psychology, Vol.60,1975,1-9.

Lin N. Foundations of Social Research (New York: McGraw-Hill,1976).

Moser, C.A. and G. Kalton. Survey Methods in Social Investigation (London: Heinemann Educational Books,1971).

Nunnally, J.C. Psychometric Theory (New York: McGraw-Hill,1978).

Peter, J.P. "Construct Validity: A Review of Basic Issues and Marketing Practices," Journal of Marketing Research, Vol.18,1981,133-145.

Schwab,D.P. "Construct Validity in Organizational Behavior" in B.M. Staw and L.L. Cummings (eds.), Research in Organizational Behavior, Vol.2 (Greenwich, Connecticut: JAI Press, 1980).

Tryon, R.C. "Reliability and Behavior Domain Validity: Reformulation and Historical Critique," Psychological Bulletin, Vol.54,1957,229-249.

Zeller,R.A. and E.G. Carmines. Measurement in the Social Sciences (Cambridge: Cambridge University Press, 1980).