提案論文3

家族的類似の概念にもとづくテスト得点のモデル

名古屋大学教育学部　村　上　　隆

1．個人差測定の問題

テストの作成過程

通常，テスト（あるいは，一般に心理学的個人差を測定するための尺度）の作成過程は次のようなものである。まず最初に，測定しようとする構成概念についてのおおまかな理論，あるいは少なくともイメージが，既存の理論，設定された教育目標，現象に関する非公式的な観察，過去のデータ，といったものにもとづいて存在する。それに依拠して項目が書かれ，収集されて，一つのテストが構成される。

次に，そのテストが多数の被験者を対象にして実施される。そのデータにより，まず，テスト得点の信頼性が，α係数等で表現される内的整合性や，再テストに関する安定性にもとづいて確認される。時には，一つの構成概念が複数の下位概念に分割されていることもある。その場合，（広い意味の）因子分析によって，その分類が確認されることも多い。さらに，探索的な因子分析にもとづいて，新たな下位概念が提唱されることもある。それぞれの因子に目立った負荷をもつ項目得点の単純和で，新たな尺度が定義され，それらについて再度，信頼性の検討がなされる。また，テストの妥当性を確認するために，テストは他の尺度と同時に実施され，被験者の属性が記録される。他の尺度との相関や，被験者の属性にもとづく群間の差が，あらかじめ想定された理論と矛盾しなければ，妥当性はとりあえず確認されたことになる。因子分析による新しい尺度については（新たな理論の構成とそれにもとづく）妥当性の検討が必要になる。

しかしながら，こうした手続きとその背景となる論理には，いくつかの問題がある。本稿はそうした問題点に関するものである。

用語と記号の定義

まず，若干の用語と記号を定義しておく。複数の問題項目（あるいは質問項目）を集めて，多数の被験者に対して実施できる形にしたものをテストと呼ぶ。テストを受験した個人の個々の項目に対する反応を，決められた方法で数値化したものを項目得点，項目得点を定められた方法で合成，総合して得られる得点の次元（数値が定義される座標軸）を尺度，個々の得点を尺度得点と呼ぶ。何らかの個人差に関する仮定された量的な次元のうち比較的永続的なものを特性（trait），一時的で可変的なものを状態（state）と呼ぶ。ある特性に一定の理論的枠組みの中で意味を持った命名がなされた場合には，それを構成概念（construct）と呼ぶ。特性にも構成概念にも，必ずしもその操作的定義（それを測定するためのテストと尺度の定義）が，一意的に対応しているとは限らない。

次に，項目iに対する反応をx_i（$i=1,\ldots,p$）と書き，項目得点と呼ぶ。通常は，被験者が項目iに正答したとき1，誤答のとき0となる。ただし，ここで考えるテストは，必ずしも能力を測定するものに限定しないから，質問項目に対して，測定をめざす特性について得点の高くなる方向での回答がなされた場合に1，逆方向の回答を0とする，というコード化も考えられる。また，何段階かのグレードのついた回答選択肢がある場合には，それらをそのまま整数値でコード化したものも含めて考えることにする。いずれにしても，x_iの値は被験者ごとに変動する確率変数とみなされる。

テスト得点とその性質

尺度得点Xの定義としてしばしば用いられるのは，項目得点の単純和，すなわち，

$$X = x_1 + x_2 + \cdots\cdots + x_p \qquad (1-1)$$

である。Xはしばしばテスト得点と呼ばれる。

Xには，変動が存在するが，それがすべてランダム誤差によるものでないことを，ある意味で保証するのが，項目間の相関の存在である。それらは，通常，項目得点の背後にそれらの変動を支配している見えない次元の存在を仮定することを正当化する。その点で，項目間相関は，それらの項目によって何らかの特性が測定できることの根拠である。テスト得点の信頼性係数の推定値として用いられるクロンバックのα係数は，項目数pと項目間相関係数の平均値$\bar{r}$を用いて，近似的に次のように表現される。

$$\alpha = \frac{p\bar{r}}{1+(p-1)\bar{r}} \qquad (1-2)$$

このαは，項目間相関の平均値$\bar{r}$の増加関数である。また，αは，$\bar{r}$が正の値をとる限り，項目数pの増加関数で，

$$p \to \infty \quad \text{ならば} \quad \alpha = 1$$

でもある。これは，多数の項目が加算される結果，個々の項目得点に含まれるランダム誤差が，テスト得点においては相殺され縮小することを意味している。

尺度構成上の問題点

こうして，大きなαの値をもつことは，その尺度がともかくも何らかの特性（少なくともランダム誤差以外のもの）を反映していることの証拠となりうる。しかしながら，なお幾つかの問題が残る。たとえば，次のような現象はしばしば経験される。

(1) 古典的テスト理論では，合成変量であるテスト得点には簡単な1次元モデル，

実測得点 ＝ 真の得点 ＋ ランダム誤差

が仮定される。この想定が正しいという前提の下で，信頼性係数は，

$$\text{信頼性係数} = \frac{\text{真の得点の分散}}{\text{実測得点の分散}}$$

で定義される。先のモデルが正しいなら，この値が小さいほど，(実測) 得点と他の変数との相関は低くなる（希薄化）はずであるが，しばしば信頼性の低い得点の方が高い相関を示すことがある (bandwidth-fidelity dilemma と呼ばれる)。

(2) 因子分析の実際のデータへの適用において，以下のことがしばしば観察される。

① 確認的因子分析における有意な因子数は，サンプルサイズが大きくなると，増加する傾向がある。また，変数を追加すると，有意な因子数は増加する。

② 変数を追加した場合，因子の数が同じであっても，抽出される因子の性質は変化する。つまり，因子の内容は，選択された変数に依存する。

③ ある項目に極めて類似した内容の項目（つまり，その項目と極端に相関の高い項目）を分析に追加すると，因子数は増加する（村上，1989)。これは，通常，因子分析の不適切な使用とみなされるが，その根拠ははっきりしない。

(3) 項目間の相関係数は通常かなり低い。このことは，項目反応に極めて大きなランダム誤差が含まれていることを意味するが，統制されたテスト条件の下では，通常，そのような大きなランダム誤差の

原因が見当たらない。たとえば，多肢選択形式のテストに対する反応をよく見ると，いかに低得点の被験者も，ランダムに反応しているわけではないことがわかる。実際，多くの項目では，低得点者の反応が集中する選択肢が存在する。ともかく，被験者の反応の中で，ランダムと言える部分の大きさは，モデルから算出されるより小さいと考えるのが自然である。

モデルの問題点

項目水準のデータを，古典的テスト理論，あるいは，因子分析法を用いて分析することについて，従来，主として問題にされてきたのは，量的連続体である特性への，離散的な項目反応の回帰が非線型にならざるを得ないという問題であった。実際，古典的テスト理論のモデルを項目水準に適用すると，

$$x_i = t + e_i \qquad (1-3)$$

となる。ここで，tは真の得点，e_iはランダム誤差である。また，因子分析モデルを項目得点にあてはめると，

$$x_i = \sum_{l=1} a_{il} f_l + u_i + e_i \qquad (1-4)$$

となる。ここで，a_{il}は項目iのl番目の因子における因子負荷量，f_lは対応する因子上での得点（因子得点），u_iは項目iの固有成分の得点である。いずれにせよ，右辺のモデルは連続量として定義されるにもかかわらず，左辺は，2値，あるいは，せいぜい数段階のカテゴリーであって，無理にこのモデルを適用すれば，ランダム誤差に不自然な性質を付加することになったり，因子分析における難しさの因子（difficulty factor）の出現のような問題が生起する（たとえば，Comrey, 1973, 芝訳, 1979）。

これに対して，項目反応が特定の値（通常は1）をとる確率が，対応する特性に対して非線型に回帰することを仮定する項目反応理論（item response theory，たとえば，芝，1992）等があらわれ，少なくとも1次元の場合について問題は解決したかに見える。

しかしながら，先にあげた尺度構成上の問題点(1)〜(3)等は，この方法によって完全には解決できないと思われる。問題は，個々のモデルを越えて共有されている「一つあるいは少数の特性次元だけが，項目反応を（組織的に）規定している」という想定にある。項目反応に対するモデルが不合理なものである結果として余分な次元が出現することよりも，本来的に多数存在する次元を少数の固定した存在として認識しようとする枠組みの方が，テストと，その結果を用いた研究の解釈や教育の評価に，悪影響をもたらしていると思われる。

本研究では，心理測定尺度や因子分析における因子に関する，先にあげたような問題を，少なくとも定性的に理解できるような，より自然なモデル，あるいは，思考の枠組みを提案したい。

項目反応に影響する要因は多数である，ということを認めるところから出発したら，どのようなモデルができるかを次に考えよう。

2．家族的類似モデル

項目反応を規定するものとしての特徴

項目反応が，多数の要因によって決定されていることは，多くの心理学的個人差測定の研究者が常識的に想定しているところである。たとえば，Messick（1975, p. 955）はこれを，「単一の……項目反応は……ほとんど確実にmultiply determinedである」と表現している。

そうした状況を表現するモデルは，何通りも考えられるし，考えられるあらゆる側面を盛り込もうとすれば，極めて複雑なものになってしまう。ここでは，比較的一般的な形のモデルから出発して，ある程度意味のある命題が引き出せる水準まで単純化をはかっていくことにしよう。

まず，ある心的特性の測定を目指すp個の項目からなるテストを考える。これらの項目に対する個々の被験者の反応を規定する，n個の量的な確率変数，y_1，y_2，...，y_nを考え，これらを特徴（feature）

と呼ぶ。その際，すべての項目が，すべての特徴を反映するわけでなく，反応に影響する特徴は，項目ごとに異なるとする（図1）。ここで，ある（テストが実施される）時点における，テストに対する反応と関係する被験者の属性は，それらの特徴の値の列と同一視できる。ここで，特徴の数nは，項目の数pと比較してはるかに大きいものとする。

なお，心理測定では，個人差（分散）だけが問題だから，個々の特徴の平均値はすべて0としても一般性を失なわない。すなわち，特徴を要素とするn次元ベクトルをyとするとき，

$$E(\boldsymbol{y}) = \mathbf{0} \qquad (2-1)$$

と仮定する。

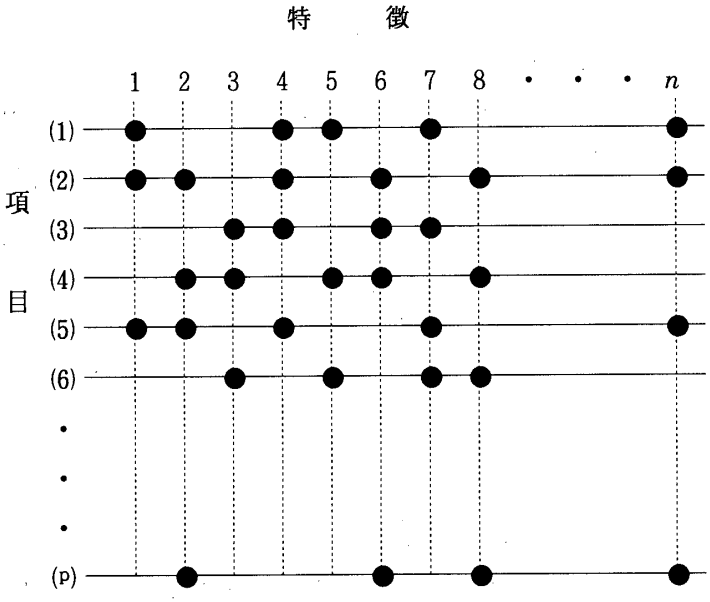

図1　項目反応と特徴との関係に関するモデル。項目反応は，（2－8）により，原理的にはすべての特徴によって規定されるが，ここでは特に大きな重みを●によって表現している。

特徴の分類

ここで，個々の特徴のすべてが，測定を目指す心的特性と関連するものであるとは限らない。それぞれの特徴は，おおよそ，図2のように分類することが可能であろう。

まず，通常仮定されるように，項目得点は真の得点とランダム誤差をともに反映しているであろう。ランダム誤差は，原因を特定することもコントロールすることもできない偶然的かつ一時的な変動である。尺度得点が，真の得点を反映する度合いが，理想的に言えばα係数ということになる。

つぎに，真の得点は，個人差のうち，ある程度永続的な次元である特性と，一時的なものである状態とを，ともに反映するであろう。この特性部分だけの分散の大きさを示すのが（諸仮定が成り立つとして）再検査信頼性係数である。

また，特性の中にも，測定を目指す構成概念を反映する部分と，それとは別の特性（ここでは一括して恒常誤差と呼ぶ）を反映する部分が区別される。構成概念を反映する分散の大きさを推定できたとすれば，その全分散の中に占める割合が（理想的な意味での），妥当性係数（coefficient of validity）ということになろう。

さらに，構成概念も複数の下位概念からなり，恒常誤差も複数の特性（単なる反応傾向を含む）からなっているであろう。状態もまた，いくつかの特徴からなるはずである。それらの個々の成分を本稿ではいずれも特徴と呼ぶわけである。

後述するように，これらのカテゴリーの境界は必ずしも明確ではないが，この分類をあえて示すのは，とりあえず概念を整理しておくためである。たとえば，古典的テスト理論の基本モデル，

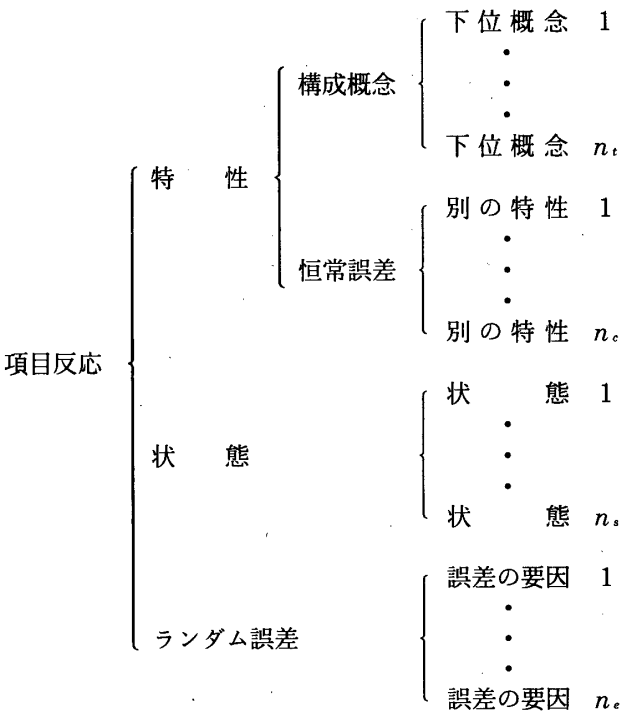

図2　項目反応に影響を与える多数の要因とその分類。特性と状態を合わせたものが真の得点である。

（1－3）や因子分析の基本モデル（1－4）において，真の得点，あるいは因子得点には，測定を目指す特性以外に，恒常誤差と状態が含まれている可能性がある。因子分析における独自成分は，特性と状態のうちその項目にのみかかわる部分を原理的に分離できると想定したものである。それが，その項目だけの恒常誤差であると考えるのは，希望的観測に過ぎない。

特徴の正準型

各特徴の母分散は相互に異なり，特徴相互間の相関も高低さまざまである可能性がある。しかしながら，適当な直交変換を通じて，相互に直交する正準型（canonical form）を作ることができる。適切な直交行列を$\boldsymbol{T}$とするとき，正準型に変換された特徴，c_1，c_2,...，c_nを要素とするn次元ベクトル$\boldsymbol{c}$は，

$$\boldsymbol{c} = \boldsymbol{T}\boldsymbol{y} \qquad (2-2)$$

と表すことができる。

このTの各行は，$\boldsymbol{y}$の共分散行列，$E(\boldsymbol{y}\boldsymbol{y}')$の全ての0でない固有値に対応する固有ベクトル，およびそれと直交する任意のベクトルである。ここでは，簡単のために，固有値の中に0のものはなく，かつ固有値はすべて相異なるとしよう。そこで，$\boldsymbol{c}$は$\boldsymbol{y}$のすべての主成分得点に対応し，かつ（符号を除き）一義的に定まる。

なお，（2－1）により，

$$E(\boldsymbol{c}) = \boldsymbol{0} \qquad (2-3)$$

であり，次で定義される$\boldsymbol{c}$の共分散行列Δ_cは対角行列となる。

$$\Delta_c = E(\boldsymbol{c}\boldsymbol{c}') \qquad (2-4)$$

項目反応を決定する関数

つぎに，特徴と項目得点の関係について規定しておく必要がある。ある特徴のパターンを示す個体が，テストの項目iにおいて値x_iをとる確率は，一般的に次のように述べることができよう。

$$\text{値}\ x_i\ \text{の生起確率} = f_i(y_1, y_2, \ldots, y_n) \qquad (2-5)$$

ここで，f_iは項目ごとに異なる関数である。

さらに，f_iは2つの段階に分けて考えることができる。第1の段階は，特徴をそれぞれの項目に対応する特性値τに変換するもの，

$$\tau = g_i(y_1, y_2, \ldots, y_n) \qquad (2-6)$$

である。このg_iを特性関数（trait function）と呼ぼう。第2段階は，τと反応x_iの関係を規定する反応関数h_{xi}である。

$$\text{値}\ x_i\ \text{の生起確率} = h_{xi}(\tau) \qquad (2-7)$$

ここで，特性関数も反応関数も，項目ごとに異なったものである可能性が高い。反応関数は，特性値を反応確率に変換する役割をもつ。

モデルの単純化

一般的なモデルのままでの議論は複雑になるので，単純化を図ろう。まずここでは，反応関数hの形態については考慮外とし，τをx_iと等価と見なす。このことにより，以後の議論の中では，項目得点は連続変量として扱われる。

そして，g_iについても簡単に，特徴の1次関数と仮定する。項目iにおける特徴kへの重みをv_{ik}と

書く。したがって，

$$x_i = \sum_{k=1}^{n} v_{ik} y_k \qquad (2-8)$$

これを，テスト得点の定義，（1－1）に代入すると，

$$X = \sum_{k=1}^{n} \left(\sum_{i=1}^{p} v_{ik} \right) y_k \qquad (2-9)$$

となる。また，特徴の正準型の得点 c_l は y_i の1次関数だから，

$$x_i = \sum_{l=1}^{n} w_{il} c_l \qquad (2-10)$$

のように，x_j は c_l の1次関数となる。再び，これをテスト得点の定義，（1－1）に代入すると，

$$X = \sum_{l=1}^{n} \left(\sum_{i=1}^{p} w_{il} \right) c_l \qquad (2-11)$$

となる。

以下においては，このような単純化された想定のもとに，こうしたモデルから導かれる事実を検討していこう。

内的整合性の規定因

前述のように，式（1－1）で定義されるテスト得点が何らかのランダム誤差以外のものを反映していることを，データの上で保証するものは，項目の内的整合性，すなわち，ゼロでない項目間相関の存在である。通常，そうした整合性は少数の特性によって説明できることが期待されており，その特性を表現するのが尺度である。Messick（1975）は，「項目間の……整合性は，……しばしば少数の決定因を，時には単一の主要な決定因をもつ。」と言う。個々の項目に関わる要因（本稿でいう特徴）は多数あるが，それらは，結局少数の特性へと要約できると考えるわけである。多数の特徴を想定した場合に，こうした様相はどのように実現するかを考えよう。

古典的テスト理論の想定

各項目ごとのモデル（2－8）において，どの項目についても，2つの重みだけが1，それ以外はすべて0であるとしてみよう。さらに，重みが1である特徴の一方は，必ず y_1 であり，他方は，項目ごとに異なるとしよう（図3）。この場合，項目反応に関わる特徴の数は（$p+1$）個である。さらに，それらは相互に無相関であるとする。

そうすると，（2－8）は，古典的テスト理論の基本モデル，（1－3）と全く同一になる。テスト得点（2－9）は，その分散の多くを，py_1 の形をとる「真の値」が占めるから，項目数 p が増加するほど，（1－2）で定義される信頼性は増加することになる。

この平行測定の仮定は，すべての項目間共分散が一定であるという非常に強い条件が，データの上で満たされていることを主張するものである。実際のデータで，こうした仮定が満たされることはほとんど期待

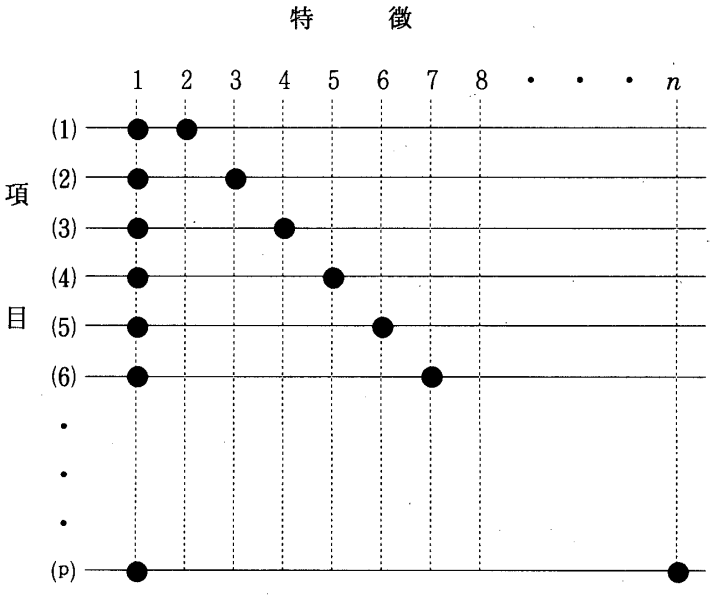

図3　古典的テスト理論における平行測定モデル。全ての特徴は相互に独立である。また，この図の条件では，$n = p + 1$ となる。

できない。しかも，y_1 が単なる状態でなく多少とも永続的な特性であるかどうかは，測定の反復によって確かめないかぎり，1回限りの測定データでは明らかでない。さらに，その条件が満たされても，y_1 が測定を目指す特性そのものであるかどうかはわからない。すなわち，図2における「構成概念」のカテゴリーに含まれるかどうかはわからない。つまり，得点の妥当性は保証されない。

また，他の特徴と独立である限り，個々の項目にかかわる特徴の数は，いくら多くてもかまわない。先の Messick（1975）による「個々の項目反応は，multiply determined であるが，合計点としてのテスト得点の決定因は単一でありうる」という命題は，ほぼこうした事態に対応すると考えられる。なお，それらの特徴の中に単なるランダム誤差でないものがあるとすれば，その項目得点の安定性にそれも寄与することになる。

このように，1回の測定では，平行測定の仮定を満足する単純なデータと見えるものにも，かなり複雑な様相が隠されている場合もありうる。そうした複雑性を，少なくとも概念的には整理できるという意味で，図2のような分類には一応の意義が認められよう。

特徴間の相関にもとづく特性

次に，正準化された特徴，c_1，c_2，……，c_n の中に，際立って分散が大きいものがあるとしよう。それが c_1 であるとしても一般性を失わない。一方，各項目ごとのモデル（2－8）においては，多くの特徴に対して0でない重みが存在するが，それらは項目ごとに異なるとする。ただし，特徴は2つのグループに分割でき，その一方は，c_1 に対して大きな寄与をなすもので，他方は，そうでないもの（すなわち，相互にほぼ無相関なもの）とする。どの項目についても，2つのグループのどちらにも0でない重みが存在する特徴がある（図4）。

この場合，項目は相互に異なる特徴に対応しているけれども，特徴間の相関が高いことを通じて，実質的に同一の意味をもつことになる。すなわち，（2－10）は，正準化された特徴の1つである c_1 を真の得点とする古典的テスト理論のモデルと近似的に一致し，対応するテスト得点の信頼性は項目数の増加とともに高まる。ただし，この場合，すべての項目が，信頼性に対して同等の寄与をするわけではない。

さらに，この場合，特性自体の性質に，一種の「幅」（3節参照）が出てくることに注目すべきである。特性の内容は，選ばれた項目によって微妙に変化する可能性がある。

なお，幾つかの大きな分散をもつ c_l がある場合，因子分析の直交解のモデルがほぼあてはまり，かつ単純構造に近い負荷行列が得られるはずである。その結果，p 個の項目から，複数の尺度が得られることになる。

本稿の最初に述べた尺度構成の過程で，暗黙裡に想定されているのは，このような事態であろう。同一の c_l に関わる特徴に支配される項目によって一つの尺度が定義され，他の項目の特徴と高く相関しない特徴に支配される項目は，どの尺度にも

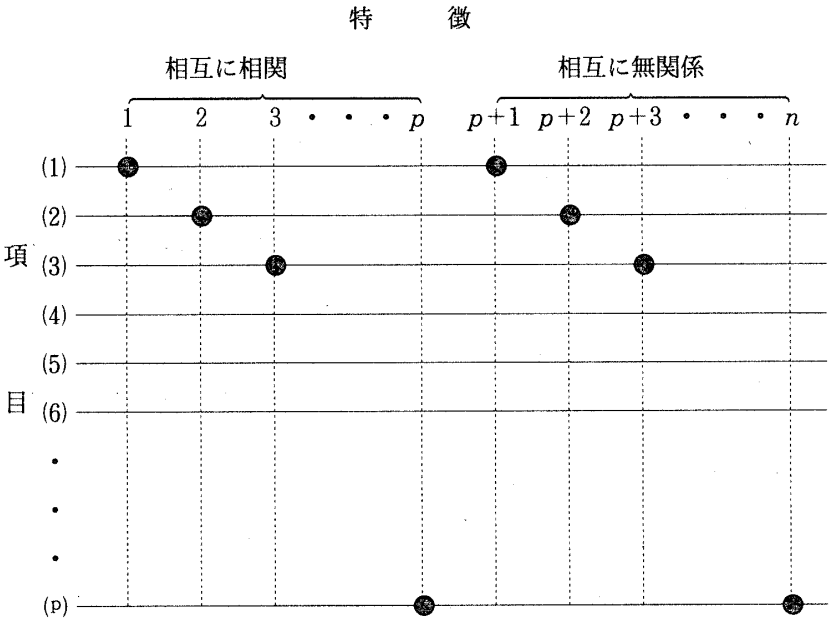

図1　特徴間の相関にもとずく準平行測定モデル。特徴は相互に相関するものと相互に独立なものにわかれる。この図の条件では，$n = 2p$。

含められず，削除される。

変数間の特徴の共有にもとづく特性

次に，特徴間全体が相互相関をほとんどもたず，正準化してもさほど大きな分散をもつものが出現しない場合を考えよう。この場合，各項目が非ゼロの重みをもつ特徴が2つ程度であるとすれば，ほとんど項目間相関は0となり，（1－2）による信頼性係数は低いものとなろう。しかし，図1のように，複数の項目が多数の特徴を反映し（すなわち，多数の重みがゼロでなく），かつ，それらが複数の項目に共有されているとすれば，それによっても項目間相関は生ずることになる。

形式的には，項目間相関は，次のようにあらわされる。

$$r_{ij} = \frac{\sum_l w_{il} w_{jl} \lambda_l}{\sqrt{\sum_l w_{il}^2 \lambda_l} \sqrt{\sum_l w_{jl}^2 \lambda_l}} \qquad (2-12)$$

ここで，λ_lは，正準化された特徴c_lの分散，すなわち，yの共分散行列の大小準にl番目の固有値である。すなわち，この条件では，二つの項目が同じ正準化された特徴（特に，分散の大きいもの）に大きな重みをもっているほど，相関が高くなる。すなわち，多少厳密さを犠牲にした言い方をすれば，多くの特徴を共有しているほど，項目間相関が高くなる。このようなメカニズムによる項目間相関は，単純な構造をなさない可能性が高いが，それでも，さまざまなモデルが適合するパターンが生成されうる。

一つのテストを構成する項目は，こうした性質をもっている可能性が高い。これらの項目は，Wittgensteinの意味における家族的類似をなしていると見なすことができる。家族的類似の概念は，村上（1989）にも引用した，Bloor（1983，戸田山訳，1988）に明快に表現されている。

「一つの概念に属するものの実例を調べてみると，我々はしばしば，それらを一まとめにするための手掛かりになるような，成員全てに共通する性質なるものは存在しないことに気付く。その代わりに見出されるものは，時に，全面的な類似であり，時に，細部における類似であるという具合に『互いに重なりあったり，交差し合ったりしている複雑な網の目なのである』。……家族の成員は互いに似通っているが，それは全員が，例えばだんごっ鼻といった，ただ一つの性質を共有しているからとは限らない。似通っているという印象は，例えば娘は父親譲りの顎を持ち，他方その息子は彼女と同じ鼻を持つが髪の毛の色は叔父と同じであるという風な事実に由来する方が多いのである。」（p. 47）

相互にほぼ無相関な複数の特徴を，項目間で共有しあうことによって項目間相関が生ずるという想定に依って立つモデルを，家族的類似モデル（family resemblances model）と呼ぶ。すなわち，（2－8）において，すべてのy_kが相互に（ほぼ）無相関であると仮定するのが，家族的類似の基本モデルである。

この場合，共有される特徴は，少なくともテストの実施時間中は安定したものでなければならない。

テスト理論の仮定の何が受入れ難いか？

家族的類似モデルは，従来のテスト理論，すなわち，古典的理論のみならず，新しい項目反応理論における，項目得点の背後にただ1つの（あるいは，少なくとも特定できる少数の）特性が存在する，という前提の否定の上に成り立っている。この前提は，先に言及した，項目反応はmultiply determinedである，という想定と直ちに矛盾するものではない。共通する一つ（あるいは少数）の特徴以外は，すべてその項目独自のもの（ランダム誤差を含む）であるとすれば，1次元性は成り立つ。あるいは，共有される特徴の数は多くても，それらが相互に高い相関をもつとすれば，すなわち，共有される正準化された特徴は一つであるとすれば，従来の理論はほぼ守られる。

しかしながら，実際問題として，多数の特徴のうちの一つだけが，複数の項目によって共有されるといった仮定は，認めがたいものである。データを素直に見る限り，家族的類似モデルの正当性は疑う余地がないように思われる。たとえば，かなり高いα係数をもつ一つの尺度を構成する項目間の相関を調

べてみると，全般に正の相関は維持されているものの，その大きさは様々であり，幾つかの項目間の相関は0であり，あるものは負であることに気づく。項目間相関行列の固有値は，最大のものが他を圧して大きいとしても，2番目以降のそれも無視できない大きさであり，かつ，明確な切れ目もなく漸減している。結局一つのテストに含まれる項目は，相対的関係のネットワークを成していると考える方が自然であるように思われる。

家族的類似モデルの反証不能性

しかしながら，家族的類似モデルは，その正当性を統計的に立証できる性質のものではない。変数の数よりはるかに多い個々の特徴の値を，データから推定することは不可能であるだけでなく，特徴の数そのものをデータから決定することもできない。すなわち，家族的類似モデルは，そもそも反証可能な体裁（どのようなデータが出現したら，理論の誤りが証明されたことになるかを明示すること）をとっていないから，科学的理論たりえないという批判は成立しうる。

それにもかかわらず，このモデルによって，尺度構成における多くの問題点が説明可能になるし，心理測定尺度の性質について新たな視野が開けるように思われる。たとえば，次のような単純な例を考えてみよう。

3つの項目があり，それぞれ，3つの正準化された特徴を1つずつ共有するものとしよう。すなわち，

$$\left.\begin{aligned} x_1 &= c_1 + c_2 \\ x_2 &= c_1 + c_3 \\ x_3 &= c_2 + c_3 \end{aligned}\right\} \qquad (2-13)$$

である。また，それぞれの正準化された特徴の分散は同一であるとしよう。すなわち，

$$\lambda_1 = \lambda_2 = \lambda_3 \qquad (2-14)$$

である。このとき，項目間の相関係数は，

$$r_{12} = r_{13} = r_{23} = 1/2 \qquad (2-15)$$

のようにすべて同一の値をとる。なぜなら，2つの項目iとjが，ともに，相互に独立で，一定の分散をもつ変量の和として表現され，それぞれに関連する変量の数を，n_i，n_j，両項目が共有する特徴の数をn_{ij}とするとき，項目間の相関係数は，

$$r_{ij} = \frac{n_{ij}}{\sqrt{n_i\, n_j}} \qquad (2-16)$$

となるからである（たとえば，芝，1975，pp.10-11）。式（2－13）においては，

$$n_1 = n_2 = n_3 = 2$$
$$n_{12} = n_{13} = n_{23} = 1$$

だから，（2－15）が得られる。

この例の注目すべき点は以下の通りである。この相関行列は，全ての変数の負荷が，

$$a_i = 1/\sqrt{2} \qquad i = 1,\ 2,\ 3 \qquad (2-17)$$

であるような，1因子モデルによって完全に説明できる。すなわち，

$$x_i = a_i f + e_i \qquad i = 1,\ 2,\ 3 \qquad (2-18)$$

という基本モデルが完全に適合することになるわけだが，これは（2－13）とは両立しない。特に，（2－18）では，モデル部分と誤差部分が完全に分離されているが，（2－13）においては，それぞれの

得点の構成要素を，モデルと誤差に区別することができない（ただし，反復測定のデータがあれば，また事情は変わってくる）。

通常，家族的類似モデルから予測されるような相対的ネットワークから生ずる項目間相関は，単純で一貫したパターンは示さないが，特殊な場合として，このような1次元モデルに一致するケースも生じうる。このことは逆に，1因子，あるいは有限のある因子数が十分な標本の大きさの下で受入れ可能であるようなデータが存在したとしても，家族的類似モデルを対立仮説として考える限り，必ずしも，真の得点＋ランダム誤差，という形のモデルが立証されたことにはならないことを示すものである。

そのことは，因子分析や古典的テスト理論の基本モデルにおいて，共通成分と独自成分の境界が，一般的にかならずしもはっきりと定められないことを意味する。この点については，3節で再度検討する。

1次元性が成立しない理由

ここで，1次元性，あるいは固定された有限次元性が，実際のデータにおいてなぜ不可能なのかという点について，多少具体的に考えてみよう。繰り返しになるが，項目反応に影響する要因は極めて多い。たとえば，Lazarsfeld（1966，西田他訳，1984）は，

「われわれは指標の使用にあたって固有の偶然的な要素を克服できるように努める。ひとびとが政府の経済状況に関する役割をどのように感じているかにしたがって，彼らを順序づけてみたいと仮定しよう。われわれは彼らに鉄道・鉱山・銀行などの公共所有について一連の質問をする。政府の無干渉主義を好めば好むほど，その人が（公共所有を）肯定する項目数はより少なくなると当然仮定するだろう。それにもかかわらず，われわれは個人の特異な面が回答に入りこんでくることを知っている。たとえば，無干渉主義を強く信奉する人が鉱山事故について新聞で読んだばかりで，この影響のもとで鉱山の項目に対して彼はイエスまたはノーと断定的に回答をする。一方，干渉主義を強く信奉する人がたまたま立派な銀行の頭取を知っていて，銀行についての項目にイエス・ノーの断定的な回答を控えることがある。」

という例をあげる。Lazarsfeld自身は，個々の要因を偶然的な要素と呼び，こうした要素を相殺するために，尺度構成の必要性を説いているのである。しかしながら，ここで述べられているような原因は，単純なランダム誤差，あるいは，特定の一項目のみに影響すると考えることが難しいものである。

能力テストにおいても，こうした要因は多い。たとえば，Vernon（1962）は，文理解のテストについて，その妥当性を損なう幾つかの要因をあげている。それはたとえば，1）読みの速度，2）動機づけの強さ，3）機械的記憶力，4）語彙能力，5）受験技術，6）内容に関する既有知識，等である。こうした側面は，もはや偶然的なものと考えることのできない，被験者の特性そのものであり，複数の項目にわたって影響を与えるであろうものである。

次元に関する性質としての家族的類似

念のために，次の点は指摘しておきたい。本来の家族的類似の概念は，個物とその集合としての概念の関係について述べられている。それに対し，ここでは，項目とその総合としての尺度に対応する特性との関係について論じている。すなわち，先のBloorの例で言えば，個々の人に当たるのがテストの項目であり，家族にあたるのが，テストが測定する特性である。測定の対象となる個人，すなわち個体の分類を論じているのではない。

3．家族的類似モデルの含意

家族的類似モデルは，現在のところそれ自体として何らかの新たな分析方法や，得点化の方法を提供するものではない。しかし，テストの作成や得点の解釈に対する示唆には無視できないものがあると考える。この節では，それについて論じよう。

特性の境界の不明確化

家族的類似モデルの第一の主要な帰結は，特定の特性の測定のための項目のプール（項目ユニヴァー

ス）が明確に規定できなくなることである。

項目ユニヴァースとは，ある特性の測定に用いられる項目全体の（仮想的な）集合のことである（Cronbach, 1951）。たまたまテストを構成している項目（にもとづく尺度）を越えて，特性一般について語ろうとすると，こうしたものを想定せざるを得なくなる。もちろん，項目ユニヴァースは操作的に具体化できるものではない。すなわち，いかなる特性についても，それを測定するための質問項目，あるいは問題項目を列挙しつくすことは誰にもできない。しかし，もしどんな項目に対しても，(1－3)のような平行測定モデルが当てはまるなら，すべての存在しうる項目は，相互に重複のないカテゴリーのどれかに属することになろう。

このような想定が可能になるのは，項目ユニヴァースが，古典的カテゴリーとして概念化されているためであると考えられる。Rosch（1987）によれば，古典的カテゴリーとは以下のような性質をもつものであるとされる。

(1) カテゴリーは正確でなければならない。すなわち，境界がはっきりしていなければならない（つまり，ある対象がカテゴリーの成員であるかどうかは，一義的に判断できなければならない）。

(2) カテゴリーのすべての成員は，何か（同一物）を共有していなければならない。

(3) カテゴリーへの帰属度に関して，成員間に差はない。

平行測定モデルに一致する項目からなるユニヴァースは，この古典的カテゴリーの特徴をもっている。すべての項目は，同一の真の得点を共有しており，ある項目が特定の項目ユニヴァースに属するか否かは（原理的には）一義的に判定できる。一因子モデルは，項目ごとに負荷量が異なり得るから，(3)の性質は失っているが，全項目は共通の因子に負荷しており，(2)の特徴はもっている。多因子モデルでも，個々の項目は単一の因子にのみ目立つ負荷をもつ，という典型的な単純構造の下では，(2)の性質は維持される。

Rosch（1987）は，上記の(1)～(3)の性質をすべて否定したところに，家族的類似にもとづくカテゴリー化の概念が生まれるとする。実際，相互に異なった特徴を共有することによる項目間のネットワークは，事実上どこまでも繋がり得るから，必然的に一つの特性に対応するカテゴリーの境界は不明確なものとならざるを得ない。

項目の側から見れば，家族的類似のモデルは，一連の項目反応の背景にある特徴の数が確定できないことを意味している。さらに，図2に示したような特徴の分類をもあいまい化してしまう。たとえば，先の（2－13）では，ランダム誤差とそれ以外の組織的変動の区別すら，原理的に困難であることが例示された。

尺度の「幅」としての領域範囲

このように，項目ユニヴァースが明確に定義できなくなる結果，心理測定尺度には，α係数等によって評価されるような内的整合性とは別の評価基準として，いわば，領域範囲（domain coverage）とでもよぶべきものが存在することになる。領域範囲とは，尺度得点が反映している特徴の数である。より正確に言えば，（2－11）の形をとるテスト得点の，比較的分散の大きい正準化された特徴c_lに対し，大きな値をとる重み（$\sum_{l=1}^{p} w_{il}$）の数（の多さ）である。すなわち，尺度の内容が純粋でなく，多様なものを反映している度合いのことである。心理測定データは，しばしば，「大量のノイズを含む」と言われることがあるが，領域範囲という考え方からすれば，むしろ「多数の相互に独立なシグナルを含む」という方が正しいであろう。このことは，1節に(3)としてあげた，統制された条件下で余りにも大きな無作為誤差の原因が見当たらない理由の説明にもなろう。

領域範囲は，必ずしも内的整合性と相反する基準ではない。もし，尺度の定義に用い得る項目が相互に共有している特徴が少ないならば，領域範囲は大きく内的整合性の低い尺度ができあがることになろう（図5－(a)）。しかし，項目が相互に共有している特徴の数が多ければ，項目間相関は高くなる結果，（1－2）で定義される内的整合性の度合いは高くなる（図5－(b)）。他方，相互に極めて類似度の高い

項目を尺度定義のために採用すれば，項目間の相互相関は高いから内的整合性は高くなるが，それらは少数の特徴を共有しているにすぎないから，領域範囲は狭くなる(図5－(c))。

尺度構成にあたっては，内的整合性のみならず，本稿で言う領域範囲も考慮すべきであることは，実践的には従来も強調されてきた。実際，データ解析の現場感覚としては，これはほとんど常識ですらある。たとえば，もし古典的な想定が正しいならば，一組の変数の主成分分析の結果として得られる合成変量（主成分）は，それらの変数の，何らかの外在基準に対する重回帰分析による合成変量（予測変量）とほとんど同じものになるはずである。しかしながら，通常これらの間には著しい違いがあるのが普通である。

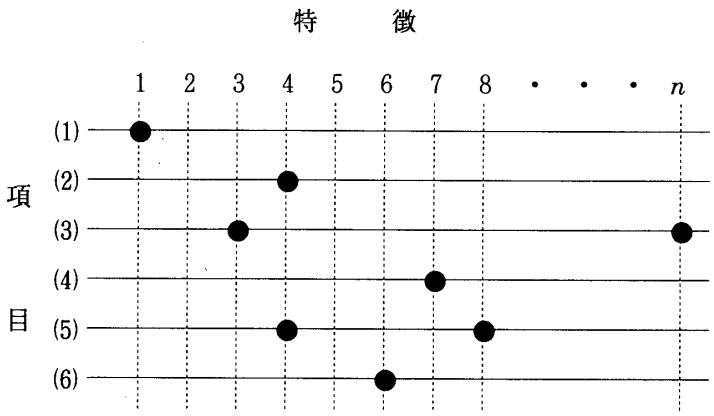

図5－(a)　領域範囲が広く内的整合性の低い尺度を構成する項目

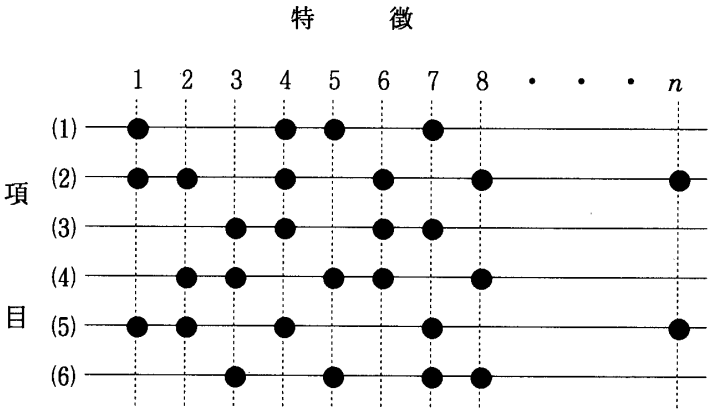

図5－(b)　領域範囲が広く内的整合性の高い尺度を構成する項目

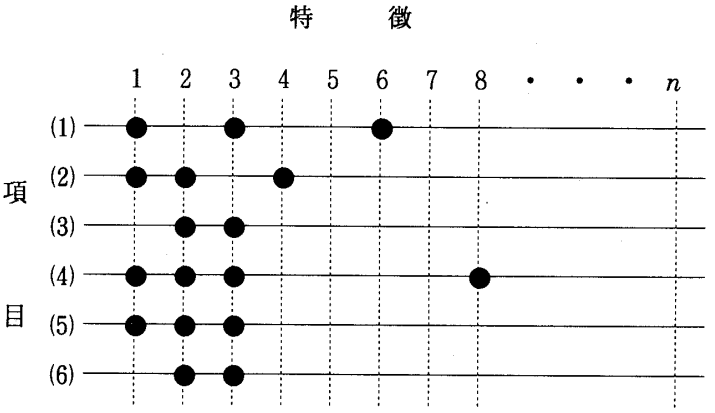

図5－(c)　領域範囲が狭く内的整合性の高い尺度を構成する項目

領域範囲はなぜ無視されてきたか

しかしながら，従来，領域範囲について，明確に指摘されたことはあまりなかったと思われる。その理由の一つは，領域範囲の大きさをデータから推定する確実な方法がないことにある。尺度の領域範囲は，他の複数の尺度との相関の大きさ等から間接的に推測しうるにすぎない。しかし，それとともに，領域範囲自体を明示的に表現するモデルが存在しなかったことも一因であろう。

そして，現在でも内的整合性と領域範囲が異なる側面であることを全く無視した議論も存在する。たとえば，最近でも入門的テキストに書かれている「項目－テスト相関は，その項目の妥当性の指標と見なし得る」（たとえば，Henning, 1987, p. 24）という命題がそうである。常識的に言えば，項目－テスト相関は，その項目が他の項目と全般的に高い相関を有するかどうかを示し，その項目がテスト得点の内的整合性の向上に寄与している程度を示すものである。家族的類似モデルの観点からすれば，その項目が，特徴を他の項目と共有している度合いである。しかしながら，テストが総合的に表現するものが，単一の次元以外にありえないと考えれば，その項目が当該の次元を反映する度合いとして，項目－テスト相関を妥当性の指標と見ることは確かに正当化できる。

また，McDonald (1985) は，(2－4) が成立するなら，次で与えられるω係数，

$$\omega = \frac{p(\bar{a})^2}{p(\bar{a})^2 + u^2} \qquad (3-1)$$

は，(1)テストの信頼性係数の下限であり，(2)構成概念妥当性の指標であり，(3)一般化可能性の指標でもある，と論じている（p. 222)。ただし，$\bar{a}$ は因子負荷量の平均値，u は特殊成分の平均値であるから，これは全項目の合計点の分散中に占める，因子によって説明される分散の大きさに相当する量であり，一種の信頼性係数と見ることができるから，(1)の性質は承認できるとしても，(2)と(3)の根拠は，専ら1次元性の仮定にある。すなわち，こうした議論は，項目反応の本質的多次元性をあえて無視するところに成立していると考えざるを得ない。

考慮すべき項目の範囲

さらに，項目ユニヴァースの範囲が確定できないという問題は，実は本稿のここまでの議論にも大いに影響する。実際，2節における形式的な議論は，すべて p 個の項目が選択された段階を出発点としていた。項目数が増加した場合には，当然，それに対応して特徴の数も増加するし，正準化された特徴の内容も変化する。もし，項目ユニヴァースを原理的にせよ確定することができるならば，対応する特徴の数も原理上は確定し，それに対する正準化された特徴の「枠組み」も固定したものとして議論することができる。しかし，項目の選択される範囲がはっきりしないとすれば，（2－10）や（2－11）の表現する内容は，大きく変化する可能性をもっているわけである。

特性は実在するか

しかし，問題はそれにはとどまらない。上記の議論では，いずれにせよ，それぞれの特性の存在自体は前提とされていた。しかしながら，家族的類似モデルでは，カテゴリーを定義するための手掛かりが失われていることに気づかざるをえない。「家族的類似」についての議論が可能なのは，「家族」そのものが，個々の成員の特徴に先立って存在していたからである。要するに，誰が家族であってだれがそうでないかは，個人のもっている特徴とは独立に決められているのである。

しかしながら，心理測定において，少なくとも客観的に観測可能な存在は，項目に対する被験者の反応，あるいは，それにもとづいて計算される項目間相関だけである。すなわち，個人の類似，非類似の関係だけにもとづいて，家族を再構成しなければならない事態に我々は直面しているのである。

古典的カテゴリーに関しては，共有される特徴にもとづいて，カテゴリー自体を定義することができる。しかし，家族的類似に関しては，前述のように，異なった特徴を共有することによる項目間のネットワークは，事実上どこまでも繋がり得るから，特徴の範囲を限定するための手掛かりがどこにもないのである。

因子分析の意味

このように考えてくると，単純構造へ向けての回転を含む因子分析に対して，通常付与されている意味も再検討の余地がある。因子分析は，相互に異なった類似度で結ばれている個体（ここでは項目）の集合から，家族を発見していく手法として実用的には，大いに有効であると考えられ，実際，多用されている。しかしながら，家族的類似モデルを認める限り，そこから得られる因子は，必ずしも単一の特徴に対応するとは限らない。すなわち，因子分析を通じて得られた尺度は，既に一定の「幅」，すなわち領域範囲の広がりをもっているに違いない。その際，得られる因子は，不変かつ固定的なものと言うよりは，用いられた項目に含まれる特徴の分布に依存した，流動的，文脈依存的なものと見なさざるを得ない。

また，個々の分析過程で得られた因子に対して，何らかの（理論的存在としての）特性が対応するとしても，その特性自体が幅のある（ぼやけた）ものとならざるを得ない。

近年，探索的因子分析の乱用を戒め，アプリオリな理論を検証するために因子分析を用いること，すなわち，確認的因子分析への移行を促進しようとする議論があるが，そこで検証の対象となっいる理論

は「幅」のない古典的カテゴリーにもとづくものであり，心理測定データに当てはめるには，いささか無理強いの感がある。このことが，1節の(2)の①～③としてあげた，因子分析における問題点を説明してもいる。項目反応に影響する特徴は数多いから，サンプルサイズを大きくすれば，有意な因子数は増えていく（①）。変数を追加すれば，関連する特徴の数もその分布も変化するから，当然，因子の性質は変化する（②）。また，たまたま，ある項目と多くの特徴を共有する別の項目が分析に加えられれば，それまで，独自成分に分類されていた変動が共通因子に移り，因子数は1つ増加することになる（③）。

一方，（単純構造に向けての直交回転を含む）探索的因子分析を行い，各因子ごとに高い（salient）負荷をもつ項目を選択して，複数の尺度を構成するという方法は，こうした限界をわきまえている限り，適切な方法であると言える。ただし，こうした使い方を公然と推奨している因子分析のテキストは，ほとんどないのであるが。

項目の純粋化は可能か

こうした事態の根本的原因は，測定の基本になる個々の項目得点を規定する特徴の数が多いこと（つまり，multiply determinedであること）にある。むしろ，項目内容をもっと純粋化することによって，事態を古典的カテゴリーの想定に近づけることはできないだろうか。多分，それは困難であり，現実的に有用でもない。たとえば，Loevinger（1957）は，

「すべての行為には複数の意味（reference）があることを忘れてはならない。不幸にして日常的行動は，一義的に項目化することができない。行動の研究に自然な単位は存在しない。」

と述べる。

テストの項目を被験者に提示し，それに対する応答を調査するというプロセスを一種のコミュニケーションの過程と見るならば，そうした対人的コミュニケーションの内容には適切なレベルがあるのであろう。人間のカテゴライゼーション（の能力）にもユニヴァーサルな基本レベルが存在しているのと同様である（Lakoff, 1987, 池上他訳, 1993）。純粋化された項目とは，もはや（素人の）被験者に解答（あるいは回答）してもらえるようなレベルの言語では記述できないようなものであろう。

日常の量的概念

こう考えてくると，心理測定は，極めて特殊な測定であるように見える。しかし実は，われわれが日常使用している量的概念の多くが，同様のあいまいさを免れない。

たとえば，心理測定のあいまいさは，それが項目に対する反応という「主観的」なデータにもとづいているせいであると考えられるかもしれない。しかし，必ずしもそうは言えない。たとえば消費者物価指数は，「円」単位で表示された店頭価格という，明確な次元を持つデータに基づいているにもかかわらず，どのような品目が選択されるかによって，異なった数値を得ることになる。また，消費者物価指数が消費者の実感に合わない（つまり，心理測定理論の意味での妥当性を欠く）という議論は常にある。消費者物価指数のような構成概念は，社会科学の中には無数にある。

また，個々の測定値が物理量として得られる場合にも，こうしたあいまいさはつきまとう。たとえば，身長や胸囲には，明確な測定方法の定義がある（保志, 1977）。しかし，「体の大きさ」といったもう少し日常的な概念を考えると，それは身長と完全に一対一で対応するとは限らない。にもかかわらず，「体の大きさ」が，ほぼ身体の物理的に測定可能な側面によって規定されていることは間違いないであろう。こうした概念を用いたコミュニケーションは通常可能であるが，やはり，個人ごとに多少とも違った（暗黙裡の）定義がなされているのであろう。「体の大きさ」は，このようにあいまいな概念である。

もう一つ「足の速さ」というのを考えてみよう。これも，一定の距離をどれだけのラップタイムで走ることができるかで定義できると考えられるかもしれない。しかしながら，距離を何メートルにとるかで，意味は異なったものになる（マラソンランナーが，短距離走も得意であるとは限らない）。さらに，走路その他の条件まで考えると，「足の速さ」もまた，かなりあいまいな概念である。

こうして，われわれの日常的な量的概念の大部分は，対象の複数の側面が，個人ごとに少しずつ違っ

たやり方で合成されたものであると言えよう。「大きさ」,「速さ」のような明確な物理的次元が対応すると考えられる場合にも，考慮の対象となる次元は一意的には定まらず，複数の次元を合成する方法も一定ではない。これらと，「頭のよさ」,「美しさ」等との違いは程度の問題にすぎないとも思える。

このように考えてくると，「長さ」,「質量」,「時間」,「温度」,「電流」といった物理的次元が，いかに特殊な形で純粋化された概念であるかがはっきりしてくる。少なくとも，これらの次元には，複数の次元への分解の余地がない。(もちろん，ある物理的次元が，多くの原因となる変数によって規定されていることはありうる。たとえば，成人の身長が，親の身長や他の幾つかの変数を独立変数とする重回帰方程式によって説明されることは可能である。しかし，これは身長を決定する要因の分析であって，身長の概念そのものが分解されるわけではない。統計モデルとしては同形であっても，区別して扱われる必要がある。)

概念と尺度の分割可能性

こうした議論は，通常の1次元の得点が出るテストの採点方法を無意味であるとするものではない。どのようなものであるにせよ，それに対応する何らかの（もちろん家族的類似にもとづいた「幅」のある）構成概念が存在するであろう。その構成概念の名称が，常識的な日常言語の意味とだいたい一致しているかどうかは別であり，そもそも，それに対して短い言語表現を与えることができるかどうかも疑問である。それでもともかく，尺度に対して実用的な，あるいは，一定の理論的な意味をもたせることは可能であろう。しかしながら，そうした構成概念を，それ以上分割不可能な実体と考えることはできない。

尺度の側から見れば（用いる項目を固定して考えたとしても）同様のことが言える。どんな尺度も(十分大きな標本を用いれば)，複数の意味のある尺度に分割することができるはずである。さらに，そのテストの領域内に含まれると想定される項目も加えれば，事実上無限の下位尺度が存在しうるであろう。

この種の「次元内構造」は，長さや質量のような（純粋の）物理量においては考えることができない。こうして，心理学的な構成概念の物理量とは決定的に異なる特徴が明らかになったと思われる。そして，こうした量的概念への反省は，当然，テストの妥当性の概念にも影響を与えることになる。

4．尺度の妥当性検討への含意

妥当性の意味

妥当性（validity）とは，テストが測定しようとしているものを，本当に測定している度合いであり，究極的には，経験的に得られる尺度と，理論的に想定される特性，あるいは構成概念が一致している度合いのことである。尺度が，(ランダム誤差以外の）組織的変動としては，完全に単一の特徴だけを反映し，構成概念もまた完全な一次元的性質をもつならば，妥当性は尺度と構成概念との間の相関係数によって評価することができる。もちろん，構成概念上での個体の「測定値」それ自体を得ることはできないから，これはあくまでも仮想的な事態にとどまる。しかしながら，多少の誤解を恐れずに言えば，基準関連妥当性，構成概念妥当性（特に，多特性－多方法行列）といった，さまざまな妥当性検討(validation）の概念や方法は，近似的（あるいは概念的）に，その相関の推定値を得るために存在してきたと言えよう。

しかしながら，尺度も構成概念も，いわば複数の「特徴の束」といった様相を呈することを認めざるを得ないとすれば，両者の関係も俄然複雑なものとなってくる。再度，図2のような区分について考えてみよう。尺度に関して，注意すべきことは，以下のようにまとめることができる。

㋐ 個々の項目は，それぞれの「成分」を異なった度合いで反映しているが，それを具体的数値として推定することは不可能である。

② もっと基本的な問題として，関連する下位概念の数 n_c と別の特性の数 n_c を特定することは不可

能である。

③ 構成概念と恒常誤差の境界を決定することは不可能である。ある理論的前提の下では構成概念に含まれるものが，別の前提の下では恒常誤差とみなされうる。場合によっては，状態やランダム誤差を区別することすらできない。

これらの点を念頭に置きながら，具体的な分析方法について考えていこう。

内的整合性にもとづく項目分析の「効用と限界」

前述のように，項目－尺度得点相関は，項目の妥当性の指標ではないし，尺度の領域範囲を無視している。そこで，1次元性を高めるために，項目－テスト相関とα係数にもとづいて項目選択をしていく，項目分析の手続きは無条件で適切なものとは考えられない。しかしながら，領域範囲云々以前の問題として，個々の項目には，何を測定しているにしてもランダム誤差が大きいものもあり得る。能力測定において，問題の意図が受験者に誤解されやすいとか，複数の選択肢が正答となりうる場合，質問文が二様の意味にとれること等である。あるいは，そうした明らかな欠陥は含まないにもかかわらず，意図した構成概念とは別の特徴だけを反映している項目もあり得る。

これらの2種類の「具合の悪い」項目は，データの上では，他の項目との相関が低い，あるいは，他の関連する次元の尺度との相関のパターンが他の項目と異なる，といった特徴があるから，そうした項目を排除する手段としては，項目－テスト相関とα係数にもとづく項目分析は有効であろう。

しかし，この二点に関して，何らかの形で項目の棄却，あるいは改善が行われ得るとしても，なお決定的な問題が残される。すなわち，一つの尺度を構成する項目全体が複数の次元からなることから，たとえば，いかなる尺度にも，次のような可能性が生ずることが避けられなくなる。

(1) 尺度を構成する項目の範囲が想定された構成概念に対して狭すぎる可能性。

(2) (1)とは逆に，尺度が，構成概念の中に本来含まれるべきでない範囲の項目群を含む可能性。

(3) 考慮されている範囲は適当なものであるとしても，項目の「分布」に偏りがある可能性。

具体的な尺度において，どのような下位概念がクローズアップされることになるかは，項目の選択に依存することになる。

他の変数との相関の解釈

個々の尺度に，前述のような内部構造が存在することは明らかであるが，その解釈を尺度の内部だけで行うことはできない。そこで，他の変数との相関や複数の被験者群間の差等を通じて，尺度の解釈を明確にしていくのが，構成概念妥当性検討の手続きである。しかしながら，そこにも新たな問題が発生する。

たとえば，2つの尺度間の相関係数の解釈にあたって，通常，そこに生ずる希薄化（attenuation）の効果を考慮に入れる必要がある。すなわち，二つの尺度XとYのそれぞれの信頼性係数をρ_X，ρ_Yとし，それぞれの真の得点の間の相関係数をρ_{XY}とするとき，XとYの間に実際に観測される相関係数r_{XY}は，

$$r_{XY} = \sqrt{\rho_X \rho_Y}\,\rho_{XY} \qquad (4-1)$$

となるということである。すなわち，真の得点を構成概念上での「正確な」測定値であるとするならば，そうした測定値間の相関係数は，みかけの値より大きいはずだということである。

しかしながら，測定値の性質を，相対的な関係のネットワークにもとづく任意的なものと考えざるを得なくなってみると，信頼性係数の値が何を意味するかははっきりしなくなってしまう。それは，単なるランダム誤差の小ささとはかぎらず，むしろ，項目が抽出されたネットワークの範囲の狭さの指標とも考えられるからである。同様の理由で，α係数はρ_X，ρ_Yの推定値としては過小となるため，それらを用いて，（4－1）によって希薄化の修正を行うと，ρ_{XY}は1を越えることがある。

さらにより深刻な問題として，尺度間の相関にも，二つの尺度の定義に含まれる項目のネットワーク

の間の重なりという問題が生じてくる。つまり，次元間の関係を問題にする場合，二つの尺度を定義する項目得点を規定する特徴には，初めからある程度の重なりがあるから，尺度間の相関は，実は単に二つの概念の定義の類似度を反映しているにすぎないのではないか，ということである。こうした問題は，より進んだモデルである共分散構造分析等を用いて，構成概念間の因果関係を検討するような場合にも回避することのできない問題である。

特徴自体について語ることの困難

われわれが個々の特徴自体を観測することができ，それらの間の相関的，あるいは因果的関係を明らかにすることができれば，個々の特徴の意味が明らかになると同時に，特徴間の法則的関係に関する命題も検証できる。それが，本来の構成概念妥当性検討が目指していたものである。しかしながら，われわれは，すでにそれ自身が「特徴の束」である項目反応，あるいはそれを合成したものとしての尺度得点を観測することができるだけである。関与する特徴をある程度具体的に想定していたとしても，必ずしも意図どおりの項目が作れるわけではない。

妥当性の意味再考

心理学的個人差を規定する基本的「要素」としての特徴については，それらの値を推定することが不可能であるばかりでなく，その数すらわからないとすれば，心理測定の妥当性について語ることは不可能であり，心理測定という営み自体が無意味なのではなかろうか。

しかし，こうした状況を認識した上で，あらためて心理測定という行為を，「ある文化的状況の中において，基本的かつ自然なレベルにおけるコミュニケーションを通じて，同じ文化的状況の中に成立している概念に量的表現を与え，それらの意味と相互関係を明らかにすること」といったように，再定義するなら，そこに少なくとも幾分かの光明を見い出しうるように思われる。

心理学的個人差に関する理論は，通常自然言語によって記述され，自然言語の中に埋め込まれている。多くの場合，内容的にもその文化圏の中で一般人のもつ常識からさほど離れたものではない。しかし，そうした理論の枠組みにもとづいて見直すと，それまで一見無意味に見えた因子分析の結果が解釈可能になり，理論のネットワークの中に位置づけられるようになる場合がある。この種の場合，因子分析の見出した次元（特性）は，そうした過程を経て既成の理論の中に組み込まれる。

逆に，因子分析の結果が，それまで持っていた理論的枠組みに大きな影響を与え，理論の修正をせまる場合もある。通常，理論というものは，データとは独立に，論理的にアプリオリに与えられ，それが経験的にデータによって検証（あるいは反証）されるという一方向的な過程が強調されることが多い。しかし，理論とデータは，どちらが先行するのでもなく，相互に影響を与え合い変化していく。そうした相互的な影響関係を通じて，尺度の「幅」も次第に適切な範囲に拡大あるいは限定され，項目の「分布」もその概念の中心的部分をとらえられるように変えられていく可能性がある。

むしろ，こうした過程を通じて，初めて一つの次元が認識されると考える方が，意義があるように思われる。そうした相互作用の上でなら，その次元を，（とりあえず）想定した構成概念の名称で呼ぶことは，さほど問題ではあるまい。人文・社会科学の場合，この種の科学的研究の結果としての言語の意味の変化は，比較的日常言語にもフィードバックされやすい性質を持っているから，研究の出発点において構成概念を支えていた文化的文脈自体を，研究の結果が変化させていく可能性もある。日常言語の文脈そのものは，そうした影響抜きでも絶えず変化しつづけるから，その結果として，理論の方も変化しなければならないことの方が多いであろうが。

結局，妥当性とは，そうした文脈を共有する個人間で話が通じるように，辻褄のあった形でその概念が定義されることである。

心理測定と言語とのアナロジー

こうした過程は，結局のところ，F．ソシュールの言語学に示されているような，人間の認識の過程そのものに近いと思われる。村上（1989）にも引用したが，丸山（1981）によれば，

「我々の生活世界は，コトバを知る以前からきちんと区分され，分類されているのではない。それぞれの言語のもつ単語が，既成の概念や事物の名づけをするのではなく，その正反対に，コトバがあってはじめて概念が生まれるのである。」(pp. 117-118)。

「ソシュール以前は，コトバは表現でしかなく，すでに言語以前からカテゴリー化されている事物や，言語以前から存在する純粋観念を指し示す道具と考えられていたが，ソシュール以後の考え方では，コトバは《表現》であると同時に《意味》であり，これが逆に，それ自体は混沌たるカオスの如き連続体に反映して現実を非連続化し，概念化するということになる。」(p. 120)。

これらを読んで，心理測定は，個人差を認識するための言語と似た位置にあると感じられないだろうか。いわば，「混沌たるカオスのごとき連続体」である特徴の集合を，理論的探究と経験的手続きを通じて，概念化することに他ならないのである。

心理測定の方法への家族的類似モデルの意味

最後に，ここまでの議論が，具体的な心理測定の方法に対してもたらす意味をまとめておこう。

(1) 心理測定尺度を定義する項目の選択にあたっては，高度の任意性が存在する。

(2) 尺度は，たまたま選択された項目群の中で優位な「下位概念」を中心とした次元を「測定」するものとなる。

(3) 因子分析の因子数を決定することはできない。正しい因子数という概念は存在しない。

(4) 構成概念を，測定方法とは独立した客観的実在と見ることはできない。尺度も対応する構成概念も，常に暫定的な性格をもつ。

(5) 構成概念間の相関関係を解釈するにあたって，概念間の法則的関係と，それらの概念を測定するテストの項目内容の重なりを区別することは困難である。

5．終わりに

本稿の議論は，従来の方法に対する批判ではあっても，何ら建設的な視点に立っていないという見方があるかもしれない。しかしながら，一連の問題項目が単一の個人差次元にだけに関わっていると想定し，1次元性からのズレを修正しようと試みるよりも，個々の項目は，相互に共通の成分はもちつつも，それぞれに異なった次元に対応していると考える方が自然ではないだろうか。項目をどのように選択するかによって，その中のある成分が強調され，別の成分は埋没し，時には相殺されることによって，尺度の意味も流動しうると考える方が，テストを解釈し利用する上でも，より実際的ではないだろうか。すなわち，データの背景に唯一の（あるいは一定数の）不変の次元の存在を仮定することよりも，さまざまな文脈とともに可変的な次元が設定できると考えた方が有効ではないだろうか。幾つかの問題項目は，教育制度や社会・文化的状況の変化とともに使用不能になるかもしれない。「物理的」に同じ項目が，異なった意味を持つようになることもあり得るであろう。

それは，必ずしもデータとは無関係に恣意的に任意の次元を設定できるということを意味しない。また，理論的に存在を予測することのできない次元に対しては，それを測定する手段（項目）を思いつくことはありえない。また，理論的な想定は可能であっても，測定のための項目が見つからないこともありうる。

悲観的に言えば，概念間の関係について新しい発見といったものがありうるとしても，それは単に見掛け（言語表現）上，著しく異なっているように見える二つの項目間に，高い相関があることの発見であるにすぎないと見られるかもしれない。しかし，ある理論的基盤を共有する研究者間では，ある研究が，特徴の重なりにだけ依拠した単なる同語反復的なものであるか，新たな関係を認識するものであるかは，意見が一致する場合が多いであろう。ある研究の結果が「面白い」ものであるかどうかは，理論的文脈に依存する。

また一方，個々の項目が多数の特徴を反映しているということは，テストにとってある種の安全装置

でもありうる。特に能力のテストは決して測定者の意図した通りのものを測るものとはなり得ない一方，全くの見当違いなものであることもあまりないであろう。いかなる項目も，何らかの relevant な特徴を含むであろうからである。

ここで示したような見方は，理論検証の手段として心理尺度を利用する際により柔軟な態度を可能にするであろう。一つの尺度に固定的に一つの構成概念を対応させて，硬直した議論を展開するよりも，他の尺度との関連で尺度が反映していると想定される構成概念を考え直す余地を残す方が，結局は現象をよりよく理解することにつながる。先験的な思考にもとづく理論がデータによって検証，あるいは反証されるという一方的な方向ではなく，データを通じて理論はその構成要素から全面的に問いなおされる。だが，そうした過程を経て生まれる理論も法則も常に暫定的なものにとどまる。

しかし，そうした柔軟性こそ，人類が人間や社会を自然言語を通して認識してきた長い歴史を受け継ぐ正当な方法であると，少なくとも著者は考えているのである。

文　　献

Bloor, D. 1983 *Wittgenstein : A Social Theory of Knowledge.* Macmillan Press. (戸田由和訳 1988 ウィトゲンシュタイン：知識の社会理論　勁草書房)

Comrey, A. L. (1973) *A Course in Factor Analysis.* Academic Press. (芝祐順訳 1979 因子分析入門　サイエンス社)

Cronbach, L. J. 1951 Coefficient alpha and the internal structure of tests. *Psychometrika,* **16**, 297-334.

Henning, G. 1987 *A Guide to Language Testing: Development, Evaluation, Research.* Newbury House.

保志浩 1977 人種特徴　人類学講座編纂委員会（編）人類学講座　7　人種　雄山閣出版　27-45.

Lakoff, G. 1987 *Women, Fire, and Dangerous Things : What Categories Reveal about the Mind.* University of Chicago Press. (池上嘉彦・河上誓作他訳　1993　認知意味論　言語から見た人間の心　紀伊国屋書店)

Lazarsfeld, P. F. 1972 *Qualitative Analysis: Historical and Critical Essays.* Allyn and Bacon. (西田春彦・高沢健次・奥川櫻豊彦訳 1984 質的分析法　岩波書店)

Loevinger, J. (1957) Objective tests as instruments of psychological theory. *Psychological Reports,* **3**, 635-694.

丸山圭三郎 1981 ソシュールの思想　岩波書店

McDonald, R. P. 1985 *Factor Analysis and Related Methods.* Lawrence Erlbaum Associates.

Messick, S. 1975 The standard problem: Meaning and values in measurement and education. *American Psychologist,* **30**, 955-966.

村上隆 1989 心理測定の理論と家族的類似の概念　名古屋大学教育学部紀要—教育心理学科，**36**, 101-108.

Rosch, E. 1987 Wittgenstein and categorization research in cognitive psychology. Chapman, M & R. A. Dixon (Eds.) *Meaning and the Growth of Understanding.* Springer-Verlag, 151-166.

芝祐順 (1975) 行動科学のための相関分析法　第2版　東京大学出版会