日本語教育学・研究方法論
——量的研究の可能性と課題(3)：調査法の信頼性の向上の枠組み——

岡崎敏雄

1．はじめに

質問紙調査法は，一般にアンケートと呼ばれて安直な捉え方がなされることが多い。しかし，同法は，厳密な実施を旨とするテストの形態として開発され，さらに，人間諸科学に適用される中で生態的妥当性を獲得し改善されてきた本来精度の高い研究方法である。ところが日本語教育の他いくつかの分野においては手続きの諸段階でズサンであったり，また多方面・多次元的な情報を含んでいるデータからごく限られた範囲の情報しか得られないという効率の悪い実施が見られる現状にある。

このズサンさ・非効率性は二つの点に集約される。

第一は，質問紙の質問項目作成を含めた執行の各過程が適切かどうかに関わる厳密な吟味がほとんどなされないまま実施されているという点である。これは質問紙調査法の信頼性の検討がなされていないことに由来するものである。

第二は，質問紙調査を行ってもその結果の処理において，せいぜいパーセントの違いを比較して論ずる範囲での情報入手に終わっていることが多い点である。このため調査対象の事象の性格を多面的にまた多次元レベルでとらえるに至っていない。具体的には事象の多面性・多次元性の把握のための枠組みに，質問紙調査結果を組み込み得ていないことからくるものである。

ごく最近の日本語教育学研究には日本語教育諸現象のダイナミズムを捉え，得られたデータの本来持っている多面的・多次元的な情報を引き出し，信頼性，妥当性の高い研究が出始めている。このような研究を散発的なものとせず，日本語教育学の研究全体がそのような性格を持っていくためのシステマティックな追求が求められている。

以上の視点から本稿は，上記の問題点のうち質問紙調査執行の各過程に関わる信頼性の検討に焦点を当てる。

具体的には以下について取り上げる。

(1) 質問紙調査法のズサンな実施の改善が進まぬ構造的根拠
—測定手段の問題を検討対象として明示的に持たぬ従来の信頼性の把握—

(2) 信頼性概念そのものの見直し
—従来の信頼性の定義の問題点—

(3) 信頼性と測定誤差
——一般化可能性理論 generalizability theory に基づく信頼性の把握—

(4) 測定手段の誤差の重要性
——一般モデル general model の特色—

(5) 測定手段としての質問紙調査法の信頼性引き上げの枠組み
—質問紙調査法の測定誤差の引き下げの枠組み—

1．質問紙調査法のズサンな実施の改善が進まない構造的根拠

——測定手段題点を検討対象として捉える根拠を明示的に持たない従来の信頼性の把握——

質問紙 questionnaire は元来テストの形式の一つとして開発されたものである。それが言語教育，教育学一般，心理学，社会学などの社会科学や人間科学において適用されてきているものである。それがズサンな使われ方をされ，質問項目の作成を含めて執行の各過程が適切かどうかの厳密な吟味がほとんどなされないまま実施され，改善が積極的に追求されないケースが多いのは先に述べたように質問紙調査法の信頼性の検討が追求されていないことに由来する。

質問紙調査法がこのように信頼性の検討の対象として積極的に取り上げられないのは，質問紙がそもそも誕生したテストの分野における従来の信頼性の把握にその構造的根拠を持つ。一言で言えば，従来のテスト理論において，テスト，つまり（質問紙を含む）測定手段，の問題点を，信頼性を左右する要因として捉え，それらを検討対象として捉える根拠をモデル上明示的に持っていないことに基づくものである。

2．信頼性概念そのものの見直し

——従来の信頼性の定義の問題点——

信頼性を定義する際，一般に従来なされてきたのは例えば次のような内容である。(東他1988,p．343)。

> 同一の集団に対して，同様な条件の下でテスト実施を繰り返す時，一貫したテスト得点が得られる程度を，それらのテスト得点の信頼性，または単にテストの信頼性という。

この限りで大きな問題はない。ところが，ここで「同様な条件の下でテスト実施を繰り返す」という場合，具体的には，「同一のテストを同一の条件の下で繰り返すことから，内容的に類似している別のテストを実施すること」として解釈されてきた。また，テスト実施を繰り返すとは，再テスト法，平行テスト法，折半法，アルファ係数を使って内的整合性を見ることであり，そこでテスト得点を左右する要因として想定されていたのは，「不注意や当て推量における運・不運」，「その日の健康状態や気分，動機付け」，「テストとテストのあいだの学習による変化」，「個々の項目に対する得意・不得意」などである。(以上引用は何れも東他1988,p.343－344)。

ここで注目すべき点は，「不注意」，「健康状態」，「学習」など受験者側について述べられているのに対して，テストそのものの側で直接信頼性に関わる部分が明確な形で述べられていない点である。同様のことは，信頼性を数値で表す指標として用いられてきた信頼性係数の基本的概念が次のようなものであることにも見られる（東他，同上p．344)。

信頼性係数＝集団における真の得点の分散／集団における得点の分散

$$\rho = \sigma_T^2 / \sigma_X^2$$

(ρ：信頼性係数；σ_T^2：真の得点分散；σ_X^2 テストの得点の分散)

信頼性係数＝実際の得点と真の得点の間の相関の二乗

$$\rho = \sigma_{XT}^2$$

(σ_{XT}：実際の得点と真の得点の相関)

何れにおいてもテストそのものの側で持つ要因は係数のどの部分をも構成し

ていない。

ただしKuder・Richardsonの第20の公式によって表された信頼性係数（後にCronbachのαと呼ばれるようになった）はテストの項目つまりテストの側の要因を取り上げている。

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma i^2}{\sum_i \sum_j \sigma i \sigma j \rho ij}\right)$$

（n：テスト問題項目数，$\sigma i, \sigma j$：項目i, jの標準偏差，ρij：項目iとjの間の相関係数）

この式によってはじめてテストそのものの側で直接信頼性係数に関わるものとしてテスト項目が明確な形で取り上げれた。内的整合性に基づく信頼性係数とよばれるこの式は，項目間相関の高さを示すものであり，これに基づいて，テストのいわゆる項目分析がなされている。（同時に外在基準とテスト得点の相関として定義される基準関連妥当性の概念に基づいて出される妥当性係数を高めるような外在的基準と高い相関を持つ項目を明らかにする項目分析の行われている。）しかし，この項目間相関を高めることを理論的に根拠としたテストそのものの信頼性の引き上げの把握は次の点でテストそのものの信頼性引き上げ指針としては未だに十分な基盤に立ったものとは言えない。

第一に，項目間相関が高くなることつまり内的整合性が大きくなることはテスト項目同士が似たものばかりが寄せ集められることを意味する。これは測ろうとする能力の真の値TRUE VALUE（後述）に近づくことを妨げる結果を導くことになる。

言い換えればこれまでの信頼性の捉え方をつきつめていくと，信頼性を高めようとして項目間相関を高めれば高めるほど，妥当性を低めてしまうことになる。一般に良質のテストを作る際に信頼性と妥当性の確保が不可欠だとされるが，実際の場面で信頼性を高めていくことはこのような捉え方の下では妥当性との間に矛盾を来すものであり信頼性の基準としては未だ不十分である。この点については，テストの開発者からの指摘はなされていたものの（cf.「信頼性と妥当性のディレンマ」Bachman 1988など）信頼性と妥当性を同一次元の問題として捉える理論的枠組みが存在しなかったこともあり，従来テストの根幹に関わるものとしては取り上げられてこなかった。APA（American Psycho-

logical Association), AERA (American Educational Psychological Association), NCME(National Council on Measurement in Education)が後述する一般化可能性理論の強い影響の下にこの点に注目し始めたのも近年のこと（1985）である。

第二に，項目間相関が事実上テストの信頼性を引き上げる論拠である場合，何がその相関を高くするか，テストそのもののどこをどうすることがそれを可能とするかの指針は明確には示されない。極言すれば，項目間相関が低かった場合それは項目そのものに問題があった故か否かの論拠も存在していないと言ってよい。何らかの形でテストそのものが問題発生の根拠になり得ることを構造的に説明し得るモデルがない限り，テスト自体が信頼性を引き下げる可能性のあるものとして位置づけその上で具体的に何をどう改善するかの指針を得ることはできないのである。

3．信頼性と測定誤差

――一般化可能性理論 generalizability theory に基づく信頼性の把握――

a．測定誤差の捉え方と信頼性

上に述べたような信頼性の捉え方は基本的に次のような真値測定モデル true score measurement model に基づく測定誤差の捉え方に基づく。一言で言えば，誤差を全てランダムエラーとして捉え，どのような要因によって構成されているかに関するシステマティックな分析を構造上組み込んでいないモデルであることに起因しているのである。先に述べたテスト得点に影響を与える要因（例「不注意や当て推量における運・不運」,「その日の健康状態や気分」,「動機付けなど」・・・）も，テストの項目間の非整合性も，ひっくるめてランダムエラーとして捉えられている。学習者に起因する誤差と測定誤差と測定手段であるテストに起因する誤差を構造上区別し，厳密にそれぞれを分析することが位置づけられていないのである。

言い換えれば従来の信頼性の把握においては，測定誤差の構造が未だ十分に把握されていなかったといえる。これに対して，一般化可能性理論は誤差の構造的把握こそをモデルの中心部分とする。以下で誤差の把握についての二つのモデル－従来のテストモデルである真値測定モデルと，本稿で取り上げる一般化可能性理論の基礎とする一般モデル－のそれぞれについて，測定誤差がどのように捉えられているかを見，その把握に立って信頼性の新たな定義を提出す

る。

b．測定誤差の把握に関する二つのモデル

——真値測定モデル true value measurement model と一般モデル general model ——

真値測定モデルの下では次の公式が成立する。

観察された得点 observed score＝真値 true score＋測定誤差 error

このモデルにおける誤差の捉え方の特徴は次の二点に集約できる。

1．誤差は全てランダムエラーとして捉える。

2．従って誤差の要因 error source としてどのような要素が有り得るかを検討する基盤がない。

これに対して一般モデルでは，次のような公式が成立する。

観察された得点＝言語能力＋測定手段の要因 test method factors＋認知上の諸要因 cognitive factors＋予測不能で非系統的な要因による要因 random factors＝系統的誤差 systematic error＋random error

このモデルは一般化可能性理論 generalizability theory （いわゆる G-theory）（Cronbach1984）に基づくもので，次のような特色を持つ。

1．誤差の要因として測定手段の要因 test method factor 及び認知上の要因 cognitive factor を明示し，ランダムエラーから区別している。

2．測定手段の要因や認知上の要因の交互作用による効果をも考察の対象とする。

c．一般モデルにおける測定誤差の把握

一般化可能性理論に基づく一般モデルの下では，測定誤差は測定者の関与することのできない単なる偶然によるエラーとしてではなく，規則性を持つものであり厳密な分析の対象として捉えられる。

系統的誤差は一回のテストの実施と次の回の実施の間で一定して規則的に発生するエラーである。例えば系統的誤差を構成する二つの要因のうち，測定手段の要因を取り上げて考えた場合，同じ受験者の能力を測る場合でもインタビューテストで得られる得点の傾向と多肢選択の形式で得られる得点の傾向の間には規則的な違いがある。本来同一の能力を測っているとすれば同一の得点

の傾向がでてよいはずであるが，これが異なっているというのは測定手段つまりテスト形式によってバイアスが持ち込まれているのだとする。このバイアスが系統的誤差と呼ばれるものである。

認知上の要因の場合には，例えばある個人の field independence などの認知上の要因は一つのテストの中で規則的にテストのパーフォーマンスに影響を与えると考えられる。これもまた系統的誤差の一つである。

また，系統的誤差には二つの違った効果が想定される。一つは同一の受験者全てに渡って影響を与える系統的誤差の効果で一般的影響 general effect と呼ばれるものであり，他方はテスト受験者のうち特定の個人あるいはグループのみに見られる特殊効果 specific effect と呼ばれるものである。例えば読みのテストとして経済学に関するテキストを取り上げてテストをした場合，テストの内容である経済学は読みの能力全体のうち経済学というトピックに関する読みの能力を測っているという点では全ての受験者に共通のいわゆる一般的影響を与えている。これに対して，ある受験者は経済学に不慣れであり他のものは習熟しているといった違いがある。この場合不慣れな受験者の得点を下げ，習熟した受験者の得点を上げるという特殊効果を持つ。

ここで注意すべき点は特殊効果はランダムエラーとは規則性を持つという点で異なるということである。予測不能で非系統的偶然の要因によるランダムエラーは，「予測不可能であり規則性を持たず一時的な諸条件によってもたらされるエラー」であり，例えばいわゆる頭がさえているか，感情的にどういう状態にあるか，テストの関係に受験を妨害したり何らかの影響を与えたりするものがないかどうかなど一回のテストから次回のテストまで不規則に変わり得る要素によって引き起こされるものとして捉えられる。

4．測定手段の誤差の重要視

—— 一般モデルの特色 ——

a．一般化可能性理論の下での信頼性の定義

一般化可能性理論に基づく一般モデルの特色の中心は測定手段の効果を誤差の要因として最大限重要視して取り上げるという点である。これは一般化可能性理論が次のような性格のものに由来することによる。一般化可能性理論は，テストの得点の多様な「分散をもたらす要因」source of vairiance の相対的な効果を特定する事を目指すものである。その場合テストの得点をいわゆる可

能尺度のユニバース universe of possible measures における標本として捉え，その得点を尺度のユニバースに照らすことによって一般化して考える。言い換えれば，ある受験者があるテストで示すテスト上のパーフォーマンスをその受験者が他の様々の条件下で見せるパーフォーマンスの一つとして一般化して考える。その場合，そこで取り上げられる受験者のあるテストでのパーフォーマンスが「一般化の可能性の高いもの」であればあるほどそれは信頼性の高いものと考える。

この意味で，一般化可能性理論の下では信頼性を「一般化できる可能性の程度」と定義する。

b．尺度のユニバース universe of measures における測定手段の重視

このような捉え方の基本をなすのが尺度のユニバースでの概念である (Bachman 1988)。一般化可能性理論において測定手段がいかに重視されているかはこの尺度のユニバースの捉え方に明確に示されている。

いま与えられた測定対象について一つのテストを考えるとする。その場合どんな性格のテストの得点を測定対象能力の指標として受け入れるべきかが問題となる。例えば，聴解能力のテストを考えてみよう。

その場合テストの得点として択一式テストによる得点，デイクテーションによる得点，口頭インタビューテストによる得点何れもが聴解能力の得点として利用し得る。また択一式及びデイクテーションのテストに取り上げるテキストとして特定の専門領域に偏らず内容選択を考えようとする場合，例えば経済に関する領域のテスト，歴史に関するもの，あるいは物理学に関するものなど幾つかの話題のものを選ぶことができる。さらに試験の実施に当っても，聴解テストであるから，その選択肢やデイクテーションのテキストをどのようなスピードでそのような読み方で読むかで得点に影響が与えられる可能性を考えれば，試験官が誰であるか，及び口頭インタビューテストであれば面接官が誰であるかも考慮の対象とすることが必要になるとなる。

これらの対象は言い換えればテストの条件を特定していることになるが，これらの異なった諸条件を特定することを一般化可能性理論では，「対象となっているあるテスト固有のコンテクストに関連のある尺度のユニバースを規定すること」と捉える。

上の例で言うと，択一式，デイクテーションテスト，口頭インタビューテストのうちどの形式で行うか，あるいはテキストの内容として経済，歴史，物理のどれをとるか，試験官としてだれを採用するか，面接官として誰を採用する

かを問う場合,「テストの形式，テストの内容，試験官，面接官」の項目はfacet，また「択一式，デイクテーションテスト，口頭インタビューテスト」はそれぞれテスト形式というfacetを構成する条件となる。そのうえでfacet総体は与えられた一つのユニバースを構成するのである。

このようなユニバースを前提としてテストを開発する場合，テスト形式のうち択一式かデイクテーションテストか口頭インタビューテストの中から実際には一つを選び出すことが必要となる。その場合三つの異なったテスト形式から得られる得点の相互比較能力を調べる。これは一般化可能化の視点から考えればそれぞれのテスト形式がどの程度一般化可能なのかその程度を調べることである。同様に，試験官や面接官についても，試験官や面接官によってどの程度テストの得点が影響されるかを見ることになる。また三種類のテキスト内容の一つを取り上げて行われた得られる受験者の得点がどの程度他の二つのテキスト内容を取り上げた場合に得られる得点内容にも一般化できるかの程度を認知する必要がある。

このような手続きにそって，例えば聴解テストの場合可能尺度のユニバースを構成する諸条件の下である受験者について幾つかの測定値が得られるが，これらの幾つかの測定値の平均値がその個人の聴解能力の最も適切な指標であると考えることができる。先に真値測定モデルについて受験者の示す「観察された得点」のうち測定誤差の中のランダムエラーを除いた測定対象能力そのものを表す部分を真値true　scoreと呼ぶとしたが，一般化可能性理論に基づく一般モデルgeneral　modelにおける真値は上記の測定値諸値の平均値として表される。(これはユニバース値universe scoreと名付けられる)。

この場合前提となるのは，可能尺度としてどのような内容をその構成物とするかの規定の仕方によってユニバース得点は変わってくるということである。言い換えれば可能尺度のユニバースの中に択一式とデイクテーションの組み合わせのみを取り上げた場合のユニバース得点と，口頭インタビューテストも加えた場合のユニバースの下でのユニバース得点は異なるということである。つまりある得点の一般化可能性はそれがどのような測定手段，より正確には測定手段よって構成されるどのようなユニバースの下で得られたものかによって異なり，画一的に述べることができないという点である。要するに，測定手段，測定形式，試験官などの諸条件を超えて存在する真値などというものは一般化可能性理論の下では真値測定モデルの場合と異なって存在しないということである。

c. 信頼性係数に表現された測定手段の重視

－一般化可能性係数における測定手段の分散－

一般化可能性理論において測定手段をいかに重要視するかについてはまた，この下での信頼性係数の定義式に端的に表現されている。真値測定モデルにおいて信頼性係数 reliability coefficient が信頼性を示す指標として用いられたのに対して，一般化可能性理論においては一般化可能性係数 generalizability coefficient が用いられる。その場合，この係数はある得点のユニバースでのみ与えられる次のようなものとなる。

$$\gamma_{xp} = \frac{S_p^2}{S_x^2}$$

S_p^2：ユニバース得点分散；S_x^2：観察得点分散（ユニバース得点分散と誤差分散を加えたもの）

この場合，測定手段に由来する誤差分散及び読み上げられるテキスト内容に由来する誤差分散は次のように定義される。

測定手段による誤差分散：$S_{\text{テスト形式}}^2 / S_x^2$

テキスト内容に由来する誤差分散：$S_{\text{テスト内容}}^2 / S_x^2$

以上のように一般化可能性理論に基づく一般モデルでは，真値測定モデルによっては明らかにされなかった多様な測定誤差を識別し，そればかりでなく上記の尺度のユニバースの枠組みに基づいてなされる測定誤差の交互作用を調べるという真値測定モデルでは複雑過ぎて捉えることのできなかった諸相が明らかにされる。このように一般化可能性理論は測定誤差の要因に当るものを特定し，特に測定手段に関わる誤差の諸要因をシステマティックに引き下げる為の枠組みを提出するものである。

5．測定手段としての質問紙調査法の信頼性引き上げの枠組み

――測定手段，質問紙調査法の測定誤差の引き下げの枠組み――

本稿では，上のような一般化可能性理論に基づき，測定手段としての質問紙の信頼性を捉えるにあたり次の点に注目する。

第一に，誤差を信頼性の引き下げの最も重要な部分として捉える。

第二に，誤差をランダムな誤差一般とした真値測定モデルに代わり，測定手段を誤差の重要な要因として位置づける一般化モデルの視点に立つ。

第三に，質問紙の信頼性の引き上げを，測定手段としての質問紙から誤差を引き下げ除去することと捉えその実現を目指す。

具体的には，質問紙の作成から執行に至る過程のそれぞれの段階における誤差発生の可能性をおさえ，その除去をシステマティックに目指す形をとる。

以上の考察を踏まえ質問紙調査法の測定誤差引き下げの枠組みは次のようなものである。(詳しくは次稿)。

A．仮説設定段階における誤差の引き下げ：剰余変数の統制による誤差の引き下げ
　1．恒常化，無作為化，ブロック化
　2．実験計画法
B．調査対象決定段階における誤差の引き下げ：標本誤差の引き下げ
　層別抽出法　stratification
C．質問紙調査表作成段階における誤差の引き下げ：非標本誤差の引き下げ
　1．質問の内容に関わる誤差の引き下げ
　2．質問の量に関わる誤差の引き下げ
　3．解答形式に関わる誤差の引き下げ
　4．質問項目の配列に関わる誤差の引き下げ
　5．ワーディングに関わる誤差の引き下げ

6．結　語

一般にアンケートと呼ばれ安直な捉え方のなされることの多い質問紙調査法を取り上げ，本来厳密な実施を旨とするテストの形態として開発され人間諸科学に適用される中で生態的妥当性を獲得してきた精度の高い研究方法である同法について，質問紙調査法執行の各過程に見られるズサンさ・非効率さを除去

するための枠組みを提出することを目指し，以下について取り上げた。

1．質問紙調査法のズサンな改善が進まぬ構造的根拠

2．信頼性概念そのものの見直し

3．信頼性と測定誤差

4．測定手段の誤差の重要視

5．測定手段としての質問紙調査法の信頼性の引き上げの枠組み

次稿以降で枠組みの詳細及び質問紙調査法実施をめぐる非効率性の改善について取り上げる。

参考文献

Bachman, L. 1988. Fundamental Considerations on Language Testing. Reading Ma : Adddison-Wesley.

Cronbach, L. J. 1951. Coefficient Alpha and the internal structure of tests. Psychometorika, 16,

———1984. Essentials of psychological testing. New York : Harper and Row

American Educational Research Association, American Psychological Association, National Council on measurement in education 1985. Standards for educational and psyhological testing. Washington : America psychological association, Inc.

Messick, S. 1975. The standard problem : meaning and values in measurement and education. American psychologist, 30.

Novick, N. R. and Luwis, C. 1967. Coefficient alpha and the reliability of composite measurements. Psychometrika, 32.

岡崎敏雄 1989。日本語能力試験のシラバス作成に伴う「変化」の体系的把握と「移行」の体系的実施。国際交流基金

———— 1991. 一般化可能性理論 generalizability theory による「移行」による影響の査定の可能性。日本語聴解問題の改善に関する考察。(日本語教育学会調査研究委員会平成3年度報告書)

村上 隆 1991. 日本語能力テストの評価—計量心理学の立場から—日本語聴解問題の改善に関する考察。(日本語教育学会調査研究委員会平成3年度報告書)