測定における妥当性の理解のために
――言語テストの基本概念として――

清　水　裕　子

ABSTRACT

This article provides a review of changes in conceptions of validity in psychometrics and in educational measurement. Validity, as well as reliability, is a fundamental construct in testing theory. In *Standards for Educational and Psychological Testing* (*Standards* hereafter), which addresses technical and professional issues of test development and use, validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (1985, p. 9). This concept, combined with Messick's conceptualization of validity (the foundation of *Standards*), has been influential in current validity theory. In order to provide background knowledge for those who are concerned with language testing, a history of *Standards* is introduced briefly in this article, followed by a comparison of the traditional view of validity and of the current view conceptualized by Messick. Finally, the influence of the changes of validity issues to the field of language testing and various implications for the field are presented.

0.0　はじめに――信頼性と妥当性

心理測定および教育測定の分野で登場する用語に信頼性（reliability）と妥当性（validity）がある。前者は，同じ対象（例えば受験者）に対して，ある観測（例えばあるテスト）をした場合，同じような測定値が得られているかどうかを観察する指標であり，実際に観測した得点の中に真の得点が占める割合のことを言う。それに対して，妥当性とは，ある尺度（scale）やテストの，特定の変数（例えば，語彙力など）に関する測定道具としての適切さを示すものであり，意図した特質や能力を測定できているかどうかを示す指標と言える。そして，これらはテストが備えるべき重要な要素として考えられている。信頼性については，その推定方法として，再テスト法（test-retest method），折半法（split-half method），クロンバック・アルファ（Cronbach's alpha），キューダー・リチャードソン公式（KR-20，KR-21）等の手法がある。本稿で取り上げる妥当性に関しては，その検証のための手法として因子分析や相関などが用いられるが，その方法論以上に，妥当性そのものの概念についての議論がなされてきている。一般に妥当性と言った場合，内容的妥当性（content validity）や併存的妥当性（concurrent validity）などの各種の妥当性に言及することが多いが，その一方で統合的な妥当性という考え方が存在する。つまり，テスト項目の質の決定やテスト原理の基本となる妥当性の分類については，後述

するように，三位一体的に各種の妥当性をカテゴリー化する伝統的な見方に対して，統合的あるいは一元的に妥当性を捉える考え方が20世紀後半に登場してきたのである。ただし，後者の考え方については，妥当性理論に大きな影響を与えてきているものの，心理測定や教育測定の分野でも議論が進行しているのが現実のようである。

言語テストの研究分野は，応用言語学の立場に立つと共に，心理･教育測定の分野としての知見をも必要とし，様々な研究や調査，あるいはテスト開発に関わる際の背景知識のひとつとして，妥当性の概念とその歴史的な流れを理解しておくことは有効であると思われる。そこで本稿では，心理･教育測定の分野における妥当性の概念の変遷と密接な関わりをもち，また米国における心理･教育測定に関するガイドラインとも言えるStandards for Educational and Psychological Testingの改訂の歴史を紹介することから話を進めることにする。次に，妥当性について，測定と評価に関する多くの入門書で示されている伝統的な捉え方およびMessickによる妥当性の概念を中心に，最近の心理･教育測定における観点を紹介し，最後に，言語テストの分野への影響に触れることで，本稿が言語テストの作成や分析に関わる者に，なんらかの示唆を与えることを期待する。

0.1　基本用語について

テストの妥当性について論を進めていく前に，最も基本的であり，且つ一般的に用いられている「テスト」ということばの概念を明確にしておくことにする。

教育に関わるテストは，心理学の分野の心理測定（psychometrics）の伝統から直接取り入れられてきている（Cziko, 1980, p.28）。「テスト」と言った場合，大きな枠組みの中では心理テスト（psychological test）のことを指し，教育テストも教育現場における心理測定として捉えることもある（Murphy & Davidshofer, 2001）。それは個人がもつ種々の特性や能力を測定するために用いる道具であり，一連の質問や問題，タスクによって構成されたものと定義できよう。テストの分類方法は種々あるが，理解の度合いや記憶，問題解決能力のような力を測定するものと，何かに対する態度や興味，動機などの個人の特性を引き出すためのテスト（性格テストやインベントリーと呼ばれることが多い）に分けることもできる。心理学の専門の立場では，両者に対する明確な用語があろうが，Brown（1983）が心理・教育測定の入門書で紹介している方法が，我々にとっては理解しやすい。つまり，前者（到達度テストや適性テスト，さらに教育現場におけるテストをも含む）を<maximal performance tests>とし，後者を<typical performance tests>としている。筆者自身は，外国語としての英語力に関して，言語使用や言語理解の分野に関心を持つ者の一人であるが，本稿で論じる妥当性の話の中では，広い意味での「テスト」を想定している場合が多いことを断っておく。

ところで，「一連の質問や問題，タスク」は個々の「項目（item）」から構成されている。「項目」という用語については，例えば，項目分析や項目難易度，項目応答（反応）理論などの用語の中に登場するにも関わらず，明確な定義を示した文献は少ない。「外国語教育リサーチとテスティングの基礎概念」（靜等，2001）の中では，項目とは「受験者から一定の応答を引き出すことを目的として設定された一つの課題。テストの，いわゆる『小問』の一つ一つを指す。項目の有機的集合体がテストとなる。」としている。また，心理測定を含む複数の専門分野の観点

から，より専門的，包括的な定義としてOsterlind（1990）の定義が優れている。すなわち，以下のとおりである。

> A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an ability, predisposition, or trait) may be inferred.(p. 3)

この定義の解釈についてOsterlind（1990）の解説をまとめると，テスト項目というものは，（1）（主観的および客観的）測定機能を備えており，何からの数値化されたデータを導き，（2）所定の反応を引き出すための刺激を与えるものであり，（3）その反応は，受験者の特性について情報を提供するものであることになる。

テスト作成に当たり最も重要なことは，測定しようとしている内容領域が明確で，構成概念（construct）を反映した項目によって構成されていることである。そのためには，項目の形式の特質や機能を理解し，作成にあたってのガイドラインが明確であるなどの，実際的な問題が関わってくるわけであるが，それは項目作成などの実践書に委ねることにし，以下，本稿の中心である妥当性の話に論を移して行く。

1.0　妥当性の概念とStandards for Educational and Psychological Testing

教育にまつわる決定のために様々なテストが使われる。テストの作成・開発にあたる者や利用者，あるいはあるテストの採用を決定する者として心得ておかなければならないことが数多くあるが，それらを専門的な立場からまとめ，米国におけるテスト開発等の公式の規約となっているのがStandards for Educational and Psychological Testing Standards（以下，Standards）（APA, AERA, & NCME, 1985および1999）である。これには，テスト開発から実施および結果の活用，その後の改訂などの様々な段階で下す評価や判断のためのガイドラインが<standard>として列挙されている。テスティングの分野で必須となる情報や基準を総括したものであり，その意図するところは，“to promote the sound and ethical use of test and to provide a basis for evaluating the quality of testing practice”（p. 1）としている。妥当性の概念の歴史を見る場合にも，Standardsの存在は大きな意味をもつと言えるし，さらに心理・教育測定の分野の延長として言語テストを捉えた場合，この存在と背景を知っておくことは意義がある。そこで，1999年版Standardsに至るまでの歴史と，その関連から妥当性の概念の変遷を概観する。

1.1　Standardsの歴史

Standards はAmerican Educational Research Association（AERA），American Psychological Association（APA）およびNational Council on Measurement in Education（NCME）の3団体の共同で編集されたもので，その歴史は半世紀前にさかのぼる。Geisinger（1992）によると，まずAPAの委員会が1954年に準備したTechnical Recommendations for Psychological Tests and

Diagnostic Techniques（以下，Technical Recommendations）に始まり，翌年AEREおよびthe National Council on Measurement Used in Education の委員会により作成され，National Education Associationによって Technical Recommendations for Achievement Tests として出版される。 次いで1966年のAPA出版により，前掲２種の文書をもとにAERA ，APAおよびNCMEの代表委員会によってStandards for Educational and Psychological Tests and Manuals（以下，Tests and Manuals）が準備され，さらにそれを改訂した形で，同３団体により1974版Tests and Manualsが出版されている。また1970年代後半における技術の進歩やテスティングに関する諸状況の変化により，1974年版 Tests and Manualsの改訂の必要が生じ，100名以上の専門家の助言等をもとに，1985年のStandardsの出版に至った。この経緯の詳細についてはStandards（AERA, APA, & NCME, 1985）のPrefaceに譲ることにするが，その後1990年代に入り，1985年版Standardsの改訂の必要性が提言され，現在の1999年版が誕生したのである。Standards 両版の詳しい比較は本稿では行わないが，1985年版が約100ページであったのに対し，1999年版はその２倍に渡り，それぞれの章がより詳細な情報を提供するようになった。提示するstandardの数も多く，巻末の用語解説や索引もより豊かになり，さらに，米国の連邦法の変更をはもとより，妥当性に関わる測定理論の動向をも反映している。また，障害者や異なる言語の背景を持つ者に対するテストの公平さに関する章や，コンピュータを用いたテスト等の新しいタイプのテストの使用への言及もある。

心理・教育測定の研究における長い歴史をもつ米国において，このようなガイドラインが存在することは当然とも言えよう。また米国のテスト作成や開発，測定と評価に関する各種文献や定期刊行物の多さには目を見張るものがある。さらに，理論面のみでなく，教育現場にも反映できる実践書も数多く出版され，そこにはStandardsの少なからぬ影響と貢献を窺い知ることができる。

１.２　Standards改訂の歴史と妥当性概念の変遷

妥当性に関する文献の中には，その定義にまつわる歴史を概観しているものがある（Angoff, 1988; Cronbach, 1988; Messick, 1988, 1989b; Sireci, 1998）。1950年代初めまでの定義については，Guilford（1946）がしばしば引用されるが，そこでは “in a very general sense, a test is valid for anything with which it correlates”（p.429）としており，テストによって観測された得点と，外的基準となる得点との相関を示すことで妥当性を捉えている。しかし，用いる外部基準や基準そのものの問題が指摘されるようになり，また1950年代に入り，テスティングや測定に関する基準が専門家により設定されはじめると共に，妥当性の概念や検証方法等についても多くの議論がなされるようになっていった。ここでは，先に紹介したStandardsの改訂の歴史と共に，妥当性の概念の変化を概観することにする。

1954年に出されたTechnical Recommendationsでは，妥当性を４つに分類して紹介している。すなわち，併存的（concurrent）・予測的（predictive）・構成概念的（construct）および内容的（content）妥当性である。（詳細については後述）これらの妥当性は＜aspects＞として示されているが，実際には，むしろ＜types＞として捉えられていたと言える（Angoff, 1988, p.25）。Technical Recommendationsの作成では，委員としてCronbachやMeehlが関わっており，彼等

の考え方の影響から，構成概念妥当性が前面にでてきているものの，基準関連妥当性（criterion-related validity）を重視する当時の風潮が継承され，これを併存的妥当性と予測的妥当性の２種の妥当性に分け，構成概念的妥当性および内容的妥当性と共に４種類の妥当性が並列的に示されていた。しかし，1966年版 Tests and Manualsでは，併存的妥当性および予測的妥当性は，基準関連妥当性としてひとつにまとめられ，３種類の妥当性への言及に変化し，さらに複数の妥当性検証の必要性を説くようになった。このことは，後に見られる妥当性の＜統合化＞の前触れとも考えられよう。

1974年版Tests and Manualsでは，時代の流れに伴って，テスト･バイアス等への言及もみられるようになったが，妥当性の定義に関わる大きな変化は1985年版Standardsになってからであった。つまり，ここで初めてテストの妥当性の概念が，今までのようにカテゴリー化されたものではなく，一元的（a unitary concept）なものであると明言されるようになったのである。さらに，妥当性は“the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores”（AERA, APA, & NCME, 1985, p.９）と定義され，テストの得点自体に意味があるのではなく，得点をもとに引き出す的確な<推論>の重要性を強調するようになった。

2.0　伝統的妥当性観から新たな妥当性へ

米国のテスティングの分野に関わる者にとって必携書であるStandardsが，心理･教育測定の第一人者たちによって編集されていたことからも，改訂の歴史にその分野からの影響を受けていることは当然と言えよう。その中で，妥当性の考え方の変化が1985年版以降に見られたことは先に触れたが，ここでは，その変化に大きく関わったMessickによる考え方を中心に，それ以前の伝統的な観点と新たな妥当性の捉え方について見ていく。

2.1　妥当性とMessick

1985年版Standardsにおける妥当性の定義については，Standardsの委員として関わっていたSamuel Messickの影響が大きいと言える。彼は，妥当性の考えを行動科学や社会科学と結びつけ，新たな妥当性の概念を打ち立てた人物であり，またテクノロジーの発展と共に学習媒体や方法，さらに測定方法における大きな変化を予測した上で，基本的な測定理論，とりわけ妥当性の本質というものは普遍的であることを説いている。Messick（1988）は，妥当性というものを“an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores”（p.33）としており，ここにStandardsの定義との関連を見ることができる。つまり，妥当性を検証することによって，テストの結果として現われるテスト得点（test scores）[1)]の解釈が当該の目的に対してどれだけ適切で意味のあるものであり，また有用性があるかということと，その結果を用いることや解釈の仕方が社会や個人にどのような影響を与えるかということに対する答えを得ることになる。なお，1985年版Standardsでも，construct-related・content-related・criterion-related という，伝統的な用語を残しているが，妥当性の諸側面を解明するための＜証拠＞になる要素について論を進めており，types of validityではなく，validity evidenceと言う表現を用いている。そして，

妥当性が２値的にその有無を問題にするのではなく，検証された証拠と測定の目的により，妥当性の異なる度合いが存在することを示唆している。

２.２　伝統的な観点からの妥当性

妥当性に関するMessickの影響が示唆に富むものであるとしながらも，その考えが浸透していないとする者（DeVellis, 2003）や，社会科学者や行動科学者の中には，伝統的な立場に従う研究者も存在するのが現実のようである（Hubley and Zumbo, 1996）。また構成概念的妥当性を統一的なものとして捉えた上で，それを示す根拠として，伝統的な妥当性の分類を含めた種々の妥当性の下位構造を示す者もある（Netemeyer, R.G. et al, 2003, pp.71-87）。

表１は伝統的な観点での妥当性を分類したものである。この３分類が，妥当性の＜aspects＞なのか＜types＞なのかについては一致した見解がないようであるが，伝統的な観点では，それぞれ個々の存在として三位一体的に捉えられている（Hubley & Zumbo, 1996, p.210）。また，Guion（1980）は，この３つの妥当性を，“something of a holy trinity representing three different roads to psychometric salvation”（p.386）と捉え，Cronbach（1988）のことばでは，“separate but maybe equal”（p.４）と表現されている。一元的に捉えた妥当性（あるいはその意味での構成概念的妥当性）を理解する上でも，伝統的な個々の妥当性について知ることが必要であると考え，次に簡単な説明を加える。

表１　伝統的観点からの妥当性の分類

基準関連妥当性	
併存的妥当性	そのテストが，同時に測定されるある基準をどれだけ適切に推定しているか。当該テストとほぼ同時に収集された基準との相関により検証。
予測的妥当性	そのテストが，そのテスト実施以降の変化等をどれだけ適切に予測しているか。当該テストと，それより後の基準との相関により検証。
内容的妥当性	そのテストを構成している項目が，全体を偏りなくどれだけ適切に代表しているか，また構成概念をどれだけよく反映しているか。その分野の専門家により判断，検証。
構成概念的妥当性	そのテストが，測定しようとする構成概念や潜在特性をどれだけ適切に反映しているか。因子分析や相関などの種々の検証方法がある。

まず，基準関連妥当性は，テスト得点と他の基準との関係をみるものであり，その下位に併存的妥当性と予測的妥当性がある。前者は，当該テストの得点とある基準とが，ほぼ同時に測定された場合の関係を示すものであり，後者の予測的妥当性は，そのテストが，そのテスト実施以降の変化等をどれだけ適切に予測するかを示す。内容的妥当性は，テストが測定しようとしている分野や内容についての専門家の判断に基づき，内容の適切さや代表性をみるものである。構成概念的妥当性は，ある特性や概念を測定するために設計されたテストを評価するものであり，テスト得点という変数と他の変数との理論的な関係をみるものである。なお，「構成概念」とは理論的に存在し，直接観察できない抽象的な特質や能力のことであり，テストという

道具を通じてその特質や能力を数値化することになる。

2.3　新たな妥当性理論 ― “Everything is construct validation”

Technical Recommendationsの発行とCronbach やMeehlの時期に，妥当性理論の変化の兆しが見られたわけであるが，Cronbach and Meehl（1955）は，すべてのタイプの妥当性の中でも，構成概念的妥当性を最重要視し，併存的妥当性や予測的妥当性も，構成概念的妥当性の検証に有用なものとして捉えていた。またLoevinger（1957）が，“since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view.”（p.636）と表現しているが，これがMessickの考え方でもあったと言える。この1950年代後半の新しい見地が，実際に多くの支持を得られるようになったのは1970年代になってからであり（Hubley & Zumbo, 1996, p.211），Messick等の貢献に負うところが多かったと言えよう。

ところで，妥当性とは実に抽象的な概念であり，何か漠然としたものである。例えば，ある学習者の「英語力」というものは具体的に見ることができないものなので，様々のタスクやテストを実施して，「得点」という仮の媒体によって「英語力」とみなすのと同じように，妥当性というものも，直接観察することはできない。そのため，刑事が事件を解明するように，様々な＜証拠＞をもって妥当性を検証することになる。つまり，Messickでは，観察された数値から直接観察できない構成概念を推測することが測定の目的であり，用いた測定道具の評価とは，なされた推測の構成概念的妥当性を評価することとするものである。ここでは，内容的妥当性や基準関連妥当性などが，より大きなカテゴリーである構成概念的妥当性の枠組みに包括されている。また，先に述べたように，1985年版Standardsではconstruct-related evidence，content-related evidence，criterion-related evidenceという表現が用いられているように，これらの＜証拠＞が妥当性の検証，あるいは妥当化（validation）につながるとし，妥当性というものがテストに備わった静的なものではなく，テスト得点からの推論をより確実なものとするための証拠を蓄積していく過程として動的に捉えている。

Messick（1980）は，テストの妥当性の<aspects>に関して，伝統的な分類との関連を表３のように示している。

表３　Alternative Descriptions for Aspects of Test Validity

Validity designation	Descriptive designation
Content validity	Content relevance --- domain specifications Content coverage --- domain representativeness
Criterion validity 　Predictive validity 　Concurrent validity	Criterion relatedness 　Predictive utility 　Diagnostic utility 　Substitutablity

（Messick ,1980, p.1015　Table 1 より一部抜粋）

ここでは，伝統的な妥当性の分類で用いられていたvalidityという用語を，relevance やutilityということばに代えている。Geisinger（1992）は，これを“downgrade”と言う表現を使って表しているが，Messick自身は，（彼の定義における）構成概念に関連した証拠は，構成概念に基づいた推論はもとより，内容や基準に基づいた推論をも補い補強するものとしている（1988, p.40）。Content relevance/ coverageやcriterion relatednessは，伝統的な分類では，それぞれ内容的妥当性，基準関連妥当性に呼応するとも考えられるが，Messickの観点からすれば，content relevance/ coverageは，それ自体はあるテストを受けた者の能力に対する推論を可能にはしないので，妥当性を証明する情報としては十分ではなく，また criterion relatednessについても，使用する基準そのものの妥当性の整合性が問われてくることになる。これらは，テスト得点の解釈に貢献するものであるが，構成概念に関する証拠の下位概念として包括されるものであり（Messick, 1989b），用いる用語や位置付けの違いはあるが，決してこれらの下位概念を軽視したものではないと言える。つまり，伝統的な妥当性の概念が分割的であったのに対し，新たな概念では，Osterlind（1998, p.64）のことばを借りれば，各要素は統一的な意味での妥当性を支持するための<convenient categories> としてみなすことになるのである。さらに，Messickの観点ではテストがある社会の中でどのように使用され（use），どのような影響を社会に与えるかという大きな枠組みからも妥当性を捉えており，Bachman（1990）に引用されているMessickの次の表現がそれを示している。

> Examining the validity of a given use of test scores is therefore a complex process that must involve the examination of both the *evidence* that supports that interpretation or use and the *ethical* values that provide the basis or justification for that interpretation or use.（p.237）

このMessickの概念を示したのが表４である。この表が示すように，Messick は妥当性を２つの次元から捉えている。つまり，テストの結果に関する次元（Function or outcome）と，テストという行為の正当化に関する次元（Justification）である。前者では，その属性としてテスト結果の＜解釈（Test Interpretation）＞およびその＜使用（Test Use）＞を，また後者については，テストの解釈と使用を正当化するための＜証拠（Evidential Basis）＞および＜結果（Consequential Basis）＞の側面を備えている。４つの象限の内，象限Ⅰはテストの解釈を支持するための証拠であり，テストが何を測定しているかを示すものである。その他の３つの象限（Ⅱ.～Ⅳ.）は，伝統的な妥当性の観点には含まれていなかったもので，そのテストが意図したように適用され，またその利用価値があるか否かにかかわる象限（Ⅱ），ある社会の価値の中でそのテストが測定していることの適切さに関する象限（Ⅲ），およびそのテストを用いることがその社会に与える重要性や影響に関する象限（Ⅳ）に分けられる。

表４　Messick's Progressive Matrix of Validity（Messick,1989, 1995）

		Function or outcome	
		Test Interpretation	*Test Use*
Justification	*Evidential Basis*	Ⅰ. Construct Validity (CV)	Ⅲ. CV + Relevance/Utility (R/U)
	Consequential Basis	Ⅱ. CV +Value Implications (VI)	Ⅳ. CV + R/U + VI + Social Consequences

（各象限内のⅠ.～Ⅳ.の表記は筆者による）

注目すべきことは，４つの象限すべての基本要素として構成概念的妥当性が登場していることである。例えば，もしあるテスト得点の解釈を正当なものとするには，構成概念的妥当性を示すだけでなく，その解釈のもつ価値についても考えなければならない。また，ある目的で何らかのテスト得点を使用するなら，構成概念的妥当性および価値を考慮することに加えて，その使用の適切性･有用性と社会的な影響を考えなければならないことになる。つまり，テスト結果の倫理に適った使用と解釈のためには，伝統的な観点における個々の妥当性をチェックリスト的に確認するのでは十分ではなく，構成概念的妥当性，価値体系，現実に即した有用性，教育システムや社会への影響を考慮しなければならないことになる。

２.４　1999年版Standards

再度，Standardsに触れておく。1999年版Standards（AERA, APA, & NCME, 1999）のIntroductionにおいて，構成概念の測定としてのテストについての説明がされている。

> ..... no longer speak of different types of validity but speak instead of different lines of validity evidence, all in service of providing information relevant to a specific intended interpretation of test scores. Thus, many lines of evidence can contribute to an understanding of the construct meaning of test scores. (p. 5)

つまり，個々の異なるタイプの妥当性があるのではなく，以下の５つの範囲（Sources of validity evidence）にわたる情報は，“...illuminate different aspects of validity, but they do not represent distinctive types of validity. Validity is a unitary concept.”（p.11）と明言し，さらに，伝統的な考え方との違いを明確にするために，1985年版Standardsで使用していたvalidity evidence という表現も用いられなくなっている。詳細は1999年版Standards（AERA,APA, & NCME, 1999，pp.11－16）に譲り，その概要を示すに留めておくが（表５参照），これはMessickの観点を取り入れながら，1985年版Standardsよりもさらに理解しやすい区分になっている。テストの妥当性は，そのテストが何を測定しているかに関する情報だけでなく，そのテストの使用の有用性に関する情報にも関わる概念であり，また社会に与える影響にも関わる概念も考慮した概念であることは，新しい区分においても明らかになっている。

表５　Sources of evidence in the 1999 Standards

Sources of evidence *based on*.....	
Test content	項目の形式や採点方法などの変数とテスト内容の関連性について
Response processes	構成概念の性質とテストによる受験者の行動や回答プロセスの連関の重要性について
Internal structure	次元性（Dimensionality）にも関わることで，テストの項目とその結果〔得点〕が，構成概念の構造と対応しているかどうかについて
Relations to other variables	テストと他の測定との相関関係に関して
Consequences of testing	テスト得点の解釈と使用が与える影響について

3.0　言語テストと妥当性

心理・教育測定の分野での妥当性の概念の変遷について概観してきたが，ここで，言語テストの分野への影響と，最近の研究動向に見られる妥当性の検証方法について触れておく。

3.1　言語テスト研究への影響

例えば，テスト作成のための実践書や入門書では，伝統的な立場で３分類（あるいは４分類）をしている場合が多い（Henning, 1987; Hughes, 1989; Brown,1996）。しかし，1985年版Standardsの出版とEducational Measurement 第３版（1989）のMessickの妥当性に関する章が心理･教育測定分野に与えた影響は，1990年代に入り，言語テストの分野にも確実に現れていった。

Bachmanは，著書Fundamental Considerations in Language Testing（1990）の妥当性に関する章の中で，Messickの妥当性理論の展開に触れながら，一元的な概念としての妥当性を紹介している。彼は“We will still find it necessary to gather information about content relevance, predictive utility, and concurrent criterion relatedness, in the process of developing a given test.”（p.237）（下線は筆者による）と述べ，個々の要素が単独では不十分であることを強調，さらに，テストはそのおかれた教育風土や社会の中で実施されるものであり，テストが与えるそれらの影響を考慮することの重要性を説いている。またKunnan（1999）はMessickの枠組み（表４）をもとに，evidential basisにおけるテストの解釈の象現（表内のⅠ.），つまり構成概念的妥当性に対して，今まで多くの注目が集まっていたが，最近になって他の３つの象現（Ⅱ.～Ⅳ.）も注目されるようになったとしている。特に，テストを受ける際の過程やストラテジー，受験者の特質，倫理などに関する研究が十分でなかったことを指摘しているが，実際，1990年代に入り，テストの波及効果やテストに対する社会の責任等についても言語テストの話題や研究の中にしばしば登場するようになってきた。

3.2　妥当性の検証

妥当性をどのように捉えるにしても，その検証には唯一の方法というのがあるのではなく，複数の観点から検証しなければならないのが，多くの研究者に共通した見解と言えよう。具体的な方法としては，相関関係や因子分析，事前－事後テスト間の差の有意性の検証や性別や言語背景などをもとにしたグループ変数間の有意性の検証などを行うことになろう。

言語テストの分野に関しても，Bachman（1990）は，相関による分析だけではなく，グループ間に観察される差や，時間の経過による変化や種々の実験的な処理の効果などによる検証の有用性を示している（p.258）。また，近年，様々な統計手法を用いた研究が数多くみられるようになってきている。例えば，項目応答理論（IRT）を用いて，ESPのリスニング･テストを分析した研究（McNamara,1990，1991）や，ラッシュ･モデルによるリスニング･テストの分析（De Jong and Glas, 1989）に始まり，項目の難易度と受験者の能力という２つの相以外の相をも加えて分析する多相ラッシュ測定（McNamara and Lumley, 1979）なども行われている。このような量的研究に加えて，受験者の内省（introspection）や外省（retrospection）による質的な研究もみられ（Storey,1997；Green,1998），Bachman & Palmer（1996）においては，あるテスト使用の環境で得られたテストの得点からの推論には，質･量的両側面からのアプローチの必要性を説いている（p.92）。

4.0　おわりに

テスト結果として得られるテスト得点の使用や，それをもとに行う推論や決定を正当なものとするためには妥当性の検証が必要であり，Messick以降，言語テストの分野においても，テストが果たす社会的な役割や意味をも考慮した上で，様々な検証の必要性やその方法がとりあげられてきている。適切に収集されたデータは，受験者集団に関するデータをも含めて，すべて妥当性の検証に関わる貴重なデータとなり得るわけであるが，Angoff（1988）は，その中にContent　Analysisを含んでいることに注意したい。

良いテストの作成に関して，心理測定面とデザイン面とに分類して紹介した入門書があるが（Friedenberg, 1995），それによれば，心理測定面では信頼性や項目分析に加えて妥当性の話をとりあげており，デザイン面では，その特質のひとつとしてテストの目的と測定領域（domain）を明確にすることの重要性に触れている。これは，伝統的な分類のことばを借りれば，内容的妥当性に関わることである。Messickの観点では，妥当性は，テストそのものについてではなく，得点からの推論のための原理であるとしているが，実際に正しい推論を行うためには，良いテストが作成，実施されていなければならないことは当然であり，構成概念妥的当性をより的確なものにするためにも必須である。このことは，社会指標の研究分野においてSireci（1998）が指摘しており，内容的妥当性の重要性の議論を行っている。彼は，一元的な妥当性の概念においては，「内容的妥当性」が，content coverage，content relevance，content representativeness等のことばで表現されているが，標準的な英語としても<content validity>という用語が，テストの質を示す用語として，なんら問題なく，より適した表現だとし，さらに心理測定の専門でない者にも容易に理解できることばだとしている（p.104）。また，Health　Science分野のBeck

& Gable（2001, p.202）は，Waltz, Strickland & Lenz（1991）の研究をとりあげ，内容的妥当性は構成概念的妥当性や基準関連妥当性には不可欠であり，測定道具の開発では最重要視しなければならないことを強調していることを紹介している。どのような表現を用いるにしても，測定の専門家たちが意図するところは共通であり，内容的に妥当なテストを作成することは，テスト開発の過程で必須要件であり，テスト得点からの推論に影響を与えるものである。

教室で使用する場合であれ，調査研究や資格認定に用いる場合であれ，測定道具を作成することは，大きな責任を負う活動である。実際にテスト開発に関わる場合は，Hattie, Jaeger, and Bond（1999, pp.393-394）の示すように，それが短期的に使用するテストであれ，長期的なものであれ，＜測定概念の決定＞，＜テストおよび項目の開発＞，＜テストの実施＞，＜使用＞，＜評価＞を繰り返すことになる。初期の段階では，測定概念を明確にし，テスト項目の細目表（specification）を作成するなど，内容に関わる基盤を作りあげることが必要であり，また実施後は，その評価を行う必要から，妥当性理論に則った検証等が関わってくる。理論面および実践面での情報のネットワークをはりめぐらして，得られた情報のフィードバックを行いながら，テスト開発に関わっていくことが求められているのである。

本稿では，米国における心理・教育測定やテスト開発に少なからぬ影響を与えているStandardsの歴史を概観しながら，妥当性理論の変遷を中心に，言語テスト分野への影響にも触れてきた。世界最大規模の教育測定に関する非営利団体である米国のETS（Educational Testing Service）では，Center for Validity Research[2)]のもとで，特に高等教育におけるテスト（TOEFL，GRE等）に関して，教育政策や測定方法，フェアネスなどの調査･研究が進められているようであり，そのような環境に関わる研究者の観点から比べれば，テスティングに関して限られた経験しかもたない筆者の理解と洞察が十分でないことや，さらに深い考察がなされるべきであることは認める。しかし，本稿が，教室レベルを含め，さまざまな場面でのテスト作成や開発に関わる者に，なんらかの新たな示唆を与えることを期待する。Standardsの存在に加えて，教育および心理に関連する望ましいテスト開発と測定・評価を目指して1978年に設立された団体であるInternational Test Commission（ITC）[3)]により，テスト使用に関するガイドライン[4)]が作成されるなど，社会の中の個人に関わる様々な決定のための道具となり得るテストを，より精錬さらた整合性のあるものにするために，テスト使用者に対する働きかけがみられる。わが国においても，日本テスト学会が，Messickの観点に立ち，構成概念的妥当性の検証に触れながらThe JLTA Code of Good Testing Practiceを提唱し，米国のStandardsに相当するような規約がわが国の言語テストの分野でも登場しつつある。このことは意義があり，また喜ばしいことであると共に，今後，日本の教育風土とテスト環境において，影響力のある存在になっていくことを期待する。

注

1）Messickは，テスト得点（test scores）ということばを用いているが，これは必ずしも数値で現われた結果のみを意味しているわけでなく，以下の引用を参考にされたい。

The term “test score” is used generically here in its broadest sense to mean any observed consistency,

not just on tests as ordinarily conceived but also on any means of observing or documenting consistent behaviors or attributes. This would include, for instance, any coding or summarization of observed consistencies on performance tests, questionnaires, observation procedures, or other assessment devices. This general usage also subsumes qualitative as well as quantitative summaries and applies for example, to protocols, clinical interpretations, and computerized verbal score reports. (1989a, p. 5)

2）http://www.ets.org/research/he.html 参照。

3）International Test Commission に関する詳細は　http://www.intestcom.org/faqs.htm　参照。

4）International Guidelines for Test-use については http://www.intestcom.org/test_use_full.htm　参照。

参考文献

靜哲人等編著（2001）.『外国語教育リサーチとテスティングの基礎概念』大阪：関西大学出版部

Allen, M.J. & Yen , W. M. (2001). *Introduction to measurement theory*. Prospect Heights: Waveland Press, Inc.

AERA, APA, & NCME. (1985). *Standards for educational and psychological testing*. Washington, DC : American Psychological Association, Inc.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC : American Educational Research Association..

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp.19-32). Hillsdale, NJ: Lawrence Erlbaum.

Beck, C. T. & Gable, R. K. (2001). Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement*, 9 , 2 , 201-215.

Brown, F.G. (1983). Principles of educational and psychological testing (3rd). Orland, Florida: Holt, Rinehart and Winston, Inc.

Brown,　J. D. (1996). *Testing in language program*. Prentice-Hall Regents

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 4 , 281-302.

Cziko, G. A. (1980). Psychometric and edumetric approaches to language testing: Implications and applications. *Applied Linguistics*, II, No 1 , 27-44.

De Jong, J. and Glas, C. A. W. (1989). Validation of listening comprehension tests using item response theory. *Language Testing*, 4 , 2 , 170-194.

DeVellis, R. F. (2003). *Scale development* ： *Theory and application (2nd ed)*. Thousand Oaks, Ca.: Sage Publications, Inc.

Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Needham heights, MA: Allyn & Bacon.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27, 2 , 197-222.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6 , 427-438.

Guion, R.M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.

Hattie, J., Jaeger, R.M., and Bond, L.(1999). Persistent methodological questions in educational testing. *Review of Research in Education*, *24*, 393-446.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hubley , A. M. and Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123,* ３, 204-215.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

JLTA. (retrieved 2003). The JLTA Code of Good Testing Practice. from JLTA Web site: http://www.avis.ne.jp/~youichi/COP.html

Kunnan, A. J. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics, 19,* 235-253.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports,* ３. 635-694. [Monograph Supplement ９].

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing,* ７, １, 52-75.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing,* ８, ２, 139 -159.

McNamara, T. F. & Lumley, T. (1997). The effect of interlocutor and assessment more variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14,* ２, 140-156.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist. 35,* 11, 1012-1027.

Messick, S. (1988). The once and future uses of validity: Assessing the meaning and consequences of validity. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp.33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989a). Meaning and values in test validation: the Science and ethics of assessment. *Educational Researcher, 18,* ２, ５-11.

Messick, S. (1989b). Validity. In R.L. Linn (Ed.), *Educational measurement* (３rd. ed.) (pp.13-103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1996). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* ９, 741-749.

Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications (*５th *ed).* Upper Saddle River, NJ: Prentice Hall.

Osterlind, S. J.(1990). Toward a uniform definition of a test item. *Educational Research Quarterly, 14.*４. ２-５.

Osterlind, S. J.(1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats (*２nd*).* Norwell, MA: Kluwer Academic Publishers.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing,* ６, １, 77-94.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research,* 45, 83-117.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective in the discourse cloze test. *Language Testing,* 14, ２, 214-231.

Waltz, C., Strickland, O., & Lenz, B. (1990). Use of qualitative methods to enhance content validity. *Nursing Research, 39,* 172-175.