Title	教育評価とテストの妥当性
Sub Title	On the construct validity of test in educational evaluation
Author	渡辺, 恵子(Watanabe, Keiko)
Publisher	三田哲學會
Publication year	1987
Jtitle	哲學 No.84 (1987. 5) ,p.119- 135
JaLC DOI	
Abstract	Evaluation of instructional objectives is essential in education in order to make instruction more fruitful by providing feedback to teachers and students. With regard to criterion-referenced evaluation, construct validity has been emphasized more and more in educational testing during the past few decades. This paper is to clarify the relationshipe between the constructs of educational objectives and the test which are intended to measure these objectives; furthermore, the importance of construct validity is emphasized. The differences between psychometric measurements, which focus on individual differences, and edumetric measurements, which focus on withinindividual growth, are also pointed out. The effects of educational instruction are properly measured by edumetric testing. The pretest-posttest method to measure the effectiveness of instruction has been said to have some shortcomings. However, the effectiveness of the method is mentioned.
Notes	
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00150430-00000084-0119

教育評価とテストの妥当性

渡　辺　恵　子*

On the construct validity of test in educational evaluation

Keiko Watanabe

Evaluation of instructional objectives is essential in education in order to make instruction more fruitful by providing feedback to teachers and students. With regard to criterion-referenced evaluation, construct validity has been emphasized more and more in educational testing during the past few decades. This paper is to clarify the relationshipe between the constructs of educational objectives and the test which are intended to measure these objectives; furthermore, the importance of construct validity is emphasized. The differences between psychometric measurements, which focus on individual differences, and edumetric measurements, which focus on within-individual growth, are also pointed out. The effects of educational instruction are properly measured by edumetric testing. The pretest-posttest method to measure the effectiveness of instruction has been said to have some shortcomings. However, the effectiveness of the method is mentioned.

* 昭和大学教養部助教授・慶應義塾大学文学部，法学部，教職課程センター非常勤講師（心理学）

Ⅰ．教育評価の目的

教育評価は，テストを実施し点数をだし，成績をつければそれでおしまいというものではない．教育活動には，実に多くの要因が直接的あるいは間接的にかかわっている．教育場面に直接かかわる学習者（児童，生徒）および教師はもちろんのこと，カリキュラム，学校の諸施設，地域的環境，教育行財政的条件に至るまで，さまざまの要因を挙げることができる．より良い教育活動を行い，より良い成果をあげるためには，教育活動に関連した上記の諸要因の一つ一つに関して，その実態を見極め，それにもとずいて何らかの教育上の意思決定をすることが必要である．そのためには，まず，現状についてのできるだけ正確な資料を収集し，それにもとずいて関連諸要因を評価しなければならない．しかし，収集すべき資料の種類，資料の収集方法，収集された資料の評価方法，また評価結果にもとずく意思決定の仕方等は，どの要因を問題にしているのかによって異なってくる．例えば，教育予算，学級の適正規模．指導要領等に関する決定は，教育行政に関するものであるが，このような評価のために必要な情報は，特定の地域，特定の学校においてではなく，全国的規模で収集するのが望ましい．また，教師の学習指導を評価し，指導方法に関してなんらかの決定をするためには，教えようと意図した事をクラスの子供たちがどの程度理解したか，またどういう点がわからなかったのか等について，日常の指導の中での観察や問答および節目節目で行うテストなどにより，できるだけ詳しいデータを得ることが必要である（東，1979）．

また，同じ資料に対する評価も，何を目的としているかにより異なったものになることもある．例えば，或る子供の或るテストにおける得点が，その子供のこれまでの得点と比較するとかなりの改善がみられるが，クラスの他の子供の得点と比較すると劣っているとき，客観的には同一の得点でも，他の子供と比べて劣っていることを知らせるのが目的である場合に

は，その得点の意味づけつまり評価は低くなり，よく頑張ったことを知らせることが目的であればその評価は高くなる．

このように，何を何の目的のために評価するかにより，資料の収集方法，評価の仕方，意思決定等も異なってくる．したがって，評価にあたっては，評価の目的を明確にすることが必要である．Cronbach (1963) は，評価の目的を，(1) 教授方法や教材の検討のため，(2) 個人に関する決定のため，(3) 教育行政上の決定のため，というように三つに分類している．また，東 (1979) は，これをさらに細かく分類し，(1) 教育行政の資料としての評価 (2) 学校の管理・運営の資料としての評価 (3) 教師の学習指導の資料としての評価 (4) 子供に情報を与えるための評価 (5) 親の参考にするための評価 (6) 子供の処遇決定のための評価 (7) カリキュラムの改善のための評価 と分類している．この他 (Nevo, 1983) にも，いろいろな分類がみられるが，要するに，教育活動にかかわるあらゆることが評価の対象になるのである．むろん対象により，非常に長期間にわたってはじめて評価できるものと，比較的短期間に評価できるものとがある．以下では，学習指導場面における評価を中心に議論を展開する．

Ⅱ．学習指導における評価とテスト

1．診断的評価，形成的評価，総括的評価

学校教育の目的は，「人格の完成をめざし，平和的な国家及び社会の形成者として，真理と正義を愛し，個人の価値をたっとび，勤労と責任を重んじ，自主的精神に充ちた心身ともに健康な国民の育成を期して行われなければならない．」(教育基本法第一条)，と記されている．この目的を達成するためには，そのために必要な身につけるべき事柄を分析し，それにもとずいて学校における教育の諸目標が設定される．例えば，日常生活に必要な国語を正しく理解し，使用できる能力を養うこと，日常生活に必要な数量的関係を正しく理解する能力を養うこと等が目標となる．次に，数量

的関係を理解する能力を表す下位概念を設定し，それらをさらに下位の目標に細分割し，それをまたさらに下位の目標に分け，より具体的な形で目標を設定していく．また，国語を正しく理解する能力には，読解力を身につけるという目標がその一つに含まれるであろう．読解力は，単語を操作する能力，文章に示された通りの行動をする能力，書かれたものを読みそれに情緒的な反応をする能力等に分割される．これらの目標を達成するためには，多くの下位の目標を達成しなければならない．例えば，文字を正確に読めるようにするという下位目標が設定される．このように，より抽象的な上位の目標は，より具体的な多くの下位目標に分けられ，全体として教育目標は階層構造をなしている（図１）．そして，いずれかの目標を達成するべく，教授活動が開始されることになる．一般には，学習が進むにつれて，より抽象的な上位の目標へと移行する．

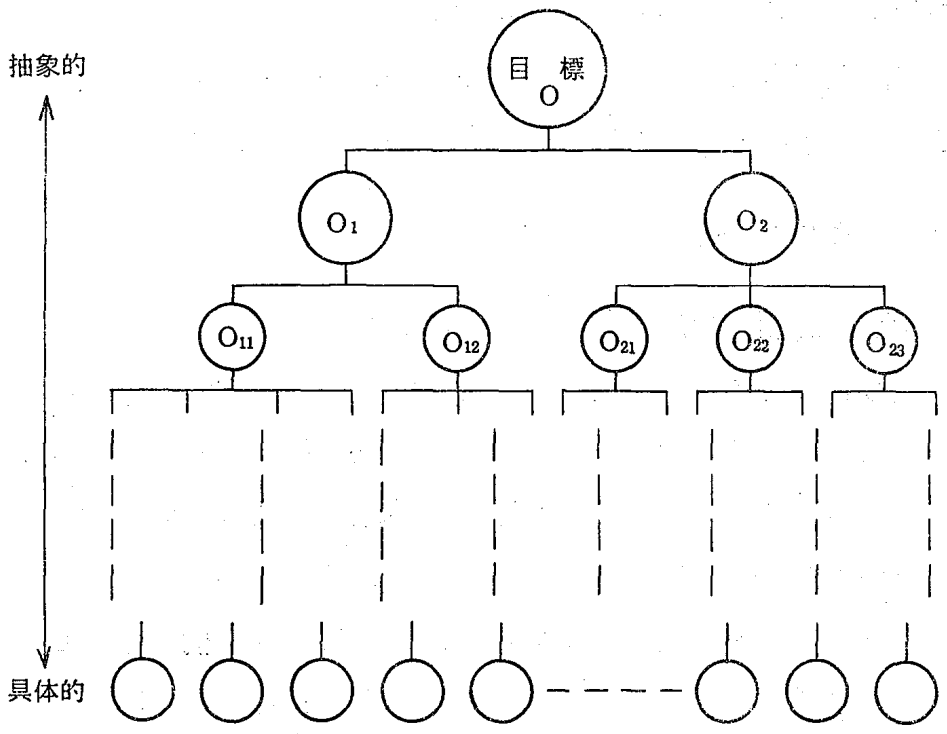

図 1　教育目標の階層構造
O：目標 (objectives)

学校教育における学習指導と教育評価を，一連の流れとして表現すれば，図2のようになる．まず，教授活動開始前に，これから教えようと意図している内容について，学習者がどの程度の知識をすでにもっているかをとらえるために，診断的評価 (diagnostic evaluation) を行う．この評価結果にもとずき，一定の教授方法で授業を開始する．指導を進めていく途中で，学習者一人一人が授業内容をどの程度理解しているか，もし理解していなければ，どういうところでつまずいたのか，それは下位目標の設定のしかたに問題があったのか，教師の教えかたに問題があったのか，診断的評価自体に誤りがあったのか等を評価する．指導の途中で行うこの形成的評価 (formative evaluation) 結果は，教師自身に，また必要に応じて生徒にもフィードバックされる．教師はこれにもとづき指導方法の軌道修正をし，より効果的な指導が可能となる．教授過程は，諸要因が関連しあって形成されていく動的なものである．絶えず形成的評価を行って，その結果を教授過程にフィードバックし，より良い教授過程を形成しながら，授業を進めていく．このようにして，いくつかの下位目標を達成したら，学期あるいは学年の終わりに，より上位の目標までを含めて，それらを達成したかどうか等まとめの評価として総括的評価 (summative evaluation) を行う．したがって，特に総括的評価は目先の目標ばかりにとらわれず，より広い能力や学力の獲得の有無を調べるように心がけることが望ましい．総括的評価は，その結果が次学期あるいは次年度の学習指導にフィー

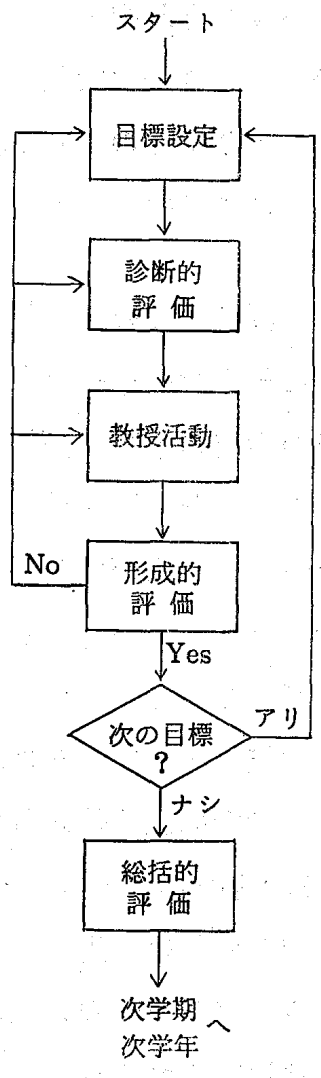

図 2　学習指導過程における評価

ドバックされ得るという意味では，形成的評価と本質的な差異はないが，後者はより抽象的な上位の目標までも含めたまとめの評価であり，前者はそれより下位のより具体的な目標についての評価であるという点を強調して，ここでは一応区別しておく．

2. 評価のための資料収集法

これまでに述べたさまざまな評価をするために必要な資料は，日常の授業の中での観察や面接およびテスト等によって得ることができる．特にテストは，非常に有効な資料収集法の一つである．この資料を或る基準にしたがって意味づけし解釈した結果が評価である．この評価結果を教育の場面にフィードバックし，教授過程に生かして初めて，テストすることの意味がでてくる．

知能テストは，20世紀の始めにビネー (Binet) が作成してから今日に至るまで実に広く使われてきたが (Binet & Simon, 中野他訳, 1982 を参照)，他の心理テストと同様に，個人差 (between-individual differences) の測定のために用いられてきた．教育の場面でも同様で，選抜，臨床的診断カテゴリーへの分類など，或る特性の上で個人を順位づける目的でテストは用いられてきた．しかし，教育の場面においては，教育的な働きかけにより個人がどの程度成長したか，最終目標にどの程度近づいたかを知ることが重要である．この個人内の成長 (within-individual growth) の測定 (すなわち教育的働きかけの効果の測定) に重きをおくテストは，個人差測定のためのテストとはおのずからその性質も，また，テスト結果の解釈のしかたも異なる．選抜のためのテストは，その目的からみて，教育的な働きかけの効果より，むしろ個人差を測定するように作成される．Carver (1974) は，個人差測定のためのテストを psychometric test，個人内の変化の測定のためのテストを edumetric test と呼んでいる．教育は，或る一つの目標に向かって個人がどのように変化していくかの過程であり，こ

の変化しながら形成されていく過程は，edumetric な測定が適していよう．プログラム学習において，目標行動がどこまで形成されてきたかを測定する場合等も edumetric な測定がふさわしい．しかし，或る時点での測定は psychometric である．したがって，教育の場面では，よりダイナミックな色彩の濃い前者に重きを置きつつ，後者も考慮しなければならない．

3．テスト得点の判断基準

テスト結果の評価方法には，集団準拠評価 (norm-referenced evaluation) および目標準拠評価 (criterion-referenced evaluation) が区別される．前者は，テストを受けた集団，あるいは標準集団の中での相対的評価である（個人内での相対評価もある）．したがって，人に比べて国語は優れているが，数学は劣るというような個人内のプロフィールはわかるが，テストが測定しようとしている目標のどこまで到達したかというような内容に関する情報は与えてくれない．したがって，個人差を明確にする選抜やふりわけのためには有効であるが，カリキュラムの改善や，教授過程の軌道修正のためにはあまり有効な評価方法ではない．

一方，目標準拠評価は，設定された目標が基準であり，学習者がこれにどの程度達したかにより評価がおこなわれる．学習指導の場面では，全生徒が目標に到達することが望ましいが，その場合にはむろん評価はすべて等しくなる．個人差を強調する選抜やふりわけの場面とは対照的である．教育の場面では，集団準拠評価から目標準拠評価への推移がみられ，後者の重要性がますます強調されてきている (Biggs & Collis, 1982; Bloom, 1956; Glaser & Nitko, 1971; Haertel, 1985)．この評価を行うためには，目標が明確にされていることが必要であるが，各教科で実際に目標をどのようにたてればよいかに関しては，例えば Bloom (1956) の教育目標の分類体系などがある．

すでに述べたように，学習指導の過程における評価は，何がどのようにわかったのかを評価することが重要であり，目標準拠評価がふさわしい．したがって，以下では，このテスト（目標準拠テスト）の妥当性について述べる．

Ⅲ．テストの妥当性

1. 三種の妥当性

テストが備えているべき性質の一つに妥当性 (validity) がある．測定しようとしていることを，テストがとらえているかどうかが，テストの妥当性の問題である．

測定したいことが明確にされているとき，テスト項目がそれを満たしていれば，そのテストは内容的妥当性 (content validity) があるといわれる．また，テストは，入学試験等のように，何らかの将来の行動を予測するためにしばしば実施される．この場合は，テスト得点が将来の行動をよく予測していれば，そのテストは予測的妥当性 (predictive validity) があるといわれる．内容的妥当性および予測的妥当性は，テストによる測定対象が具体的に明確に表されているか，または外部基準が存在している場合に確かめうる妥当性である．ところが，外部基準もなく，また，測定したい内容が学力や知能等のように抽象的な特性や属性である場合の妥当性は，別の方法で確かめなければならない．知能を例にとると，知能テストの妥当性は，知能テスト得点と，いわゆる知能といわれるものにおける個人の成績との相関により確かめられる．ところで，いわゆる知能とはわれわれが構成した概念である．概念構成体としての知能を測定するテストの妥当性は，この概念を表す他のいろいろな実際の行動における成績と，知能テスト得点との関係から検討される．パーソナリティ・テストで，“寛大さ”を測定しようという場合，不安テストで“不安”を測定しようという場合等のテストの妥当性も，同様に検討される．これが，概念的妥当性 (con-

struct validity) である.

2. 概念的妥当性

知能，学力，不安，動機づけ等は抽象的なものであり，直接観察することは不可能である．これらは，われわれが作りあげた概念であり，概念構成体 (constructs) と呼ばれる．いま，学力の中の数学の学力という概念構成体を測定するテストの妥当性を考えてみる．ただし，この学力テストは，可能なら将来の行動を予測するためにも使われるが，ここでは，学習

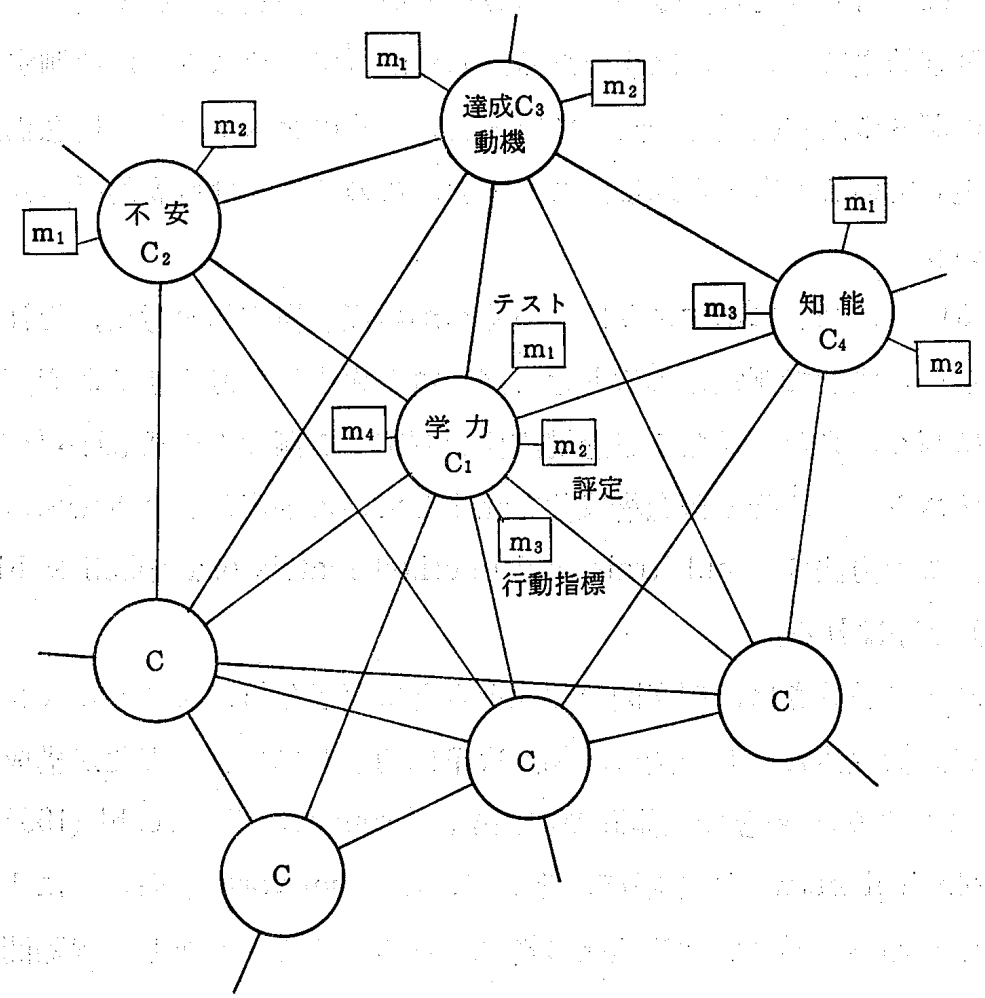

図 3　nomological network
C：概念構成体 (construct)
m：測度 (measure)

指導の過程でどの程度の学力が身についたかという，到達度を測定することを第一目的とするテストとする．数学の学力に対応する明確な内容，実体，あるいは行動は存在しないため，内容的妥当性は扱えない．

いま，数学の学力という概念構成体を測定するものに，テスト，教師による評定，その他の行動指標等があるとする．それらを，m_1，m_2，m_3,……とする（図3）．いま，テスト m_1 が妥当であるためには，同一概念構成体を測定する他の測度 m_2，m_3,……による測定結果と，テスト m_1 による測定結果が一致して（同じ傾向を示して）いなければならない．逆に言えば，もし全ての測度が同一の概念構成体を測定しているなら，すべての測定結果は一致するはずである．したがって，テスト m_1 の測定結果が他の測度の測定結果と一致しているとき，そのテストはその概念構成体を測定していると考えられる．このとき，このテストは概念的妥当性があるという．

あるいは，図3に示すように，いくつかの概念構成体を考え，それぞれが複数個の測度で測定されるとき，同一概念構成体を測定する測度間の一致性は高く，また，異なる概念構成体を測定する測度との相関が低ければ，テスト m_1 はその概念構成体を測定している妥当なテストであると考える．この方法は，multitrait multimethod matrix (Campbell & Fiske, 1959) と呼ばれる．

これらの手続きは，測度間の相関をもとに因子分析によりおこなわれるが，概念構成体どうしの間の関係も次第に明らかにされ，理論が構成される．このような概念構成体間の関係を，Cronbach & Meehl (1955) は nomological network と呼び，そして，この network を明らかにすることがすなわち科学における理論の構成であるとする．しかし，実際問題として，一つの概念構成体についての一つの測度によるデータの収集も容易なことではなく，nomological network 全体は，いく世代にもわたって長い時間の流れの中で完成されていくものである．

概念構成体によっては，テストという一つの測度しかない場合もある．このような場合には，概念構成体に影響を及ぼすような状況の変化を加え，その時の測定結果とそうでない場合の測定結果とを比較し，測定結果が期待される方向に動くことで妥当性を確かめる．しかし，例えば“不安”という概念のように，状況の変化の影響を受けやすいものは，実験的に操作することが可能であるが，“学力”は，長期にわたって変化するものであるにせよ，その時点では比較的安定していて，状況の変化の影響はあまり受けないように思われる．

概念的妥当性を調べる方法については，上に述べたが，Thorndike(1982)はこれらをまとめ，四つの具体的な方法を提唱している．(1) 概念とテスト課題の性質とを比較判断することにより確かめる方法．(2) テスト課題と，その概念を反映すると信じられている他のテストあるいは種々の事象との相関によって確かめる方法．これは，別の概念を測定するテストを作り，テスト得点にもとずき，テスト間の相関係数を計算し，これを因子分析する．同一概念を表すと考えられているテスト課題は，同一因子にたいする因子負荷量が高くなることにより確かめる方法．(3) その概念で異なっていると考えられるグループ間（被験者集団）で，テスト得点を比較することにより確かめる方法．(4) テストで測定しようとしている概念は比較的安定した特性である，という考えから確かめる方法．もし，テスト実施条件の変化などにより，テスト得点が著しく変化するようであれば，テストでねらっているのは安定した特性ではないので，概念的妥当性があるとはいえない．いずれにしても，相関をとって調べる方法は，高い相関が得られれば，テストの概念的妥当性は認められるが，相関が低い場合は，テストに問題があるのか，基準としてとったテストあるいは行動なりが，いま問題とされている概念を反映していると考えたのが間違っていたのかのいずれである．これらの中で(2)～(4)の方法は，Cronbach の提唱した nomological network の考え方の上になりたつ方法である．

3. 目標準拠テストの妥当性

集団準拠テスト（テストの結果を集団に準拠して評価するテストという意味）が妥当性をもつためには，このテスト結果を何のために使うかによるが，テストで高い評価を受けた者は，評価の低かった者より，だいたいにおいて，より有能である，より成功する，或る特性をより多くもっている，というように，テスト得点と，テストが測定しようとしているものを表す外部基準との相関が高ければよい．このことは，集団準拠テストは，もともと一つの次元上に，個人を順位づけることが目的であることを考えれば明らかである．したがって，予測的妥当性があればよいことになるが，この妥当性は比較的容易に確かめることができる．ところが，目標準拠テスト，特に抽象度の高い目標への到達度を測定するテストの妥当性を調べるのは，それ程容易ではない．

学習指導過程において，或る特定の教科の，或る特定の具体的目標（図1の下の方の目標）に，生徒が到達しているかどうかをみる形成的評価では，教科の内容が具体的に示されるので，内容的妥当性が確かめられる．例えば，算数で，くり下がりのある2桁の引き算のテストが妥当であるかどうかは，テスト項目がすべて，くり下がりのある2桁の引き算であるかどうかを確かめればよい．しかし，もう少し目標を上位にして，数量的関係を理解する能力（これには，くり下がりのある2桁の引き算の理解も含まれるであろう）を育てるという目標を設定し，この目標への到達度を調べるためのテストの妥当性を考えてみる．数量的関係を理解するという目標は非常に抽象的であり，この能力は概念構成体としてしかとらえることができない．そこで，この概念を測定すると考えられるテストを作成し，このテスト得点と，同一概念を測定する他の測度（例えば異なる量の水を入れたいくつかのビーカー等の材料をあたえ，指示通りの行動をさせる等）による結果との相関により，その妥当性を調べる．

Cronbach ら (1972) は，目標準拠テストの妥当性を以下のように考え

る．目標にあったテストをたくさん作成することが可能であり，またこれらのテストを，いろいろ異なる機会に，実施方法もいろいろに変えて実施することができる．したがって，一つの目標に関して無数に多くのテスト得点を考えることができる．この可能なテスト得点の集合を想定し，これを universe とする．この universe を構成する得点の平均値を，universe score と呼ぶ．一回のテストによる得点は，この universe からの一つの標本であると考え，したがって，テスト得点と目標を関係づけるということは，テスト得点をこの universe score に関係づけるということになる．これが可能であるためには，すなわち概念的妥当性を調べるためには，テストと目標を関係づける理論が必要である．Cronbach (1972) は，これを generalizability 理論と呼び，universe score の分散と，テスト得点の分散の比をとり，概念的妥当性について検討している．

形成的評価のために実施される目標準拠テストは，具体的な教科内容についての測定である場合には，そのなかに終局的な目標が直接的に表れていることはむしろ稀である．この場合には，内容的妥当性が確かめられ，さしあたり問題はない．しかし，形成的評価，総括的評価のための目標準拠テストの背後には，一つの大きな教育目標，言い換えれば期待される教育の成果が存在する．この目標の内容はわれわれが構成した概念である．そして，その他の目標準拠テストが測定する諸目標は，多くの目標を概念とする nomological network の中で，教育の終局の目標であるこの概念，および他の諸概念（不安，達成動機等）とも有機的に関連づけられていることが望ましい．この観点にたつと，学習指導において用いられるテストは，本来概念的妥当性を満たしていることが大切である．教育の成果の評価は，この network の中で多次元的になされなければならない．評価をする人は，目前のテストと潜在的な学力，知能等の諸特性との関連を心に描きながら評価することが大切である．目標準拠テストは，目標が抽象的

になればなるほど，概念的妥当性を満たしていることが要求されよう．また，テスト項目を越えて一般化しようという意図のもとに作成されるテストは，概念的妥当性を備えていなければならない．この場合，標本としてのテスト得点から，テストが測定していると仮定される目標（母集団）への到達度に関して推論することになる．それが可能なためには，例えば，上述の generalizability 理論のような理論が必要である．

計算能力，漢字書き取り能力等，特に skill の上達だけを目標にしていると，その skill は上達するが，そのさらに上位概念である例えば数学的なものの考え方が育たなかったり，あるいは，数学に対する興味を失ってしまったりすることがある．数学の学力という概念を表す階層構造のなかに，現在の目標を位置づけ，たえず終局の目標を意識しながら学習指導をしていくことにより，このような問題を解決することが可能であろう．行動分析的な教授方法をとると，skill の訓練ばかりになり，結晶性知能 g_c (Cattell, 1963) は発達するが，いろいろな場面に柔軟に適応することを可能にする流動性知能 g_f は育たないという実証的研究がある (Snow & Yalow, 1982)．いずれも知能ではあるが，知能の二つの側面（硬い頭と柔らかい頭）であり，g_f を犠牲にして g_c を育てることは片手落ちといわざるを得ない．Snow & Yalow の研究は，知能と学力という二つの概念構成体間の関係を nomological network の中でとらえた示唆に富む研究である．skill の上達だけでなく，いろいろな可能性を自由に考えられるような，指導および評価が必要であろう．

むすび

学習指導の流れの中での評価は，教育目標の階層構造（図1）および教育に深く関連するその他の諸概念との nomological network の中に，目下問題にしている目標を位置づけて行うようにするのが望ましい．また，

目標の習得度を測定する目標準拠テストは，この意味で概念的妥当性を備えていることが望まれる．

最後に，テストに関連して，一言ふれておきたい．客観テストばかりやっているので考える力がつかない，という批判を最近よく耳にする．客観テストと自由解答形式テストの分類は，後者はその名のとおりテスト形式を表したものであるが，前者を客観的というのは，採点が客観的に行えるようになっており，誰が採点しても同じ結果が得られる形式のテストだからである．普通は，多肢選択肢が用意されその中から一つを選ぶ，○×をつける，関係のあるものどうしを線で結ぶ等，すでに用意されている項目のなかから正解を選ぶというテスト形式になる．したがって，偶然に正解となることもあり，このようなテストばかりしていると……という前述の批判となる．確かに，二つの形式では，テスト項目に解答する際に関与する心理的過程，また測定される学力や能力の側面が異なっているという点では，上記の批判は正しい．しかし，測定手段としてのテストは，慎重な作成過程を経ることにより，評価の目的にあった有効な資料を提供してくれる．テストは学習を支えるものであるが，テストすることすなわち学習ではなく，思考力，文章力，理解力，推理力などは，テストとは別に日常の学習の過程の中で習得すべきものである．テストに対する前述の批判が我が国で起こるのは，日常の授業の中にまで，選抜のための評価が深く浸透していることを物語っている．選抜のための評価ばかりがはばをきかす教室の授業になってしまったら，他の人よりも少しでも高いテスト得点をとることだけが目的となり，そういうことには意欲のない生徒は，本来の学習意欲までも失ってしまうということにもなる．

また，運転免許試験，医師国家試験，その他種々の国家試験等は目標準拠評価であり，これらのテストは，内容的妥当性はもちろん，或る程度の予測的妥当性も備えていることが必要である．これと同様に，大学における教科の科目試験も，内容的妥当性ばかりでなく．それと関連するであろ

う分野における将来の行動も予測しうる方向にもっていくことが望まれる.

もともとテストの信頼性や妥当性は，客観テストを開発（作成）する上で重要視されたのであり，自由解答形式テストの場合は，テストの妥当性よりむしろ，自由に述べられた解答（反応）を，いかに得点化するかの方が問題となる．反応を質的に評価するための一つの基準として，SOLO taxonomy (Biggs & Collis, 1982) などは有効である.

（この論文をまとめるにあたり，慶應義塾大学文学部教授，並木博氏に貴重な御示唆をいただきました．ここに記して感謝の意を表します）

References

東　洋　子どもの能力と教育評価．東京大学出版会，1979.

Biggs, J.B. & Collis, K.F. Evaluating the quality of learning. The SOLO taxonomy. Academic Press, 1982.

ビネー，A. & シモン，Th. 中野善達，大沢正子訳，知能の発達と評価．福村出版，1982.

Bloom, B.S. (Ed.) Taxonomy of educational objectives. New York: Longmans, Green, & Co. 1956.

Campbell, D.T. & Fiske, D.W. Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Blletin*, 1959, 81–105.

Carver, R.P. Two dimensions of tests; psychometric and edumetric. *American Psychologist*, 1974, 29, 512–518.

Cattell, R.B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 1963, 54, 1–22.

Cronbach, L.J. Course improvement through evaluation. *Teachers College Record*, 1963, 64, 672–683.

Cronbach, L.J. & Meehl, P.E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281–302.

Cronbach, L.J. & Snow, R.E. Aptitudes and instructional methods. Irvington, 1977.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependabilty of behavioral measurements. Wiley, 1972.

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In Thorndike R. L. (Ed.) Educational measurement. American Council on Education, 1971.

Haertel, E. Construct validity and criterion-referenced testing. *Review of Educational Research*, 1985, 55, 23-46.

Nevo, D. The conceptualization of educational evaluation: an analytical review of the literature. *Review of Educational Research*, 1983, 53, 117-128.

Nunnally, J. Psychometric theory. McGraw Hill, 1967.

Snow, R. E. & Yalow, E. Education and intelligence. In Sternberg, R. J. (Ed.) Handbook of human intelligence. Cambridge university press, 1982.

Thorndike, R. L. Applied psychometrics. Houghton Mifflin Company, Boston, 1982.

van der Linden, W. J. A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research*, 1981, 51, 379-402.