「学習者自身による客観テストの作成と評価の試み」

MULTIPLE-CHOICE TEST MADE BY LEARNERS

栗山　容子

学習評価の主要な機能の１つはいうまでもなく学習者が自己の学習上の弱点がどこにあるかを明らかにし，学習を進めていく上での指南役を果すことである。

大学における成績評価については，評価内容が，思考方法やその内容，論理的な推論や判断など，高次の精神過程を含むところから，現状ではこの機能が充分に果されているとは言い難い。評価の基準が明確に示されることは少なく，学生達にはベールに包まれた作業と受けとめられている。大学生の学習評価についての諸研究が極端に少ないのも，ここに１つの理由があるだろう。大学における教育目標の観点からは，レポートや日常の学習態度，あるいは論文テストによる評価が適切と考えられるが，今日のように学生数が増えてくると利点よりも欠点が目についてくる。多量のレポートを読み，いくつかの評価基準と照合しながら総合的判断，評価を下していく作業は教師にとって責務とはいえ，苦痛を伴なってくる。

筆者は受講生の多い基礎コースにおいて，客観テストの利点を生かし，学習者自身にも問題の作成に参加してもらい，一種の自己評価として評価活動の一部に加わるという方法で学習後の成績評価を行なってきた。後に述べるように問題を公開しないことを原則としてきたが，'83年度より担当コースに変更が生じたので，これを機会に，これまでの実践記録をまとめ，報告す

る。

Ⅰ．方　法

⑴対象：本学教育学科学生。一部他学科生も含まれる。２年次学生と一部の３，４年次学生。人数は'80年度（冬学期）――70名，'81年度（春学期）――58名，'82年度（春学期）――56名。（但し，コンピュータで処理できた者の人数のみ）

⑵コースとその内容：「測定と評価　Measurement and Evaluation」教育学科の基礎科目の一つ。学習内容は Ebel. R.L（1979）を参考に次の12のカテゴリーとした。1. 評価の歴史　2. 評価の諸理論　3. 評価の目的　4. 評価の方法　5. 評価用具　6. 評価用具の具備すべき条件　7. 客観テストと論文テスト　8. 絶対評価と相対評価　9. 形成的評価と総括的評価　10.標準テスト　11.資料の整理（記述統計学の基礎と応用）12. その他の特殊問題。

⑶授業方法：講義形式で適宜，質疑応答を行なう。初等統計学のみ，テキストを使用し，随時参考文献の紹介。OHP使用。第１回目のクラスでオリエンテーションを行ない，授業内容，日程，評価方法について説明。１回70分。全27回前後。

⑷評価方法：４回の初等統計の課題提出を最低の条件として，中間テスト（初等統計問題）及び学期末テスト（学生自身による問題作成）の２つのテスト得点で相対的に評価，学期末テストについては，コースの終了する10日前にテスト問題の作成方法について教示し，５日後にテスト項目案を提出させた。コースの最終日にテストを実施し，（正味60分）学期末テスト期間中の割りあてられた日時に，結果についてのフィードバック及びまとめを行なった。

⑸テスト項目の作成：1）設問の領域について，B.S.Bloomら（1971）の教育目標のtaxonomyを参考にして，学習内容と行動目標の二次元マトリックス上に整理した。行動目標についてはBloomらが６つのカテゴリー――知

識，理解，応用，分析，総合，評価――について具体的に検討を行なっている。また Ebelはこれを７つのカテゴリーに分けている。ここでは問題作成上，枠組が理解されやすく，また明確になるように考慮して，次の４つのカテゴリーに分ける[*1]。A.基本用語，概念の知識　B.事実や原理の理解　C. 応用　D. 総合。前述の学習の内容とこの行動目標によって得られるマトリックスの各セル（分類記号で区別）からテスト項目を抽出する。これは出題範囲が片寄らないようにし，テストの妥当性を高めるためである。2) 形式は多肢選択テスト（４肢選択）。形式を整えたのは，作問の具体例を示しやすく，作問上の諸注意を具体的に示せるためである。3) 具体的方法；A４紙1／4大のカードを使用し，１カードにつき１問を作成する。各カードには分類記号，作成者氏名を記入する。１人３問以上６問以内作成すること，正答は選択肢の最初にもってくることを教示し，問題作成上の留意点を説明。以上により提出された問題項目を，内容領域，形式，問題の一義性などの観点から充分吟味し，必要に応じて最少限の加筆修正を行なった上で，適当と判断された項目を抽出した。これらの結果から重みづけを考慮して'80には筆者が作成した問題10問，'81,'82はそれぞれ前年度の項目分析（後述）の結果，良好と判断された項目を10問加え，テストを構成した[*2]。各年度の問題項目数は'80－69問，'81－50問，'82－70問である。

⑹評価のねらい；コースが基礎科目の１つであるため，基礎的な学習内容の理解の程度を知ることに主眼点をおいた。また学習者自身が作問に加わる過程で，学習内容を深め成績評価の過程の一端を実際に体験することによって教育評価の意義を確認することを本評価のねらいとした。

Ⅱ．分析方法と結果

⑴１問正解につき１点，誤答０点として各人の得点を算出した。得点の分布は図１～図３に示した。素点の取扱いは，T得点に換算する方法があるが，ここではグループ間の比較を行なう目的ではないので，素点のままとした。平均，標準偏差，及びテストの信頼性係数をK－R21[*3]で推定した結果を表１

に示した。

表 1. 年度別の平均，標準偏差及び信頼性係数

実施年度	項目数	平均(Range)	標準偏差	K−R21(α21)
'80	69	43.8 (26～59) 63.5 *	8.4	.819
'81	50	31.6 (7～42) 63.2 *	5.0	.628
'82	70	46.0 (28～57) 65.7 *	6.5	.701

* 100点満点とした時の換算値

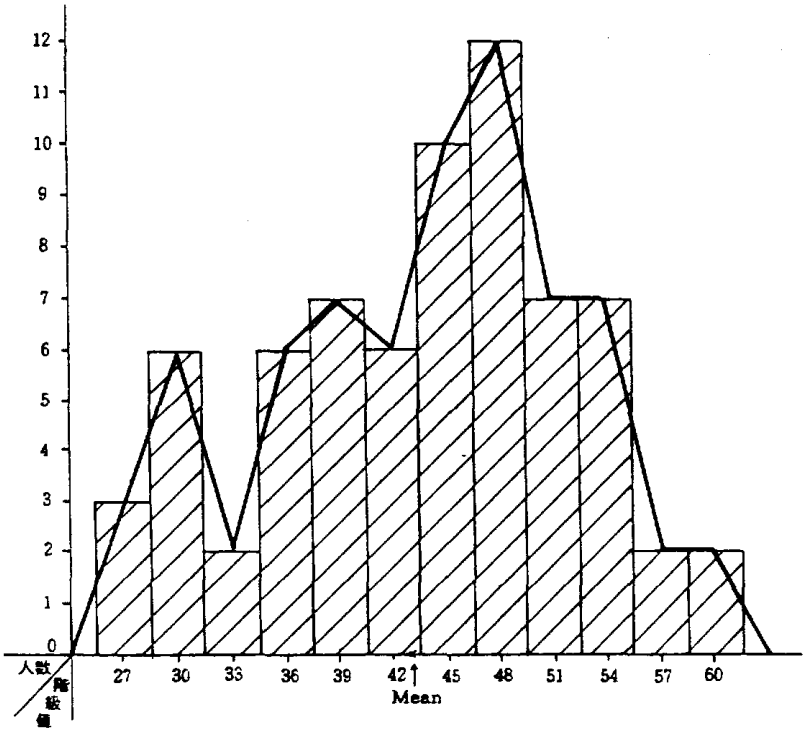

図 1. 素点分布（'80）

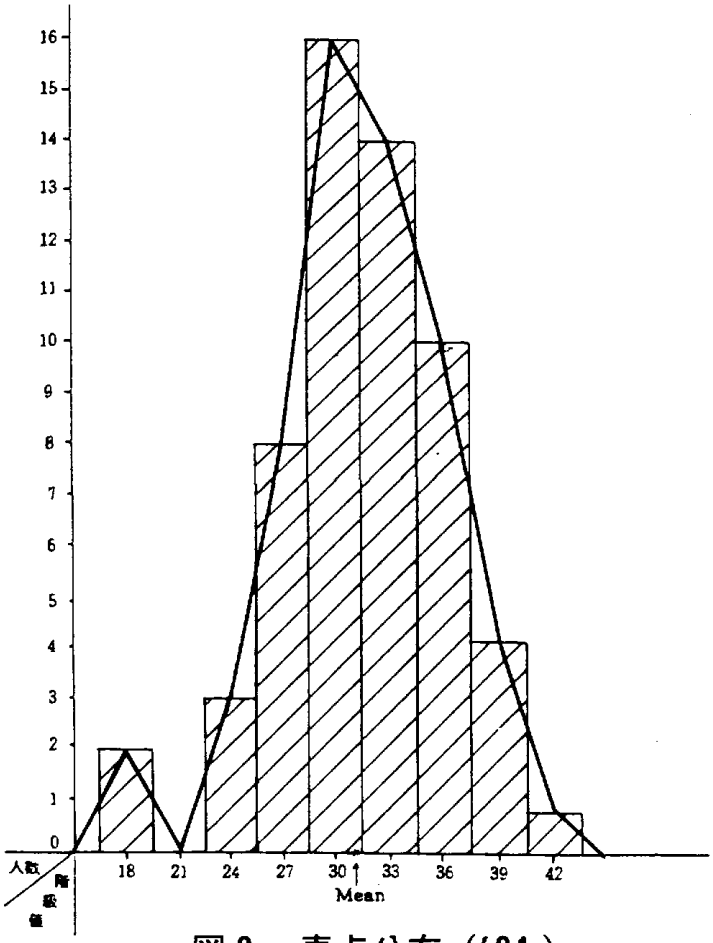

図 2.　素点分布（'81）

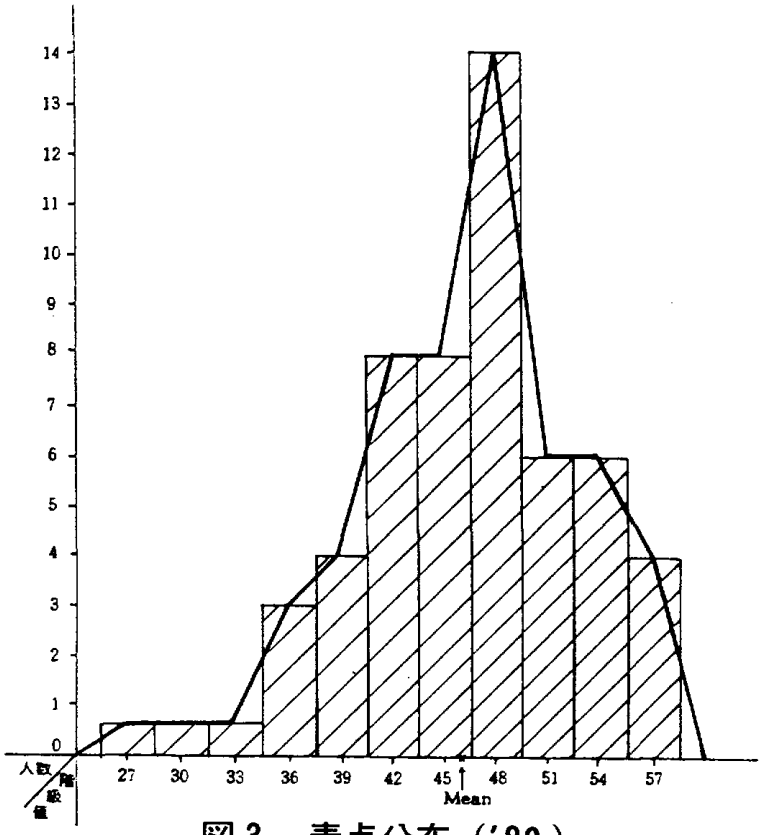

図 3.　素点分布（'82）

(2)各テスト項目について，正答率と弁別指数を算出し，テスト項目の良否を検討した。弁別指数は項目の識別力を示すもので，成績上位27%群の正答率と下位27%群の正答率の差による簡便な方法を用いた。マイクロ・コンピュータを用いて簡単に算出できるメリットがある。図4～6には水平座標に項目の難易度（正答率），垂直座標に識別力（弁別指数）として，全項目を2次元に表現した。

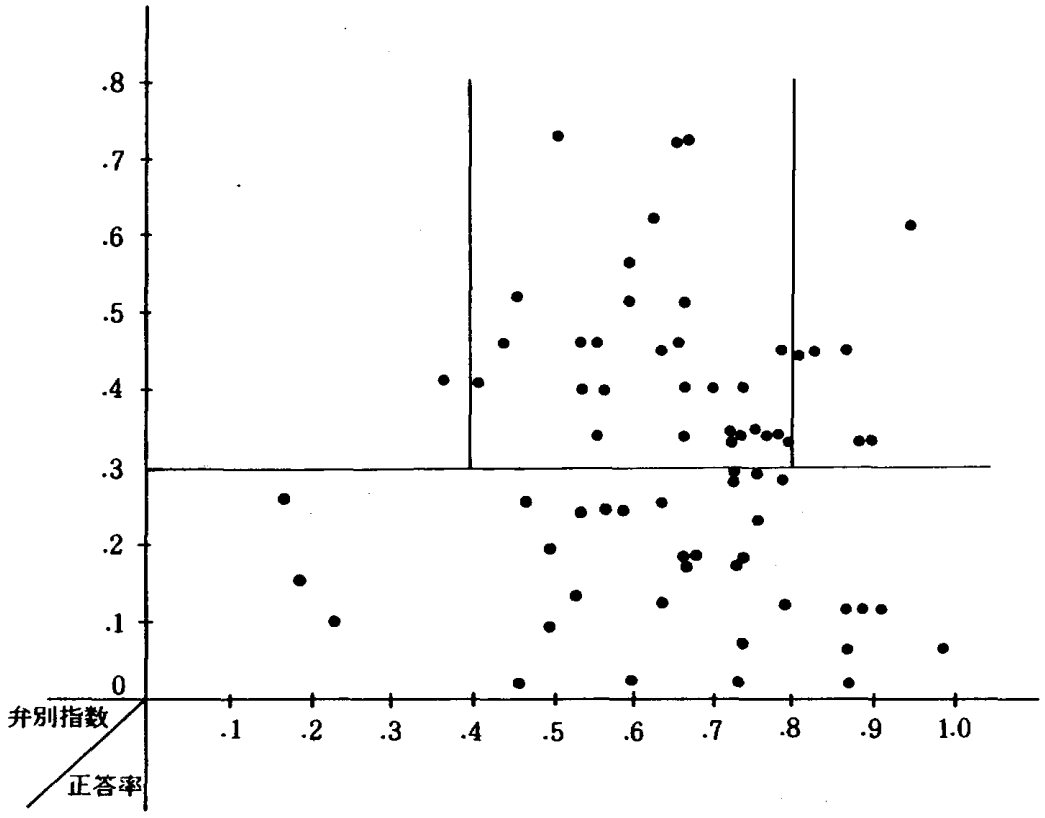

図4. 正答率と弁別指数（'80）

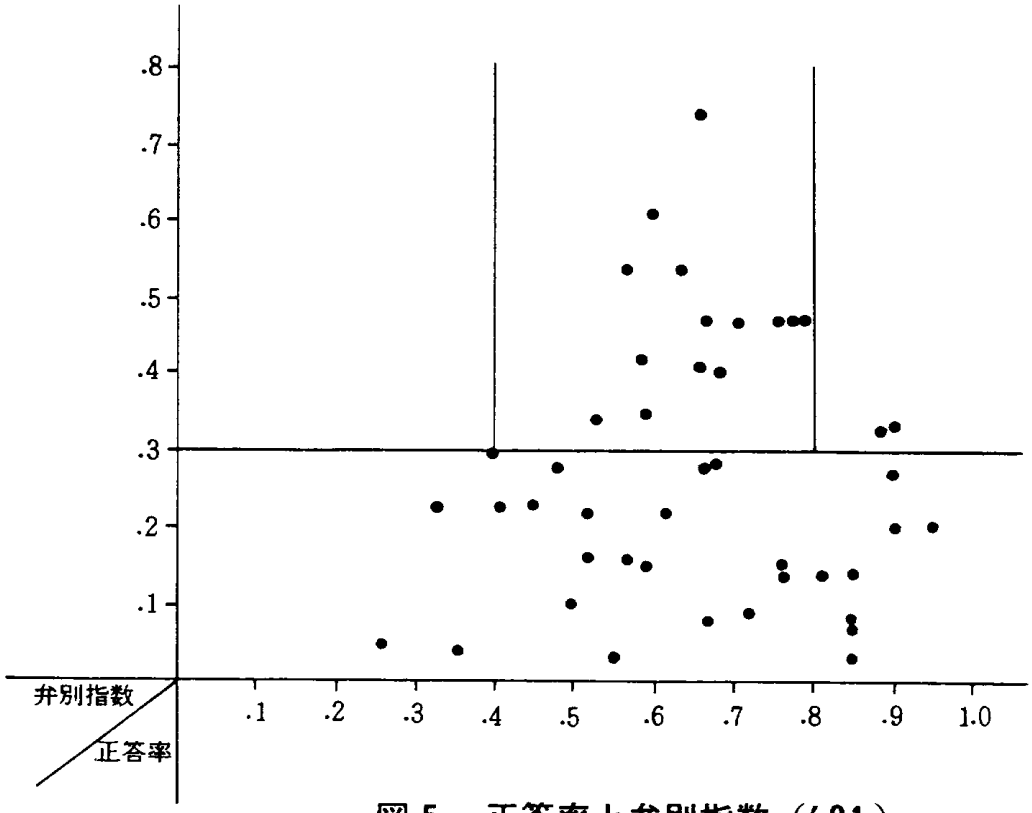

図 5.　正答率と弁別指数（'81）

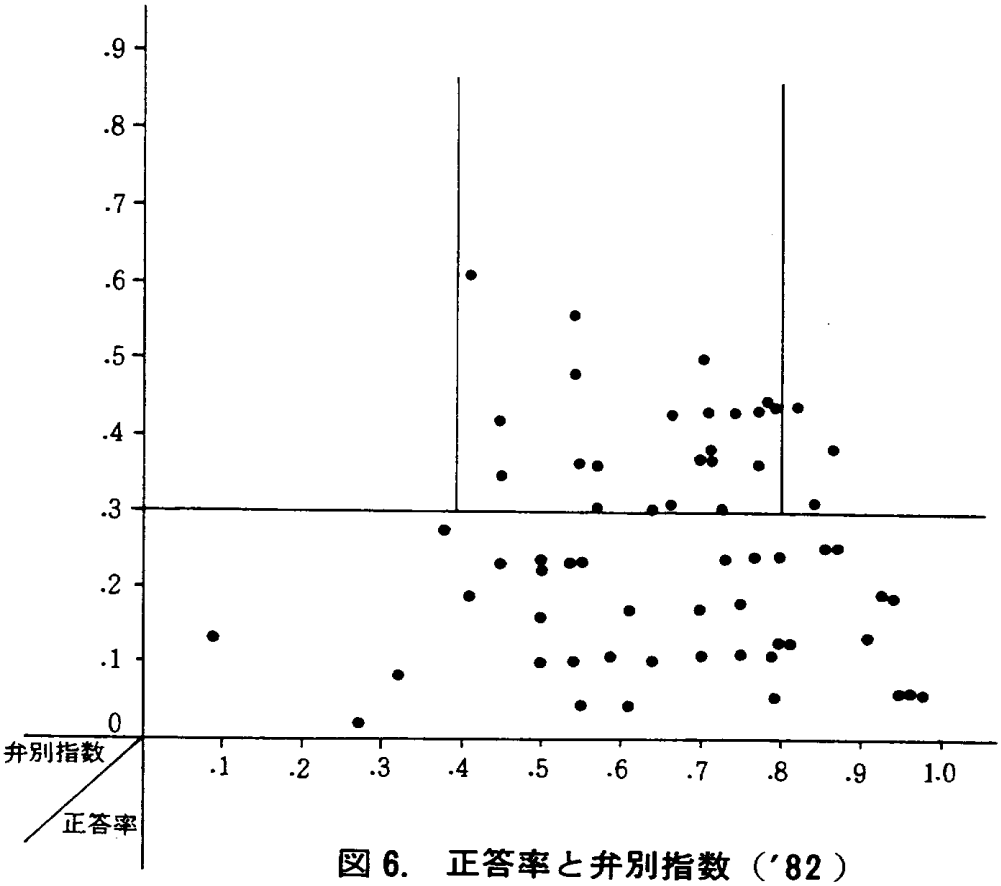

図 6.　正答率と弁別指数（'82）

(3)項目分析の結果，各テスト項目について，４つの選択肢（正解はそのうちの１つ）の選ばれ方の分析から適切と判断され，３ケ年にわたって使用された問題項目の具体例とその結果の一部を表２にまとめた。P.は正答率，d.は弁別指数を示す。

表２　良好と判断された問題例と'80年度の項目分析結果，及び３ケ年の正答率と弁別指数

（注）選択率は百分率で示してある
＊は正解

例１　A. T. I. は適性によって

a．最良の学習習得の方法が違う
b．学習習得の時間が異なる
c．学習の効果が予測できる
d．学習者の進路を決定できる

という考えに基づいている。

選択肢／選択率	＊a	b	c	d
上位27%	100.0	0.0	0.0	0.0
下位27%	27.8	11.1	33.3	22.2

	p.	d.
'80	.67	.72
'81	.78	.47
'82	.93	.19

例２　対象者が100～200人ぐらいで問題数が10～20問程度のテストでは次のどの方法を用いるのが最もよいか。

a．素点法
b．５段階法
c．偏差値法
d．指数法

選択肢／選択率	a	b	＊c	d
上位27%	15.8	5.3	78.9	0.0
下位27%	27.8	27.8	33.3	11.1

	p.	d.
'80	.56	.46
'81	.53	.34
'82	.45	.23

例3　信頼性の検証に用いられる4つの方法のうちで，実施の際に測定の独立性保持が問題となるのはどれか。

a．内部一貫性による方法
b．平行テスト法
c．折半テスト法
d．再テスト法

選択率＼選択肢	a	b	c	* d
上位27%	5.3	5.3	10.5	78.9
下位27%	27.8	27.8	22.2	16.7

	p．	d．
'80	.63	.62
'81	.50	.10
'82	.54	.48

例4　数学，国語，英語に関する実力テストについて，A，Bの得点は表のとおりであった。AとBは3科目，総合的にどのような評価をうけるか。

	A	B	クラス平均	標準偏差
数学	50	40	50	11
国語	32	45	40	5
英語	66	63	54	12

a．Aの方がBよりも高い評価を受ける。
b．Bの方がAよりも高い評価を受ける。
c．AとBは同じ程度の評価を受ける。
d．この表だけでは判断できない。

選択率＼選択肢	a	* b	c	d
上位27%	10.5	73.7	10.5	5.3
下位27%	11.1	27.8	38.9	22.2

	p．	d．
'80	.54	.46
'81	.85	.03
'82	.82	.44

例5　下の表はあるレディネステストの結果とその後の生徒の成績の相関表である。成績評価はVery GoodからPoorまでの4段階である。このレディネステストの特徴は何か。

成績＼テスト得点	20	30	40	50	60	70	
Very Good						22	7
Good				1	12	25	2
Fair			5	15	20	27	
Poor		3	3		8	2	

a．妥当性が高いといえる
b．信頼性が高いといえる
c．総括性があるといえる
d．何ともいえない

選択率＼選択肢	* a	b	c	d
上位27%	84.2	15.8	0.0	0.0
下位27%	11.1	27.2	7.1	12.9

	p.	d.
'80	.51	.73
'81	.52	.16
'82	.45	.42

Ⅲ．考察

1．問題項目、問題構成の検討

作成された問題の良否を前項の結果より考察する。問題の作成にあたって正答率は60%前後をねらうこと，但し，難易度に幅をもたせること（30～80%）によって問題に取組む動機づけを高める必要があることを事前に説明してあった。表1により平均値を各年度別にみると平均正答率はいずれの年度も，.60～.65の値となり，ねらい通りの結果が得られた。学習者自身が作成したテストの実施結果であるから平均正答率は高めとなることを予想していた。しかし，多数作成された項目の中から抽出された項目であるから，1人1人の作成者の問題は全問題数のごく一部であり，評価者作成の客観テストと同質であったことを結果は示している。またK－R21によるテストの信頼

性係数の算出結果では'81で.63, '82で.70となり'80の.82と比べやや低い。信頼性係数はその性質上，問題数によって影響を受ける。特に'81は全50問でこの種のテストとしては問題項目数が少ない。問題数をふやすことによって信頼性の高いテストに改善することができるだろう。

得点分布（図１～図３）でみると，'81は平均を交点として，２つのゆるい山形分布がみられる。人数が少ないので推測の域を出ないが学習が不十分で，また項目作成にも積極的でなかったものと学習に積極的だったものという異質の２グループが考えられ，学習の深度と関連していると思われる。'81の分布は分布の広がりが小さく尖度の大きい分布の形状である。問題数が少なく，得点範囲が狭いためである。'82はやや負に歪んだ分布であるが，客観テストの分布としては望ましい形状といえるだろう。

項目の識別力という点から項目の良否を検討するために，弁別指数を求めた結果（図４～図６）を見ると，正答率.4～.8の範囲でd.の値.3以上のものは'80で43%（30／69）'81には30%（15／50），'82には33%（23／70）である。すなわち各年度とも約1／3以上の項目が良い項目と判断された。しかし一方で，不当に正答率が低い項目や，逆に正答率が高すぎる易しい項目があり，これらは識別力も低くなっている。

表２に結果の一部が示してある各項目の４つの選択肢の選ばれ方の検討では，低得点群では４つの選択肢がほぼ同率で選ばれていること，すなわち，４つの選択肢はほぼ同じ程度に正答らしくなければならないが，中には全く選択されない選択肢をもつ項目があり，検討の余地を残していることが明らかとなった。またスモールステップの原則をとることによって問題数を増やし，誤答であった場合に，どのレベルでの理解のつまずきがあったのかを知ることができたのではないかと思われる項目もあった。たとえば表２の例４では偏差値を算出する問題項目と総合判断を求める問題に分けて設問する方が適切であった。

良い問題の一部は次年度に再度使用されたが表２にあげた５例のうち，例４の'81の弁別指数が小さくなっている他はいずれの年度でも安定した良い

項目となっていることがわかる。例4（'81）の場合には正答率が高いことを考えると，低得点群が良くできていたため弁別指数が低くなったのである。良い問題をプールして，適宜再使用することにより評価の絶対的な基準を定めていく可能性が認められた。

2．実践の意義及び問題点

学習者自身がテスト項目を作成し，構成されたテストを評価に用いるという試みの問題点をまとめる。(1)問題領域は広範囲にわたり，項目抽出の方法はテストの妥当性，信頼性を高めている一方で，行動目標のカテゴリーD.〝総合〟は，そのカテゴリーに入るものが少なく，せいぜいC.〝応用〟の範囲に収まってしまう。また，たとえそのカテゴリーに所属するものがあっても問題に一義性が保証されず，良い問題となりにくい。高次の学習内容をいかに問題構成するかが問題となる。(2)問題項目数は多い方が望ましいが一方で，実施上の困難もある。'81年度を例にとると，信頼性係数は.63であるから，これを.8まで高めるためには少くとも2.3倍の長さ，すなわち117問前後にしなければならない[*4]。これは，60分で実施するテストとしては問題数が多すぎる。時間延長を考慮するか，ある程度信頼性を犠牲にしても問題数を減らすかは実施条件を考慮すべきである。(3)この評価方法は項目のカテゴリー別抽出や，個々の項目の形式，字義などの修正，項目の独立性のチェック，正解の位置や項目順序の決定など，準備段階で労力と時間がかかる。(4)成績通知のための評価にどのように生かしていくかの検討が必要である。

では学習者自身がテストを作成することにどのような意義があるのだろうか。第一に，作問に加わる過程で，自らの学習を深めることができることである。テスト後に求めた感想で「問題を作るためには，内容の充分な理解が必要だった。」と学習の必要性を認めている。第二にテスト結果のフィードバックを充分に行なうことによって，測定と評価の諸問題を体験的に学習させることができる。自由記述で求めた感想では，この評価方式が大変好意的に受入れられていた。第三に，学習者自身によって作成されたテストでありながら，客観テストの性質を具備していることである。多数用意された項目

から抽出され，構成されたテストなので，１人の作成者の問題はそのごく一部となり，全体としては評価者により準備された客観テストと同じ性質のものとなっているのである。客観テストの利点は評価の基準が明確であり理解の程度をはっきり知ることができる。ここで作成されたテストもこのような利点をもっているといってよい。

このような試みが学習者の動機づけを高め，やる気を起こさせるに役立ったかどうかといった学習者に働く心理要因については，今後，組織的に検討していく必要があるだろう。また他の教科への一般化が可能かどうかも検討の余地があるだろう。

＊１　Bloomらは学習，評価目標を大，中，小，のカテゴリーに分類を試み，機能的に関連づけているが，そのためには，年月をかけてこのコースの教育目標に取組む必要があるだろう。

＊２　問題を非公開としてきた理由である。良い問題をプールしておき，そこからも問題項目を抽出していく方法をとった。

＊３

$$K\text{-}R21 = 1 - \frac{.8M(K-M)}{K\sigma^2}$$

但し，K　項目数
σ^2　分散
M　平均

＊４

$$K = \frac{R(1-r)}{r(1-R)}$$

但し，K　テストの長さ
R　求めたい信頼係数
r　テストにより得られた信頼性係数

参考文献

B.S.Bloom, J.T.Hastings and C.F.Madaus（eds.）

Handbook on Formative and Summative Evaluation of Student Learning McGraw-Hill 1970

R.L.Ebel

Essentials of Educational Measurement（Third Edition）Prentice Hall Inc. 1979

Summary

The process of assessing students' achievement in learning is very important and is much more complicated than it may seem to be. This is the report on university students' achievement as evaluated by a new system: an evaluation by student-made objective tests.

This treatment proved to be essentially satisfactory and beneficial as an educational procedure.

I. Method

Subjects were 184 ICU students: 70 in 1980, 58 in 1981, and 56 in 1982.

1. Course title: "Measurement and Evaluation" (foundation course in education)
2. Specifications of content: a) The essential problems of educational measurement and evaluation, b) An introduction to descriptive statistics.
3. Procedure for evaluation: Students were informed how to make test items five days before the last class.

Form of test items. Multiple-choice test (four branches) was designed.

Test construction. Test items prepared by student were refined, where necessary, by deletion and addition. Items were chosen from each two way grid assigned to 12 areas of contents × 4 ability categories. The number of test items are as follows: '80 – 69, '81 – 50, and '82 – 70

This procedure is useful for the improvement of test validity.

II. Results

The distribution of test scores and variances

The proportion of correct responses is '80 – 0.635 (mean), 8.4 (variance), '81 – 0.632, 5.1, and '82 – 0.657, 6.5. As might be expected, this was slightly easier than other ordinary objective tests.

Test reliability

The test reliabilities coefficient estimated by K-R 21 is '80 – 0.82, '81 – 0.63, and '82 – 0.70. The estimation in 1981 was slightly low.

The item analysis

To improve the items as a measuring instrument, two indeces are available; 1) the index of item difficulties, 2) the index of item discrimination. The index of difficulty is defined as the proportion of the group who do not answer the item correctly. The item discrimination is indicated by the difference in proportion of correct responses between upper and lower groups. Using the two indeces, the dispersion of difficulty values (range 0.4 to 0.8) and the higher discrimination index (higher than 0.3) for each test item were evaluated. As a result almost one-third of items in each year were found to be satisfactory.