段階反応モデルによる小論文データの解析

東京大学教育情報科学研究室　渡　部　　洋
東京大学教育情報科学研究室　平　井　洋　子

An Analysis of Essay Examination Data Using Graded Response Model

Hiroshi WATANABE/Yoko HIRAI

The purpose of this study is to investigate the reliability of essay test scores when the scores are expressed by using several rating categories and applied to graded response model. The data were obtained by assigning an essay examination to high school students. The 165 essays were evaluated holisticaly and given marks between 0 to 100 points. The marks were then classified into the 2 to 5 categories according to the two criterions : percentile and standard deviation. It was found that the reliability of the 5-category data (standard deviation criterion) was almost as good as that of the original 100-point data, which indicates validity of categorical rating. The characteristics of some of the raters were also described by their information functions.

目　　次

Ⅰ．はじめに

入学試験に小論文を課す大学は多い。国公立大学の2次試験をみても，小論文を採用する大学・学部の数は80%に達するといわれる。知識の測定を得意とする多肢選択式テストに比べ，小論文テストでは知識だけでなく，論理的思考力，文章表現力，問題意識，独創性なども測定されることが期待されている。小論文テストで測定される知識は主に能動的なものであり，受験者はそれを文章という形に組み立てて表現する。この点で小論文テストは実技試験に近い性格を持つといえよう。

小論文テストが測定している「書く能力」を，実際に文章を書かせるのでなく，客観式テストで間接的に測定できないか，という研究がある。Ackerman and Smith (1988) は，客観式テストで綴りの正確さ，句読点の正確さ，語の用法，段落の展開，段落の構成などの能力を測定した。また，Perkins, Pohlmann and Brutten (1988) は客観式テストでアナグラム，語を並びかえて文にする，文を並びかえて段落にする，段落を組み立てるという能力を測定した。しかしいずれの研究でも，客観式テストによる間接測定は，実際に論文を書かせる直接評価と測定内容が異なるので，代替はできないという結論が出ている。このことは，小論文テストは客観式テストにない独自の内容を測定していることを示している。

小論文テストの評価で常に問題になるのは，評価が主観に依存することによる評点の信頼性の問題であろう。この問題は，同じ評定者が同じ小論文を繰り返し採点するときの再評価信頼性と，異なる評定者による評点の一致度を表す評定者間信頼性に分けられる。渡部，他 (1988) は，3名の評定者が1週間おいてもう一度小論文を総合評価した場合，再評定信頼性が0.40～0.91と，評定者によってかなり異なることを見いだした。また，同じ研究で11名の評定者の総合評価間の相関係数を計算したところ，0.22～0.57と，これもまちまちの値となった。このことは，評定者によって異なる観点で評定が行われた可

能性を示唆する。

評定者をテスト項目に見立てた場合，テスト得点すなわち評点の，真の値どうしの相関係数が1ならば，その評定者たちをcongenericという(Jöreskog, 1971)。Blok (1985) は同じ評定者の繰り返し評定と異なる評定者による評点が，同じ真の値を持っているかどうかを検討し，16名の評定者に105の作文を2回評定させ，線形構造方程式モデルで分析した。その結果，同じ評定者による繰り返し測定はcongenericということができ，同じ真の値を反映していると見なせるが，異なる評定者による評点はcongenericではなく，本質的に異なる測定であることがわかった。

このように評定者によって異なる質の評定がされることを仮定して，評定者の効果を修正するモデルも提案されている。De Gruijter (1984) は加法モデルとRaschモデルを拡張した非線形モデルの2つのモデルを実際のデータにあてはめて検討し，非線形モデルが実際のデータによくあてはまるとしている。また，Houston, Raymond and Svec (1991) は，評価者ごとの評定の甘さ／辛さ，すなわち評定基準を最小2乗法，重みつき最小2乗法，EMアルゴリズムを用いた欠損値の補完の3通りの方法で修正し，無修正の場合と比較した。その結果，3つの方法とも真の値の推定精度が向上し，中でもEMアルゴリズムを用いた方法が最も優れていたことが示された。

De Gruijter (1984) の非線形モデルは，テスト項目を各評定者に見立てたRaschモデルの変形とみることができる。他に項目反応理論からの小論文データへのアプローチの可能性としては，部分的反応モデルや段階反応モデルが考えられる。たとえばDe Ayala, Dodd and Koch (1991) は部分的反応モデルを用いて書く能力の測定を行った。彼らは，1つの課題への評点として2名の評定者の評点の合計を用い，4つの課題をそのままテスト項目と見なしてモデルをあてはめた。この研究では，合計6段階の評点の順序性が一部逆転するという結果が出ており，モデルがきれいにあてはまったとはいいがたい。項目反応理論の立場からのアプローチは，現在のところあまり多くはなく，モデルの有効性に関する検討がさまざまなデータでなされることが必要である。

本研究は，粗評点を段階的なカテゴリに分けることで，各評定者の評定基準をある程度コントロールし，評定者を項目に見立てて1つの課題から被験者の能力を推定することを目指すものである。そのさい，項目反応理論の立場から段階反応モデル(Samejima, 1969)をあてはめ，近似的なテスト信頼性を求め，再検査信頼性や分散分析による信頼性と比較して，カテゴリ採点の有効性を検討する。さらに，各評定者の情報曲線を検討することで，各評定者の特徴も記述していく。

II. データ

A. 粗データ

1. 課　題

本研究で用いたデータは，渡部，他 (1988) のデータの一部である。

小論文の課題は「旅」という題で，都内の高校1年生（男子51名，女子49名）と近県の高校1年生（男子56名，女子29名)の，合計185名に，制限時間は特に設けず，字数800字前後で書かせた。評定は学校や塾で国語教育に携わっているもの5名（評定者1～4，6）と，それ以外の教師や大学院生6名（評定者5，7～11）の合計11名で行った。評点は100点満点で総合評価である。評定の系列効果が生じるのを防ぐため，各評定者に渡る小論文の順番はランダムに入れ替えてある。なお，以下の解析では最初の20名分の小論文を評価練習用とみなして除外したため，実際に用いたのは165名分のデータである。

2. 評点の分布

表1は11名の評定者による評点の，平均と標準偏差を示したものである。平均が47.0～72.3，標準偏差が4.7～15.7と評定者によって大きく異なっている。また，評点の分布の様子も評定者によって大きく異なる。このことを示したのが図1，図2である。図1は評定者1の評点の分布を幹葉表示で示したもので，連続的な分布となっている。一方図2は評定者9の評点分布だが，離散的ですでにカテゴリ分けされているような分布となっている。しかし，11名の評定者による評点の主因子解の第1固有値が第2固有値の約7.6倍の大きさを持つことから(図3)，各評定者は共通して同じ内容を評定しているとみなすことができる。

表1　各評定者の評点の平均，標準偏差

評定者	平均	標準偏差
1	53.8	8.0
2	47.0	7.5
3	59.7	12.8
4	50.1	15.7
5	48.7	15.2
6	53.8	7.0
7	72.3	4.7
8	61.9	10.2
9	49.2	12.0
10	70.4	6.3
11	47.3	15.1

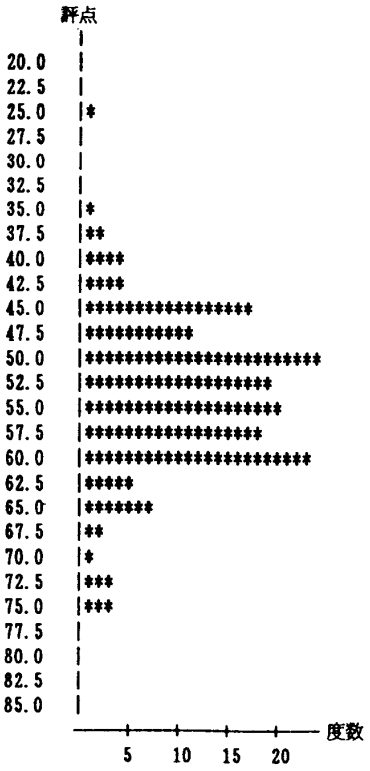

図1　評定者1の評点の分布

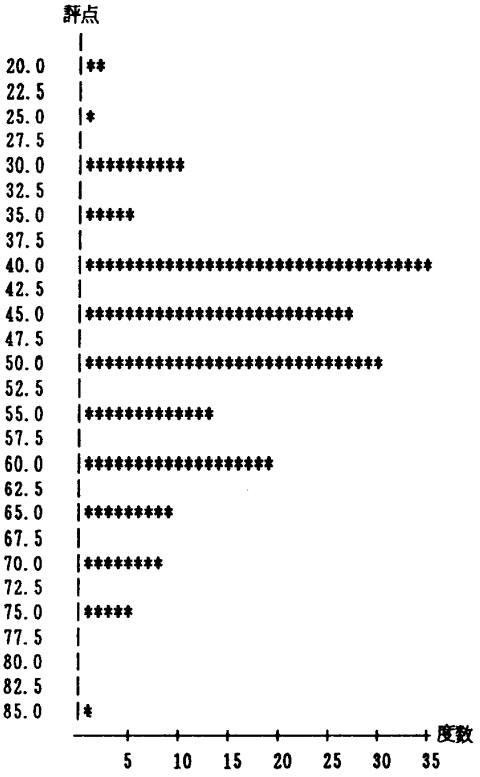

図2　評定者9の評点の分布

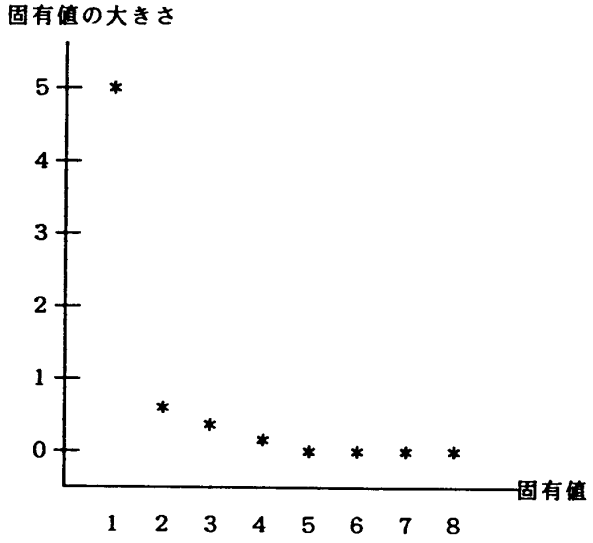

図3　主因子解による固有値のプロット

B．段階カテゴリ分け

100点満点の評点を，段階カテゴリに分類する。そのさい，各カテゴリの頻度をできるだけ均等にするように分類する方法（以下，パーセンタイル基準とよぶ）と，各評定者の平均と標準偏差にもとづいて分類する方法（以下，SD基準とよぶ）の2通りの分類方法をとった。段階カテゴリの数はそれぞれ2～5までの4種類とした。

SD基準にもとづく分類では，2カテゴリの場合，各評定者ごとに平均より高いか低いかで2分割した。3カテゴリの場合は，各評定者の平均よりその評定者の標準偏差の2分の1以上高いか，平均から上下2分の1の範囲に入るか，平均より標準偏差の2分の1以上低いかで3分割した。4カテゴリの場合も同様に，平均の1標準偏差以上高いか，平均より高いが1標準偏差以内か，平均より低いが1標準偏差以内か，平均より1標準偏差以上低いかで4分割した。5カテゴリの場合は，両極のカテゴリの頻度を確保するため，分割の基準をやや狭めた。すなわち平均より標準偏差の5分の6以上高いか，5分の2以上5分の6未満高いか，平均から上下5分の2の範囲に入るか，平均より5分の2以上5分の6未満低いか，5分の6以上低いかで5分割した。

表2はパーセンタイル基準にもとづくカテゴリ得点の平均，標準偏差を示したものである。粗評点の平均や標準偏差が評定者によりまちまちだったのに対し，カテゴリ得点の平均や標準偏差はほぼ同じ値となっている。これによって評定者の評定基準がある程度コントロールされたといえよう。また表3は，SD基準にもとづくカテゴリ得点の平均，標準偏差を示したものである。SD基準にもとづくカテゴリ得点も，平均や標準偏差がほぼそろっ

ている。ここで，パーセンタイル基準に比べて各カテゴリ得点の標準偏差が小さめになっているが，これは，もともとのカテゴリ分けにおいて，SD 基準の方がカッティングポイントが広めになっていたことによる。

表 2　パーセンタイル基準によるカテゴリ得点の平均，標準偏差

		評定者 1	2	3	4	5	6	7	8	9	10	11
2カテゴリ	平均	1.50	1.53	1.52	1.42	1.51	1.55	1.48	1.52	1.57	1.57	1.52
	標準偏差	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.50	0.50	0.50	0.50
3カテゴリ	平均	1.97	1.93	2.05	2.04	1.95	1.95	1.92	1.96	2.01	1.95	1.96
	標準偏差	0.81	0.88	0.83	0.84	0.84	0.76	0.91	0.82	0.81	0.84	0.86
4カテゴリ	平均	2.49	2.53	2.48	2.42	2.53	2.41	2.82	2.49	2.45	2.52	2.43
	標準偏差	1.10	1.10	1.19	1.12	1.09	1.15	1.04	1.11	1.19	1.13	1.27
5カテゴリ	平均	3.08	2.93	3.02	3.07	2.92	3.10	2.94	2.93	3.16	2.92	2.93
	標準偏差	1.45	1.41	1.35	1.39	1.40	1.44	1.21	1.38	1.32	1.42	1.33

表 3　SD 基準によるカテゴリ得点の平均，標準偏差

		評定者 1	2	3	4	5	6	7	8	9	10	11
2カテゴリ	平均	1.50	1.47	1.52	1.42	1.51	1.50	1.38	1.56	1.52	1.39	1.52
	標準偏差	0.50	0.50	0.50	0.49	0.50	0.50	0.49	0.50	0.50	0.49	0.50
3カテゴリ	平均	1.98	2.04	1.93	2.02	1.86	1.91	2.27	2.08	1.93	1.98	2.12
	標準偏差	0.74	0.67	0.79	0.72	0.78	0.72	0.63	0.82	0.76	0.74	0.76
4カテゴリ	平均	2.48	2.53	2.55	2.44	2.58	2.48	2.42	2.52	2.55	2.41	2.49
	標準偏差	0.89	0.81	0.93	0.96	0.94	0.81	0.78	0.91	0.87	0.88	0.97
5カテゴリ	平均	2.96	3.00	3.04	3.02	2.96	2.91	2.95	3.08	3.07	3.05	2.96
	標準偏差	1.12	0.96	1.11	1.21	1.24	1.04	1.17	1.13	1.15	1.06	1.21

III. 分　析

A. 信頼性の指標

まず，100点満点の粗評点を段階カテゴリに分類したことによって，被験者の能力測定の精度がどのように変化するかを検討する。

被験者の能力測定の精度を表す指標としてよく用いられるものは，テストの信頼性係数であろう。この指標は古典的テスト理論の枠組みで定義されたものである。一般にテスト得点Xの分散を σ_X^2，真の値Tの分散を σ_T^2，誤差Eの分散を σ_E^2 とすれば，テストの信頼性係数 ρ_X は

$$\rho_X=\frac{\sigma_T^2}{\sigma_X^2}=\frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$$

で与えられる。

測定の精度はまた，分散分析を用いて推定することもできる。MS_S を小論文による平均平方，MS_E を残差による平均平方，n を評定者数とすれば，n 人の評定者による平均の信頼性は

$$\rho_n=\frac{(MS_S-MS_E)\cdot n}{(MS_S-MS_E)\cdot n+MS_E/n}$$

で与えられる（渡部，他，1988）。

項目反応理論において，テストの信頼性に相当する概念はテスト情報量である。これは，被験者の能力の最尤推定量 θ の漸近分散の逆数として定義され，能力 θ の関数となる。信頼性係数が被験者集団全体についての精度を表すのに対し，テスト情報量は被験者の能力レベルごとの測定精度を表し，このままでは比較することができない。そこで，テスト情報関数から信頼性係数を推定し，テスト信頼性係数および分散分析による信頼性の推定値と比較する。

豊田（1989）は，項目固定型の項目反応モデルにおける信頼性の推定方法として，能力推定の誤差分散 σ_E^2 が

$$\sigma_E^2=\int I(\theta)^{-1}g(\theta)d\theta$$

で与えられることを利用し，信頼性係数 ρ_I を

$$\rho_I=\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_E^2}$$

で推定した。ただし，$I(\theta)$ はテスト情報関数，$g(\theta)$ は能力 θ の分布を表す。

本研究において，$g(\theta)$ は標準正規分布であり，$\sigma_\theta^2=1$ なので，信頼性係数は

$$\rho_I=\frac{1}{1+\sigma_E^2}$$

で推定できる。

表 4 はこうして求めた信頼性の推定値である。この表から，カテゴリ数が増えるとともに信頼性が高まること，同じカテゴリ数でもパーセンタイル基準より SD 基準の

方がカテゴリ数にして約1つ分だけ信頼性が高いことがわかる。このことは，図4，5によって説明される。図4はパーセンタイル基準のカテゴリ数別情報量を，図5はSD基準のカテゴリ数別情報量を示したものである。SD基準にもとづくカテゴリ分けの方が幅広い能力層を精度良く測定していることがわかる。情報量のピークの高さも，4カテゴリまでは2通りの基準でだいたい同じ高さだが，5カテゴリになるとSD基準の方が明らかに高い。つまり，SD基準の方が相対的に$I(\theta)$が大きくσ_E^2が小さくなっているため，信頼性が高くなるのである。

表4　段階反応モデルの信頼性推定値

基　準	カテゴリ数	信頼性推定値
パーセンタイル	2	0.59
	3	0.71
	4	0.74
	5	0.78
SD	2	0.58
	3	0.74
	4	0.78
	5	0.81

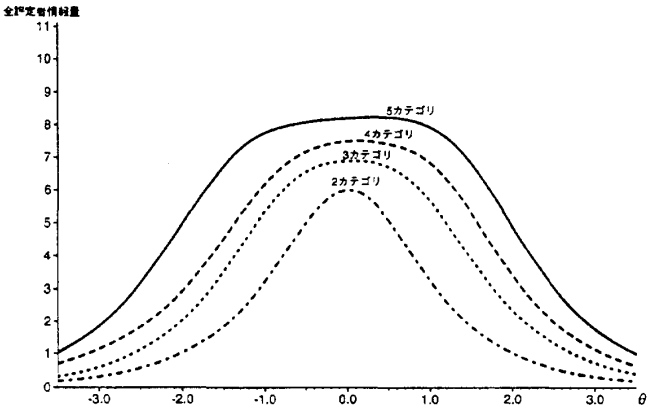

図4　カテゴリ数別全評定者情報量（パーセンタイル）

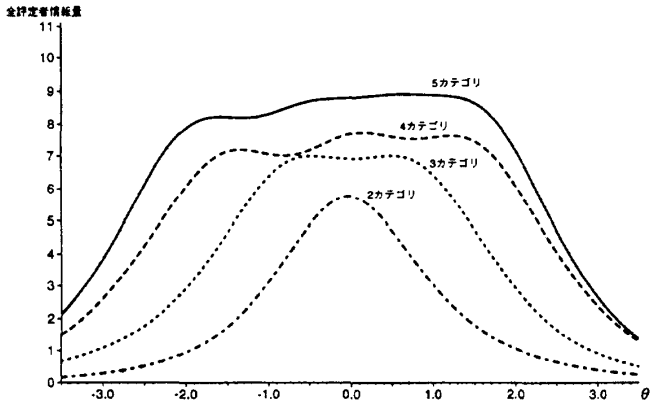

図5　カテゴリ数別全評定者情報量（SD）

B．信頼性の推定値の比較

段階反応モデルによる信頼性の推定値を，古典的テスト理論による信頼性係数や分散分析による信頼性の推定値と比較してみよう。

渡部，他（1988）によれば，本研究データの再検査信頼性係数は0.40～0.91であった。ただし再検査信頼性係数は11名の評定者のうち3名（評定者1，2，8）についてだけしか求められていない。この再検査信頼性係数は個々の評定者についてのもので，11名全体としてどのような信頼性になるかはこの数値からだけでは明らかでない。そこで，11名の評定者を5名と6名に分け，折半法信頼性係数を求めた。その結果，折半法信頼性係数は，評定者の組み合わせにより0.72～0.87の値となった。そのメディアン（中央値）は0.80であった。この値は評定者の数でいうと半数での信頼性係数である。そこでスピアマン・ブラウンの公式によって評定者数を2倍にし，11名での信頼性係数を推定したところ，0.83～0.93，メディアンは0.89となった。これらの値は段階反応モデルによる信頼性よりも全体的に高い。とはいえ，段階反応モデルでの最高値はSD基準5カテゴリの場合の0.81であり，この値は粗評点を5カテゴリに圧縮しても，被験者の能力測定上あまり情報が失われなかったことを意味する。

分散分析による信頼性の推定値との比較ではどうだろうか。渡部，他（1988）は各評定者が1つの小論文に同じ真の値を付与する一要因配置モデル，評定者の主効果も考慮に入れた二要因配置モデルにもとづいて分析し，その結果，一要因配置モデルによる信頼性の推定値が0.74，二要因配置モデルで繰り返しがない場合の信頼性の推定値が0.88であることを見いだした。本データは粗評点の平均，標準偏差が評定者によりまちまちなので，二要因配置モデルの方がよりデータを説明していると思われる。また，この0.88という値が折半法信頼性の修正

値のメディアン0.89に非常に近い値であることも注目される。

結局，粗評点を5カテゴリに圧縮しても，段階反応モデルをあてはめることで，かなりの程度信頼性が確保できることがわかった。

C．評定者により異なるカテゴリ数の採用

図2のように離散的な評点のつけ方をした評定者は，全部で4名みられた(評定者3，5，9，11)。このような評点は，評点自体が本質的にカテゴリカルになっている。そこで，この4名についてはもとの離散分布を活かしたカテゴリ分けを導入する。この場合，1つのカテゴリに入る被験者が少なくとも10%いるようにし，10%を切るカテゴリは前後の小さい方のカテゴリに合併した。その結果，評定者3，9，11は7カテゴリ，評定者5は8カテゴリとなった。この4名以外の評定者については，5カテゴリ（SD基準）を採用した。

この結果，11名全体の情報量がどう変化したかを示すのが，図6である。一部の評定者についてカテゴリ数を増やした場合，全員が5カテゴリ（SD基準）の場合に比べて，かえって全体の情報量が減少する傾向があることがわかる。これは，カテゴリ数を増やした結果かえって識別力が下がった評定者がいたためである。おそらくカテゴリ数が多すぎて，パラメタの推定が不安定になったのであろう。この図から，本研究のデータの場合，これ以上カテゴリを増やしてもあまり意味がないことがうかがえる。

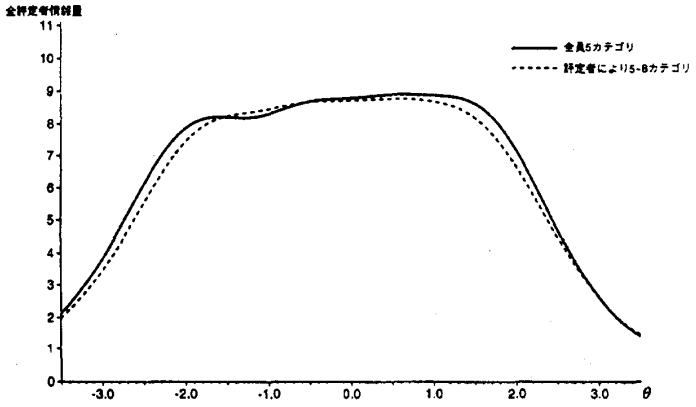

図6　評定者によってカテゴリ数を変えた場合

D．能力推定値間の相関

被験者の小論文を書く能力の推定値としてよく用いられるのは，全評定者の合計点，全評定者の最大値と最小値を取り去った残りの合計点，全評定者のメディアンであろう。そこで，これらの値と段階反応モデル（5カテゴリ・SD基準）による能力推定値との間の相関関係を調べる。段階反応モデルによる能力推定値は，最尤推定値と事後分布のモードの2通りを用いる。

表5はこれら5通りの推定値の間の相関係数である。ただし，全評定者のメディアンに関しては，順位相関係数をとっている。この表から，被験者の書く能力の推定に関しては，どの推定値を用いてもほとんど同じ結果が得られることがわかった。このことは粗評点を5カテゴリという少ないカテゴリ数に圧縮しても，同じように書く能力を推定できることを示しており，カテゴリ評定の有効性が示されたといえよう。

表5　能力の推定値間の相関係数

	①	②	③	④	⑤
①11名の合計点		0.996	0.940*	0.983	0.981
②上下1名ずつを除く9名の合計点	0.996		0.954*	0.982	0.981
③11名のメディアン	0.940*	0.954*		0.941*	0.940*
④最尤推定値	0.983	0.982	0.941*		0.999
⑤事後分布のモード	0.981	0.981	0.940*	0.999	

(注：＊は順位相関係数を表す)

E．評定者ごとの情報量からみた評定者の特徴

最後に，評定者ごとの情報量から評定者の特徴を探ってみたい。

図7は，評定者11の情報量を示したものである。データは5カテゴリ・SD基準のものを用いた。評定者11は，全体の情報量が$-1.5<\theta<1.5$の範囲で全体的に高い。また，各カテゴリの働きが均等で，しかもカテゴリ情報量のピークの位置がほぼ等間隔に並んでいることも特徴的である。各カテゴリの働きが均等ということは，この評定者は幅広い能力層に偏りなく評点を散らばらせていたことをうかがわせる。またピークの位置がほぼ等間隔ということは，粗評点においても例えば高い評点部分での10点の差と低い評点部分での10点の差が同じ意味を持っていたことを示唆する。評定者ごとの情報量が同じような特徴を持っていたのは，他に評定者1，3，5，6，8，9の6名いた。

図8は，評定者2の情報量である。この評定者の特徴は全体的に情報量が少なく，小論文テストで測定する能力の識別が悪いことである。この評定者は評定の観点が他の評定者とやや異なっている恐れがある。また，カテゴリ1，5が，極端に能力の低い被験者か極端に高い被

験者しか識別していない反面，カテゴリ３の識別する範囲はかなり広い。つまり，この評定者は多くの被験者にカテゴリ３に相当する評点を与え，それ以外の評点は，小論文の質が明らかに良いと感じられたかあるいは悪いと感じられた場合のみ与えていると思われる。他に情報量が全体的に少なく，異なる観点から評定していたことが疑われるのは，評定者10である。

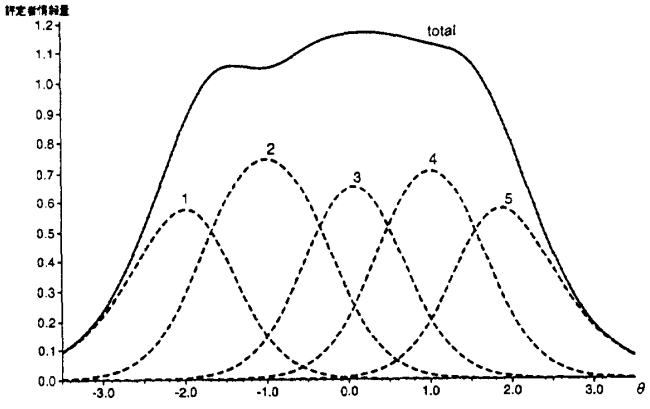

図７　評定者11の情報量

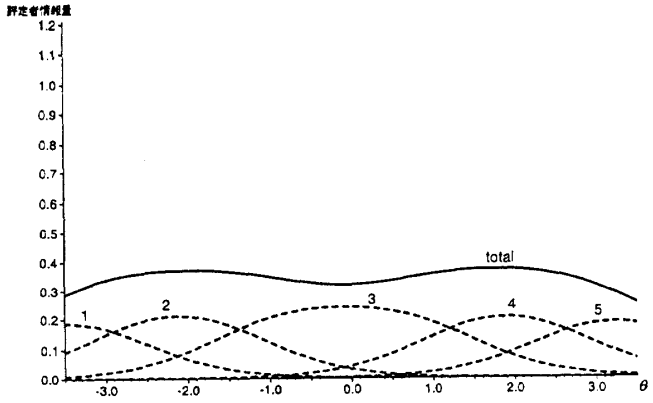

図８　評定者２の情報量

図９は，評定者４の情報量である。この評定者の特徴はカテゴリ２，４があまり機能してなく，実質的にカテゴリ１，３，５の３カテゴリデータに近くなっていることである。この評定者は中程度の質の小論文にはカテゴリ３に相当する評点を与え，カテゴリ２，４に相当する評点はあまりつけなかったことがうかがえる。このように，各カテゴリの機能が均等になっていない評定者は，他に評定者７が該当する。

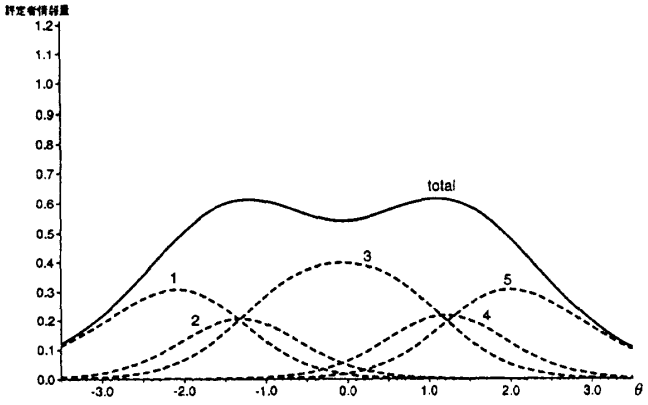

図９　評定者４の情報量

IV．まとめと考察

本研究では，100点満点の粗評点を何通りかの段階的なカテゴリに分け，項目反応理論の立場から段階反応モデルをあてはめ，測定精度がどうなるかを検討した。モデルのあてはめのさいは，評定者を項目に見立て，１つの課題文から被験者の能力を推定した。その結果，各評定者の評点の標準偏差にもとづいて５カテゴリに分類した場合に測定精度が最も大きくなり，100点満点の粗評点による測定にほぼ近い精度が得られることがわかった。また，粗評点をカテゴリ分けすることで，各評定者の評定の甘いとか辛いとかの評定基準がある程度コントロールされることもわかった。さらに，各評定者の情報量を検討することで，各評定者の評定の特徴も見ることができた。

本研究の問題としては，まず用いたデータの被験者数が165名と少なく，段階反応モデルをあてはめるには５カテゴリが限界だった点があげられる。もし数百人規模のデータが得られた場合には，カテゴリ数を増やすことも可能となり，段階反応モデルによって推定される信頼性が折半法信頼性や分散分析による信頼性の推定値を凌駕する可能性もある。

また，段階反応モデルによる信頼性の推定にも問題が残る。この方法は，本来項目数（本研究の場合は評定者数）が20項目以上のときに近似的にあてはまる方法であるとされている（Samejima，1977）が，本研究での評定者はわずか11名である。上記の問題と合わせ，今後の研究では被験者数および評定者数の両方を増やし，改めて段階反応モデルをあてはめてみる必要があるだろう。

なお，項目反応理論の長所の一つとして，θという潜在的な共通尺度の構成がしばしば挙げられる。これは，異なるテスト項目群を受験した場合でもθという共通

尺度上で直接能力を比較することができる，というものである。本研究のように課題文が1つだけである場合，異なる評定者群による評定値が与えられてもそれらを直接比較できるようにすることが可能である。もし，課題文が複数ある場合には，被験者×課題×評定者の3相データとなり，異なる課題文，異なる評定者による評点しかない被験者どうしでも，一次元的能力で測定する限り被験者にせよ，課題にせよ，また評定者にせよ相互比較することが可能となる。この場合，客観式テストと異なり，小論文では課題文が異なれば測定する能力 θ が異なってくる恐れがあるので，測定すべき特性が何かをあらかじめ明確に定めておく必要があろう。

引用文献

Ackerman, T. A. and Smith, P. L. (1988) A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*, 178-128

Blok, H. (1985) Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22*, 41-52

De Ayala, R. J., Dodd, B. G. and Koch, W. R. (1991) Partial credit analysis of writing ability. *Educational and Psychological Measurement, 51*, 103-114

De Gruijter, D. M. N. (1984) Two simple models for rater effects. *Applied Psychological Measurement, 8*, 213-218

Houston, W. M., Raymond, M. R. and Svec, J. C. (1991) Adjustment for rater effects in performance assessment. *Applied Psychological Measurement, 15*, 409-421

Jöreskog, K. G. (1971) Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133

Perkins, K., Pohlmann, J. T. and Brutten, S. R. (1988) A factor analysis of direct and indirect measures of English as a second language writing ability. *Educational and Psychological Measurement, 48*, 1111-1121

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No.17*

Samejima, F. (1977) A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247

豊田秀樹 (1989) 項目反応モデルにおける信頼性係数の推定値 教育心理学研究 37，283-285

渡部洋，平由実子，井上俊哉 (1988) 小論文評価データの解析 東京大学教育学部紀要 28，143-164