小論文評価データの解析

東京大学教育情報科学研究室　渡　部　　洋
同　上　　平　　由実子
同　上　　井　上　俊　哉

An Analysis of Essay Examination Data

Hiroshi Watanabe, Yumiko Taira, and Shunya Inoue

The purpose of this study is to investigate the structure of the data obtained from essay examinations. One set of data was obtained by assigning a small essay test on the title of "Travel" to the 185 10th grade high school students. The essays were evaluated holistically and analytically by the 11 raters, and the 3 raters among them evaluated the same essays repeatedly after more than one week. The orders of the essays were always randomized. Another set of data was obtained by assigning also a small essay test on the title of "Poverty and Richness" to the 144 12th grade girl high school students. The essays were also holistically and analytically evaluated by the other 7 raters (all were teachers), and the 5 raters among them evaluated the same essays after more than two weeks. The main findings were as follows. 1. The consistency of the holistic scores is very low among the raters. 2. The stability of the holistic scores was very different among the raters, and was judged to be low. 3. The 15 viewpoints for the analytic evaluation were classified into the two categories, "content" and "language skills". 4. The first principal component of the scores obtained by the analytic evaluation shows almost the same level of reliability with that of the holistic scores. 5. The high consistency among the raters was observed for the evaluation of the easiness of reading. 6. The close relationship was observed between the holistic evaluation and the general impression of the essays.

I　はじめに

現在では，学校から企業に至るまで，様々な目的のためにテストが利用されている。それらのテストの形式は，共通第一次学力試験でも用いられているような多肢選択型のいわゆる客観式テストと呼ばれるものから，論文テスト，面接テスト等々実に多くのものがある。しかしながら，それらのテストのうちでも専門的研究が行なわれているといえるのは，客観式テストと呼ばれる形式のものについてだけであり，他の形式のテストについては，その重要性がいわれながらもほとんど本格的な研究が行なわれていないのが実状である。とくに，入学試験に関しては，1979年に国公立大学入試に共通第一次学力試験が導入されて以来，二次試験になんらかの形で小論文試験を課す大学が増えてきており，国公立大学における一般入試の場合で全体の4分の1程度の学部で，また同じく推薦入学の場合には全体の3分の2の学部で小論文試験を実施するに至っている（国立大学入学者選抜研究連絡協議会，1983，1984，1985，1986，および1987）。それにもかかわらず，我が国における小論文試験に関する研究報告は，いくつかの例を除き，ほとんどない状況である。そこで，以下では，小論文データの基本的構造を明らかにするための基礎的研究として，2つの小論文に関するデータを様々な角度から解析することを試みる。

II　先行研究について

1．小論文評定の諸観点

小論文試験を実施する側は，受験者のある種の能力が答案に反映することを期待しているが，その結果には実際に

は様々な要因が複雑に関与している。Cooper (1984) は,「小論文がWriting Abilityを測定としているものと考えると,その得点に関して誤差要因として働くものには,書き手(writer),題目(topic),形式(mode),時間制限(time-limit),テスト状況(examination situation),そして,評定者(rater)がある」という。これらの大部分は,いわゆる「試験」に共通して見られる要因であるが,特に「評定者の要因」は小論文試験にとって重要なものである。そこでここでは,まず評定者がどのような観点で小論文評価を行なっているかという問題に焦点を当てた研究を取り上げる。

イタリアのRemondino (1959) は,作文を評価する者がどのような観点で作文を見ているかを明らかにしようと考え,複数の国語(イタリア語)教師(約20名)に識別可能な作文の性質について質問し,回答のリストから作文の評価項目として以下の17観点にまとめた。

Remondinoによる作文評価の17観点

1. 読みやすさ:書字の質—明瞭,読みやすい
2. 美的配置:文字の並び方の美しさ
3. 見掛け,きれい,注意深く提出されている
4. 綴り:綴りの誤りの数
5. 単語の形式,語形変化:単語の形,語形変化における誤りの数
6. 文法:文法の誤りの数
7. アイデアの組み立て:アイデアの並べ方とバランス
8. アイデアの豊かさ:アイデア,モチーフの数
9. 思考の適切さ:アイデアが表現されている性質がテーマに合っているか
10. 事実性:思考,事実の客観的正確さ
11. 理解可能性:意図されていることの表現が完全か
12. 簡潔性:最小限の単語を使っているか,繰り返し,冗長さがないか
13. 用語力:単語の正しい使い方
14. 文体:文の構成の平易さ,正確さ
15. 独創性:書き手の性格が,ある部分文の中に明らかになっているか
16. 成熟度:判断力,推理力,批判力
17. 想像力:創造力,投影力

その上で,これらの17観点に基づいて作文の評定に関与する複雑な要因を分離するために,因子分析を用いて,評定者の評定の構造を研究した。その結果,評定者が作文の評定の経験を持つ教師であっても,その経験を持たない人物であっても,評定の構造は4つの因子(外見上の美しさ,用語力,内容と構成,および内容の個人的側面の4因子)でほとんど説明されると結論している。また,このデータを用いて,McColly (1970) が追跡研究を行なっている。

Remondinoの研究は,作文を評定者に様々な観点から分析的に評価させたが,個々の作文に一つの得点を与えるような総合的な評価に,分析的な個々の要因がどのように影響しているかという点に関する研究も多い。

Harris (1977) は,高校の教師に対して「作文に何を求めるのか」を調査した。36人の高校教師に高校生レベルの12作文を評価させたところ,①教師が最も重要だと思う要因が評価に最も影響しているわけではない,②教師が最も重要だと思う基準と作文の評価・修正に用いる基準が負に相関する,③教師の用いる基準に個人差の存在を予測したが,教師は皆同じ様な基準で評価している,④教師は文の構造を見る時に,構造の要素や文のパターンの多様性よりも,まず,技術的正しさを見る,といったような傾向が見出された。すなわち,教師は作文の評価について,個々の基準の概念的側面については一致しているが,現実の評価における理論と実践に矛盾があり,内容と構成に重きを置くにもかかわらず,技術の巧拙に強く影響を受けると結論している。

Freedman (1979) は,説明的な論文を4カテゴリー(内容,構成,文構造,技巧)で変化させて書き直し,12人の評定者に4件法の総合評価で論文の質を判断させた。分散分析の結果,内容と構成が評価に最も影響し,文構造と技巧の影響力は小さかったが,構成と交互作用があったとしている。

Grobe (1981) は,Stewart & Grobe (1979) の追試を中心に研究した。StewartとGrobeによると,説明的作文の教師の評価を分析した結果,教師の作文に対する評価は文法的成熟度(syntactic maturity),技巧(mechanics)などより,作文の長さ(essay length),綴りの正確さ(freedom from spelling errors)などに影響を受けるということである。Grobeは,5,8,11年生の書いた物語的作文を用いて,各作文につき2人の評定者に総合評価(1～4点)をさせ,次に18人の評定者によって,技巧,文法(syntax),用語(usage)などに関する14個の観点で評価させている。総合評価を基準変数,分析的評価を説明変数として,各学年毎にステップワイズ回帰にかけたところ,充分な説明率が得られなかった($R^2=0.32\sim0.60$)。そこで,語彙について10変数を追加すると説明率が上昇した。この結果から,説明変数の中で語彙の特徴(vocabulary characteristics)が総合評価と最も強い関係が

あると結論している。

2. 小論文試験の妥当性

小論文試験が,試験の実施者が測定したいと考えている能力を，どの程度測定しているかも重要な問題である。

Coffman (1966) は，Godshalk et al. (1966) のデータを用いて，論文試験の妥当性が高いことを示そうと試みた。Godshalk らのデータは，646人の中学生に5つの作文を書かせ，1つの作文につき5人の評定者に評価させたものであるが，Coffman はそのうち2つの作文を取り上げ，新たな評定者を用いて評価させた。そして，2つの作文の元の得点（Godshalk らの研究での得点）と新たな評定との相関が非常に高いことから,論文試験の妥当性は高いと結論づけている。ただし，妥当性係数の算出にあたっては，一部は実際のデータから求めているが，この論文の結論に関わる部分は，スピアマン・ブラウンの公式や希薄化の修正公式を用いて外挿的に推測したものである。また，彼は客観テストに匹敵するほど論文試験の信頼性・妥当性を高めるためには，数多くの課題（題目）と評定者を確保する必要があるとも述べている。

3. 小論文試験における外的誤差要因

小論文の評定が,ある程度以下の信頼性しか確保することができないものだとすれば，その妥当性・信頼性を阻害する要因には，どのようなものがあるだろうか。勿論，誤差として得点に影響する要因を全て挙げ尽くすことは不可能であるが，そのうちの一つを取り上げた先行研究は，比較的多く存在する。ここでは，それらのうち，文字・綴り，評定の系列的効果，課題選択などについての研究をいくつか取り上げることとする。

① 文字・綴り

文字の巧拙などは,小論文評価のための本質的な要因であると見なす考え方もあり得るが,Chase はそれを誤差要因の一つと考え，その影響についての研究をいくつか行なっている。

Chase (1968) は，内容以外に作文の評価に影響する要因として，「文字の上手さ，綴りの正確さ，得点化の手掛かりの有無」に焦点を当てて調査した。その結果，文字の上手さは得点に有意に影響し（上手な方が高得点），綴りは影響しないが,得点化の手掛かりのある作文に評定者が高い得点を与える傾向があると結論している。

Chase (1979) は，成績に対する評定者の期待と文字の上手さが,論文試験の得点にいかに影響するかについて調査した。この結果，文字の上手さだけでは得点に影響する有意な変数とはならないが，評定者の期待の効果が，文字の上手さと論文試験の得点の関係に変化を与えるという。すなわち，文字が下手でも，評定者の成績に対する期待が高い群の得点が最も高く,文字の効果だけで考えていた研究と対照的な結果になる。

また，Chase (1983) は，前述の研究結果を受けて，文字の拙さ，綴りの間違い，文法の間違いなど論文試験の得点を減じる効果の共通要素として「論文の読みにくさ」を挙げ，綴り・文法は正確だが，理解の困難さの程度が異なる論文を評価させた。その結果，理解の困難さは，論文試験の得点に関係しているが,曲線的に相関している可能性があり，評定者は最適の困難度のものに最高得点を与え，理解が容易すぎても困難すぎても,その論文の得点は下がる傾向があるという。

誤差要因として文字の巧拙に注目した研究者は Chase 以外にもいる。Marshall & Powers (1969) は，12形式の論文(内容が同じで，文字の上手さ，文の間違いのタイプが異なる）を420人の教職志望者に評定させた。その結果,内容のみに基づいて評価するように指示があったにも関わらず，評定者は文字の上手さと文の出来 (quality of composition) に左右され，文の間違いと文字は有意な交互作用がなかったとしている。しかし，この研究では，得点の高い順に,上手な文字→下手な文字→タイプの文字→普通の文字，となり，予想に反した結果となっている。この原因については不明としている。

② 評定の系列的効果

ある小論文の評定が答案全体の中で何番目に行なわれるかということが,その論文の評価に影響を及ぼす可能性がある。例えば，特に良い論文の後に評定される場合と，あまり良くない論文の後に評定される場合とでは,論文に与えられる得点が異なるかも知れないと考えるのは自然である。前に評価された小論文の出来・不出来が得点に与える影響を，ここでは「系列的効果」と呼ぶこととする。

系列的効果の影響に関しては，Hales & Tokar (1975)，Hughes et al. (1980)，Daly & Dickson-Markman (1982)，Hughes et al. (1983b) など，一連の研究がある。しかし，これらの研究全般に言えることは，系列的効果は確かに存在するが,その影響を得点から取り除くことは非常に困難だということだけである。

③ 課題選択

異なる課題に基づいて書かれた小論文をどのように評価するかという問題は非常に困難である。日本の大学入試小論文で採用している所も若干ある（例えば，茨城大学の人文学科)が，受験者に論文の課題の選択が与えられる試験状況においては,選択した課題の内容によって結果に有利不利が生じる場合も考え得る。

Meyer (1939) は，大学の定期試験（論述式）の問題において，学生が問題（課題）を自由選択して解答する方法の是非について検討している。彼は，難しい問題を選択した学生が不利な評価を受ける可能性があることを主張し，問題の難易度に応じて得点に対する重みを考えるなどすべきであると述べている。

④ 様々な誤差要因

評定に影響すると思われる様々な誤差要因の影響を同時に取り扱って実験計画法的に見ようとした研究もある。

文字の影響の研究で前述した Chase (1986) は，書き手の性別，人種（黒人・白人），成績に関する情報を評定者に与え，文字の巧拙の要因を加えて，80人に評定させるという研究を行なった。分散分析を用いた結果によると，主効果に有意なものはなく，4次の交互作用と3次(人種×成績×文字)の交互作用が有意であり，外的要素は様々な組み合わせで論文試験の評価に影響するとしている。

Coffman & Kurfman (1968) は，4人の評定者，8グループの作文，2種類の評価方法（総合評価，分析的評価），2日に渡る評定，読む順序（グループ毎）などの要因を組み合わせた2つの実験計画法（主効果を見る実験I，交互作用を見る実験II）で実験を行なった。結果は，日の効果，評定者の効果，評定者と日の交互作用などが有意であったとしている。

4. イレブン・プラスに関する研究

大学入試に関する研究ではないが，入試における小論文の研究として，1950年代を中心とした英国イレブン・プラスに関する研究は注目に値する。イレブン・プラスとは，この当時，11才の子供の多くが grammar school 入学のために受けていた選抜試験のことである。全国的に統一的な試験科目が決められていた訳ではなく，地域によっては作文を課していたところがあったようである。イレブン・プラスに関する研究のうち，ここでは2つについて概説する。

Wiseman (1949) は，その当時の先行研究から，作文の評定の信頼性が低いことを示し，それを高める方法について述べている。当時，作文の評定方法としては，いくつかの観点について評定する分析的評価が主流であったが，この方法は作文を全体的に評定する総合評価と比べて著しく時間とコストが掛かる反面，それほど信頼性の向上が期待できず，また，作文の全体としての出来を反映していないのではないかと考え，総合評価（彼は General Impression と呼んだ）の有効性を強く主張している。これは，イングランド地方南部のデボン州のイレブン・プラスで，4人の評定者が独立に作文を評価し，その合計得点をとる方法により，信頼性を高めることに成功したという結果に基づいている。また，評定者の信頼性の概念に，「評定者間の一致」と「評定者内の一貫性」の2種類あることを指摘し，後者をその基礎とすべきであるとしている。

Wiseman & Wringley (1958) は，課題選択と評定の信頼性について報告している。イレブン・プラスの試験方法は，共通のパターンは有るものの，147地域毎に異なり，英語のテストに関しては，客観テストと作文テストの是非論が盛んであった。作文テストで受験者による課題選択がある場合，生徒の想像力が充分に働く場を与えるという利点があるが，反面，選んだ課題により優劣が決定されたり，評定者の好みなどによる不公平が生ずる可能性がある。彼らの研究では，その点の調査がなされ，課題によって平均得点に差があることが実証されたが，これらの差の多くは，子供の能力の個人差によって作り出され，評定者による違いはそう多くないとしている。また，沢山の課題の中から受験者が一つを選んで解答しても，評定の本質的な誤差要因にはならないとしており，たとえ課題選択を許さなくとも，信頼性が上がることはない，と結論している。

5. 理論的研究

必ずしも小論文試験のみに適用される訳ではないが，広く主観的に得点が与えられるテストの得点に関する統計モデルの中から，De Gruijter (1980) と Blok (1985) について，ここでは簡単に述べておく。

De Gruijter (1980) は，論述試験において，テストの信頼性に与える主に評定者の影響について，理論的に述べている。彼は，信頼性に影響する要因を「評定者 (rater)」と「問題 (question)」に分け，信頼性の算出方法や信頼性の高い合成得点を作るための考え方など，様々な点について論じている。

Blok (1985) は，論文評価の重要な問題として multiple rating（異なる評定者の評価と同じ評定者の繰り返し評価）が同じ真の得点を反映していると仮定できるかどうかということを挙げ，16人の評定者に105の作文を2回評価させ，線形構造方程式モデルでこの仮定を吟味した。その結果，異なる評定者間の推定された真の得点間の相関は0.41～0.91で，同じ真の得点を反映しているとは言えないが，同じ採点者の評価は同じ真の得点を反映していると結論している。

6. その他の研究

古くは，Willing (1926) が生徒に作文指導をする上で，作文を実際に書かせるよりも，間違いを発見する課題を用

意して与える方が効果的であると主張している。また，作文技能の開発とその評価に関する研究（Veal，1966）などもある。

論文試験そのものだけでなく，客観テスト（多肢選択テスト）と比較したものとして，英国のイレブン・プラスに関連した研究（Peel & Armstrong，1956；Wiseman，1956）や，大学生の作文能力の低下などの問題から，作文能力測定に関心を寄せたもの（Breland & Gaynor，1979；Hogan & Misher，1980）などがある。

コンピュータを用いて論文試験の評定を自動化する研究（例えばHiller et al.，1969），認知心理学的な立場から読む能力と書く能力との関係を比較分析したもの（Shanahan & Lomax，1986），作文能力（Writing Ability）の測定に関する研究（Benton & Kiewra，1986）などがあるが，本論文の主題とは異なるので，ここではこれ以上ふれないこととする。

III　方法

1．データI

「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）の計185名に，時間制限は特に設けずに，字数800前後以内で小論文を書かせ，学校や塾で国語教育に携わっている者5名（R1～R4およびR6）とそれ以外の教師や大学院生6名（R5およびR7～R11）の計11名を評定者として採点した。評定者は全員全ての小論文（順不同）を2回読み（順不同），1回目は100点満点で総合評価を，2回目にはA1からA15の15の観点に関して，7段階評定により分析的評価（表III－1参照）を行なった。また，そのうちの3名の評定者（国語教師R1とR2，および大学院生R8）は，1週間以上の期間の後に再び全小論文について総合評価と分析的評価を行なった。なお，以下の解析では最初の20名分の小論文を評定練習用とみなして除いてある。したがって，以下の解析で用いた小論文数は165である。

2．データII

「貧しさと豊かさ」という題で，東北の私立女子高校3年生144名に，60分800字以内で小論文を書かせ，7名の現職教諭（R'1理科，R'2数学，R'3～R'7国語担当）が評定者となって採点した。評定者はさきと同様全員全ての小論文（順不同）を2回読み（順不同），1回目は100点満点，2回目にはA'1からA'14までの17の観点（さきのものを改良したもの）に関して7段階評定により分析的評価（表III－1参照）を行なった。（但し，評定者R'1については総合評価のみを2回）。また，そのうちの5名の評定者（R'1～R'5）は，2週間以上の期間の後に再び全小論文について総合評価と分析的評価（但し，R'1については総合評価のみ）を行なった。なお，データIの場合と同様，以下の解析では最初の20名の小論文を除いてあるので，実際の解析に用いた小論文は124である。

表III－1　分析的評価のための観点

	データI	データII
誤字・脱字	A 1	A' 1
用語力	A 2	A' 2
文字	A 3	A' 3
文法	A 4	A' 4　a，b
文体	A 5	A' 5　a，b
課題のとらえ方	A 6	A' 6
発想	A 7	A' 7
文の構成	A 8	A' 8　a，b
表現力	A 9	A' 9
知識	A10	A'10　a，b
論理性・一貫性	A11	A'11
思考力・判断力	A12	A'12
一人よがり	A13	A'13
読後感	A14	A'14
親近感	A15	ナシ

IV　総合評価に関する結果と考察

1．総合評価の平均と標準偏差

2種類のデータ（「旅」と「貧しさと豊かさ」の各評定者ごとの1回目の総合評価の評価値の平均と標準偏差（S.D）は表IV－1の通りである。データIおよびデータIIともに，最初の20名分の小論文を取り除いた場合とそうでない場合のそれぞれについて，平均と標準偏差を示してある。表からも明らかなように，評価値の平均と標準偏差は評定者によって大きく異なり，データIおよびデータIIともに，その平均で25点以上，標準偏差で3倍以上の差異が見られ，評定者によって評価の仕方が大きく異なることがわかる。最初の20名を含んだ場合とそうでない場合とを比べると，データIおよびデータIIともに，平均および標準偏差の両方が，最初の20名分を含んだ場合の方が値がやや大きい。このことは，最初の方の小論文はやや高めに評価されかつ変動が大きいことを示唆するが，そのような系列的効果が存在するかどうかは別に統制された実験によっ

て確認される必要がある。また，データⅠとデータⅡを比較すると前者の方が平均，標準偏差ともにデータⅠより大きい。いずれにせよ，表Ⅳ－1の結果は少なくとも小論文を公平に評定するためには，あらかじめ幾つかの小論文を実際に採点してみて練習しておくことと，評定者間でその練習結果を照らし合わせて評定者間の差異をできるだけ少なくしておく努力が重要であることを示している。

2. 総合評価間の相関係数とその主成分分析

2人ずつ評定者を対にして，総合評価の評価値間の相関係数の値を求めたところ，表Ⅳ－2(a)および表Ⅳ－2(b)のようになった。データⅠでは，相関係数は0.22～0.57の間に分布しており，その全体における平均は0.43であった。評定者ごとの相関係数の平均（表Ⅳ－2(a)における最右列参照）をみると，その値が小さい評定者R2，R7，R10は，総合評価の評価値の標準偏差（表Ⅳ－1を参照）がそれぞれ7.5，4.7および6.3とかなり小さいことがわかる。データⅡでは，相関係数は0.05～0.58の間に分布しており，その全体における平均は0.26であり，データⅠと比べかなり低い。また，評定者ごとの相関係数の平均（表Ⅳ－2(b)における最右列参照）も当然ながら全体的に低いが，標準偏差が小さいからといって相関係数の値が小さいという傾向は見られない。但し，データⅡでは全体的に標準偏差の値が小さいということも，相関係数の値の大きさを吟味する際には考慮する必要がある。

表Ⅳ－2(a)および(b)に示された2つの相関行列を主成分分析した結果は，表Ⅳ－3～6に示されている。表Ⅳ－3からも明らかなように，データⅠにおいては，第1主成分と第2主成分とで全体のおよそ60％の変動を説明しており，第1主成分の負荷の大きさ（表Ⅳ－4参照）は表Ⅳ－2(a)に示された相関係数の平均にほぼ対応している。データⅡにおいては，表Ⅳ－5にも示されている通り，データⅠに比べ第1主成分と第2主成分とで説明する割合が53％でやや低いが，第1主成分負荷における傾向はデータⅠと変わらない。

3. 再検査信頼性

データⅠでは3名（R1，R2およびR8）に，データⅡでは5名（R'1，R'2，R'3，R'4およびR'5）に，一定期間後に再び同じ小論文の評価を行なってもらったが，総合評価に関しては表Ⅳ－7のような結果を得た。表中，2回目の平均のうちで＊印や＊＊印がついているのは，平均値の差についてのテイ検定を両側検定で行なったとき，5％水準で有意ならば＊印，1％水準で有意ならば＊＊印ということを意味している。すなわち，1回目と2回目の評定において，平均値について有意な差が見られたのは8人の評定中5人であり，そのうち1人の評定者のみが2回の評定の平均値が1回目のそれよりも高くなっている。

繰り返し測定間の相関係数は，いわゆる再検査信頼性と呼ばれるものに等しいが，その値は表Ⅳ－7からもわかる通り，0.40～0.91のあいだに広く散らばっている。全体として，データⅠの方がデータⅡの場合よりも相関係数の値が大きく再検査信頼性が高くなっているが，その理由としては評定者の違い，小論文の題の違い，および生徒の違いの3つの要因が考えられる。但し，データⅡを見ると評定者の違いのみによって再検査信頼性が，0.40～0.81までの差異があることが示されており，評定者によって小論文評定の再検査信頼性が大きく異なることがわかる。なお，データⅠの評定者に対しては全員に評価終了後，簡単なアンケート調査を行なったが，そのアンケートから非常に高い安定度を示した評定者R1は，数項目からなる自分なりの評価基準をつくり，配点をして，それらの合計で得点を与えるという方略をとっていたことが判明した。

4. 推定された真値との相関

ここでは，各評定者ごとの総合評価の評価値の信頼性係数を推定するために，各評定者の総合評価値と小論文の「真の評価値」との間の相関を考えてみる。そのためには，各小論文の真の評価値をまず推定しなければならないが，その方法として以下では，

(イ) 小論文ごとの評価値の平均
(ロ) 小論文ごとの評価値の中央値
(ハ) 小論文ごとの平均と全体平均との重みづけ平均
(ニ) 対数変換を伴う小論文ごとの平均と全体平均との重みづけ平均

の4つの方法を用いた。(ハ)のための具体的な計算式はNovick and Jackson (1974) の（9．5．12）式によって与えられており，このための計算式も同じく（9．5．12）式，（9．6．7）式および（9．6．10）式によって与えられている。表Ⅳ－8は，データⅠにおける評定者ごとの真の評価値の推定値と総合評価との間の相関係数を示している。但し，(ロ)の中央値との相関をとるにあたっては，スピアマンの順位相関係数を用いてある。また，参考のために，評定者R1，R2およびR8について再検査信頼性（表Ⅳ－7参照）の値も併記してある。この真の評価値の推定値と総合評価値との間の相関係数は，小論文総合評価値の信頼性係数の推定値の一つと見なし得るが，信頼性係数の推定値のなかでも特に，評定者間の一貫性に関する情報を含むものと考えられる。但し，この相関は自分自身を含むものとの相関であるために，やや高めの数値と

なっていることに注意する必要がある。11名の評定者の中で特に値が小さい者は，R2 と R10 であるが，この 2 人は評定者間相関の平均(表IV— 2(a)参照)の値もそれぞれ.36と.35と，やはり低い評定者である。また，(イ)から(ニ)までの 4 種類の信頼性係数の推定値のうち,(ロ)のいわばノンパラメトリックな信頼性係数の推定値が他と比べてやや低目の値となっているが，これは 1 つには中央値は平均値と比べ自分自身の値を直接には含まないということからくるものであろう。また，R1 と R2 においては，4 種類の推定値よりも再検査信頼性の値の方が大きくなっており，評定者によっては，評定者間の一致度よりも時間的安定性が高いことがわかる。

データ II においても同様な数値を求めた結果が表IV— 9 である。データ I の場合と比べると数値が一般的に小さく，4 種類の推定値のうち(ロ)のノンパラメトリックの推定値の値が評定者 R'2 と R'3 を除きやはり小さい。再検査信頼性の値と 4 種類の推定値，または評定者間の相関の平均（表IV— 2(b)参照）とを比較すると，R'5 は再検査信頼性の値が非常に大きく，他の評定者との一貫性よりも時間的一貫性が高いことがわかる。

5. 分散分析による信頼性の推定

古典的テストモデルのもとでのテストの信頼性という概念に限界があることは，多くの研究者が論じてきているところであり(たとえば，Cronback et al., 1963, 1972；Lord & Novick, 1968；池田, 1973；Thorndike, 1982；Brennan, 1983等々)，そこでは分散分析を用いて一般化可能性(generalizability)という概念のもとでテスト得点の性質が論じられることが多い。

(1) 一要因配置のもとでの信頼性

ここでは一要因配置モデル（変量効果モデル random -effects model)

$$Y_{ij}=\mu_i+E_{ij},\quad i=1,\ 2,\ \cdots,\ N$$
$$j=1,\ 2,\ \cdots,\ n$$

のもとで，信頼性を吟味する。但し，i は小論文を j は評定者を示し，Y_{ij}は i 番目の小論文の j 番目の評定者による総合評価値を示す。μ_iは i 番目の小論文のいわば「真の評価値」であり，E_{ij}は誤差を表す確率変数である。このモデルのもとで，データ I を分散分析した結果が表IV—10に示されている。この分散分析表より，評定表 1 人当りの信頼性は

$$\hat{\rho}_1=\frac{(MS_S-MS_E)/n}{(MS_S-MS_E)/n+MS_E/n'}$$
$$=\frac{(598.5-156.4)/11}{(598.5-156.4)/11+156.4}$$
$$=0.20$$

である。但し，上式において，MS_Sは小論文による平均平方，MS_Eは残差により平均平方を表し，n は評定者数を示している。n'は評定者何人当りの信頼性であるかを示すものである。従って，評定者11人による総合評価値の平均の信頼性は

$$\hat{\rho}_{11}=\frac{(598.5-156.4)/11}{(598.5-156.4)11+156.4/11}$$
$$=0.74$$

となる。また，評定者の数を無限に増したときの総合評価値の平均の信頼性は

$$\hat{\rho}_{\infty}=\frac{(598.5-156.4)/11}{(598.5-156.4)/11+156.4/\infty}$$
$$=1.00$$

となる。逆に，信頼性が0.90となるためには，

$$\hat{\rho}_{n'}=0.90$$

とおくことによって，n'を求めると n'≒35となることから，35人以上の評定者が必要であることがわかる。

データ II において，同様に信頼性を求めると(表IV—11の分散分析表を参照)

$$\hat{\rho}_1=\frac{(184.5-144.9)/7}{(184.5-144.9)/7+144.9}$$
$$=0.04$$
$$\hat{\rho}_7=\frac{(184.5-144.9)/7}{(184.5-144.9)/7+144.9/7}$$
$$=0.21$$

となり，信頼性が0.90となるためには評定者が231人以上必要となる。

(2) 繰り返しがない場合の二要因配置のもとでの信頼性

ここでは，繰り返し測定が無い場合の二要因配置モデル（変量効果モデル）

$$Y_{ij}=\mu_i+\alpha_j+E_{ij},\quad i=1,\ 2,\ \cdots,\ N$$
$$j=1,\ 2,\ \cdots,\ n$$

のもとでの信頼性を吟味する。但し，α_jは評定者の主効果（変量効果)を示すもので，いわば評定者の全体的な評価の甘さを表すと考えることができる。その他の変数は先のモデルと同様である。このモデルのもとで，データ I を分散分析した結果が表IV—12に示されている。この分散分析表より，評定者 1 人当たりの信頼性は

$$\hat{\rho}_1=\frac{(MS_S-MS_E)/n}{(MS_S-MS_E)/n+MS_E/n'}$$
$$=\frac{(598.5-74.6)/11}{(598.5-74.6)/11+74.6}$$
$$=0.39$$

である。但し，上式において，MS_Eは残差による平均平方

（すなわち交互作用による平均平方）を表している。また，評定者11人による総合評価値の平均の信頼性は

$$\hat{\rho}_{11}=\frac{(598.5-74.6)/11}{(598.5-74.6)11+74.6/11}$$
$$=0.88$$

である。また，信頼性を0.90にするためには計算上，少なくとも14名の評定者の評価値を平均する必要があることがわかる。

データIIにおいて，同様に信頼性を求めると(表IV—13の分散分析結果を参照)

$$\hat{\rho}_{1}=\frac{(184.5-65.1)/7}{(184.5-65.1)/7+65.1}$$
$$=0.21$$

である。また，評定者7人による総合評価値の平均の信頼性は

$$\hat{\rho}_{7}=\frac{(184.5-65.1)/7}{(184.5-65.1)/7+65.1/7}$$
$$=0.65$$

となる。さらに，信頼性を0.90にするためには計算上35人以上の評定者の評価値を平均する必要があることになる。

(3) 繰り返しがある場合の二要因配置のもとでの信頼性

上記の(1)および(2)のモデルのもとでは，繰り返しが無い場合のデータのみを取り扱ったが，データIの場合にせよデータIIの場合にせよ，それぞれ再検査信頼性を求めるために，データIでは3名，データIIでは5名の評定者に一定期間の後に再び同じ小論文を総合評定させており，そのデータを用いるならば，繰り返しがある場合のモデルのもとでの解析が可能となる。すなわち，ここでは繰り返し測定がある場合の二要因配置モデル

$$Y_{ijk}=\mu_i+\alpha_j+\eta_{ij}+E_{ijk},\quad i=1,\ 2,\ \cdots,\ N$$
$$j=1,\ 2,\ \cdots,\ n$$
$$k=1,\ 2$$

のもとでの信頼性を吟味する。但しkは繰り返しを示し，ここでは最高2回である。η_{ij}は小論文と評定者との交互作用項を表し，μ_i，α_jおよびη_{ij}ともに全て変量効果(random effect)を表わす母数であるとする。すなわち，このモデルのもとでは，小論文も評定者もともに大きな母集団からの無作為標本と考えていることになる。このモデルのもとで，データIを分散分析した結果が表IV—14に示されている。この分散分析表より，評定者1人当りの1回当りの信頼性は

$$\hat{\rho}_{11}=\frac{(MS_S-MS_{SXR})/nr}{(MS_S-MS_{SXR})/nr+(MS_{SXR}-MS_E)/n'r+MS_E/n'r'}$$
$$=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(97.6-30.9)/2+30.9}$$
$$=0.37$$

である。但し，上式において，MS_{SXR}は小論文と評定者の交互作用の平均平方を示し，rは繰り返し測定の回数(ここではr=2)，r'は繰り返し測定何回当りの信頼性であるかを示すものである。従って，評定者3人による総合評価の平均の1回当りの信頼性は

$$\hat{\rho}_{31}=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(97.6-30.9)/6+30.9/3}$$
$$=0.64$$

となり，評定者3人で2回繰り返し測定したときの平均の信頼性は

$$\hat{\rho}_{32}=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(97.6-30.9)/6+30.9/6}$$
$$=0.70$$

ということになる。また，評定者の数を無限に増やしたときの総合評価値の平均の1回当りの信頼性は

$$\hat{\rho}_{\infty 1}=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(97.6-30.9)/\infty+30.9/\infty}$$
$$=1.0$$

であるのに対し，繰り返し測定を無限に行なったときの平均の評定者1人当りの信頼性は

$$\hat{\rho}_{\infty 1}=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(97.6-30.9)/2+30.9/\infty}$$
$$=0.53$$

ということになる。逆に，測定1回当りの信頼性が0.90になるためには

$$\hat{\rho}_{n'1}=\frac{(325.2-97.6)/6}{(325.2-97.6)/6+(17.6-30.9)/n'+30.9/n'}$$
$$=0.90$$

を解いて，$n'\geqq 15$を得る。すなわち，評定者が15人以上必要ということになる。

同様に，データIIに関して分散分析した結果が表IV—15に示されている。表IV—15より評定者1人当りの1回当りの信頼性は

$$\hat{\rho}_{11}=\frac{(185.7-51.0)/10}{(185.7-51.0)/10+(51.0-30.1)/2+30.1}$$
$$=0.25$$

となる。評定者5人による総合評価の平均の1回当りの信頼性は

$$\hat{\rho}_{51}=\frac{(185.7-51.0)/10}{(185.7-51.0)/10+(51.0-30.1)/10+30.1/5}$$
$$=0.62$$

となり，評定者5人で2回繰り返し測定したときの平均の信頼性は

$$\hat{\rho}_{52}=\frac{(185.7-51.0)/10}{(185.7-51.0)/10+(51.0-30.1)/10+30.1/10}$$
$$=0.73$$

となる。さらに，繰り返し測定を無限に行なったときの評定者1人当りの信頼性の極限は

$$\hat{\rho}_{1\infty}=\frac{(185.7-51.0)/10}{(185.7-51.0)/10+(51.0-30.1)/2+30.1/\infty}$$
$$=0.56$$

である。また，測定1回当りの信頼性が0.90になるためには，n'≧35となり，35人以上の評定者が必要ということになる。

以上の分散分析による結果の一部をまとめたのが表IV－16と表IV－17である。表IV－16より，小論文の書き手の個人差，および評定者の個人差を考慮した場合には小論文の信頼性はあまり高くないことがわかる。また，表IV－17からは，小論文の総合評価において十分な信頼性を確保するためには，かなりの数の評定者が必要であることがわかる。但し，総合評価の仕方をもっと統制し，管理すれば，これらの数値よりもっと好ましい結果を得る可能性が高い。

V　分析的評価に関する結果と考察

1．分析的評価の平均と標準偏差

第III章でも述べたように，各評定者（データIでは11名，データIIでは7名）は総合評価を終了した後，同じ小論文を順序を無作為にした上で7段階評定で，分析的に再評価した。その分析的評価のための観点は，先に示した表III－1の通りである。データIにおける分析的評価の平均は表V－1に，標準偏差は表V－2に示されている。15個の観点のうち平均値が最も低いのはA13（1人よがりな点がないかどうか）である。これは，読んでもらう努力や読み手に対する配慮などを評価するための観点であり，少なくともそのような点については高い評価が与えられていないということになる。表V－2より標準偏差の最も小さい観点はA10であることがわかるが，この観点は知識を評価するための観点であり，知識の豊富さ，内容・事実の正確さなどを評価するものであるが，この点についての個人差が小さいということは「旅」というような題目のもとでの小論文では知識に関する評価は困難であることを意味するものであろう。

データIIにおける分析的評価の平均は表V－3に，標準偏差は表V－4に示されている。（表中，横線――が引いてあるのは欠測値であることを示す。）データIのように観点A13の平均が低いかどうかは，データIIにおいてA13に相当する観点がないために不明であるが，標準偏差については観点A'10 aおよびbともに平均が小さく，「貧しさと豊かさ」という題目での小論文でもやはり知識について評価することは困難であることを物語っている。

2．分析的評価間の相関係数の分布

評定者を2人ずつ対にして分析的評価のための各観点ごとに相関係数の値を求めたところ，その分布は表V－5のようになった。表からも明らかにように，データIおよびデータIIともに，書かれた文字が丁寧であるかどうかに関する評価については評定者間で相関が高い。また，データIIにおいては，とくにA'5 aで相関がおしなべて高く，文体が一貫しているかどうかについての評価の評定者間相関が高いのに対して，データIではそのような傾向はみられない。また，データIで評定者相関の最も低い観点は，文法上の誤りや句読点の位置についての評価であるが，文法に関する評価の相関はデータIIでもやはり低い。

3．分析的評価の主成分分析

各評定者ごとに分析的評価の観点間の相関行列を主成分分析したところ，その主な主成分の寄与率および固有値が1.0以上の主成分の数は表V－6のようになった。但し，「平均」とあるのは表の数値を全ての評定者に渡って平均した値のことで，「プール」とある欄には全評定者に渡って観点間の分散共分散行列を合計しプールした上で相関行列になおして主成分分析を行なった結果の数値が示されている。表より明らかなように，第1主成分の寄与率は平均40％前後であり，第1主成分の寄与率は概してデータIの方が高い。また，固有値1.0以上の主成分の数はデータIIの方がデータIより多い。

表V－7および表V－8は，データIおよびデータIIの各々について，評定者ごとの観点間相関行列を主成分分析することによって求められた第1主成分の得点間の相関行列を示したものである。（第1主成分の固有ベクトルと固有値は表IV－4および表IV－5に与えられている。）言いかえれば，評定者ごとに分析的評価に重みをつけて合計し，一種の分析的評価の合計点を求めた上で，その合計点の評定者間相関を求めたものである。合計点といっても，第1主成分スコアであるので分散が最大となるように合計されており，よって信頼性も高いことが期待される。表V－7および表V－8における最右欄の「総合平均」は表IV－2(a)および表IV－2(b)の「平均」の欄に示されている総合評価の評定者間相関の評定者ごとの平均を示すものである。これを表V－7および表V－8の「平均」の欄に示された分析的評価の第1主成分スコア間相関の評定者ごとの平均と比較すると，データIおよびデータIIともに一般に後者の方が値が大きい。このことは分析的評価を組み合わせて1つの評価値とした方が，総合評価よりも評定

者間の一致度が高まることを意味している。

4. 観点間のプールされた相関行列の因子分析

まず評定者ごとに分析的評価のための観点間の分散共分散行列を求め，全評定者に渡って合計してプールされた分散共分散行列とした上で，それを相関行列になおした結果が表V－9（データI）および表V－10（データII）である。表の最右欄に「平均」とあるのは各行の数値の平均の値を意味している。データIにおいて，平均相関が最も高いのは観点A14の読後感に関するもので，平均相関が最も低いのは観点A3の文字に関するものである。データIIにおいて，平均相関が最も高いのは観点A'11の論理性に関するのとA'12の思考に関するもので，平均相関のもっとも低いのは観点A'3の文字に関するものである。いずれにせよ，データIおよびデータIIの双方において，文字や誤字・脱字に関する分析的評価は，他の観点についての評価と非常に低い相関関係にあることがわかる。

表V－9に示されたデータIにおける観点間のプールされた相関行列を因子分析した結果が表V－11である。共通性の推定値としてはSMCを用い，主因子解を求めた後にバリマックス回転を行なった結果である。表V－11に示された2つの因子の因子負荷から，15の観点が小論文のいわば読後感想的な評価に関する観点群（第I因子）と文法や用語力等の分析的評価に関する観点群（第II因子）の2つの観点群に分類できることがわかる。同様に，表V－10に示されたデータIIにおける観点間のプールされた相関行列を因子分析した結果が表V－12である。表V－12より，17個の観点が読後感想に関する観点群（第I因子），文法・用語に関する観点群（第II因子）および文の構成に関する観点群（第III因子）の3つの観点群に分類可能であることがわかる。

VI 総合評価と分析的評価

1. 総合評価と分析的評価との相関

データIにおいて，各評定者ごとに，総合評価の評価値と分析的評価の各観点についての評価値との間の相関係数の値を求めたところ表VI－1のようになった。それら相関係数のうち，最小値は.02，最大値は.66であった。観点別にみると，相関係数の平均値（表VI－1の右から2つ目の欄参照）が最も大きいのは，観点A14（読後感），次いでA15（親近感），A9（表現力）であり，総合評価と小論文を読んだ印象との間の相関が比較的高いことがわかる。また，総合評価との相関が平均的に最も低い観点はA1（誤字・脱字）であり，誤字・脱字がないかどうかは総合評価とはあまり相関がないということになる。なお，表VI－1において最右欄にプールとあるのは，まず評定者ごとに総合評価の評価値と分析的評価における各観点の評価値との間の共分散（および分散）を求め，それを全評定者に渡って合計してプールされた共分散（およびプールされた分散）とした上で相関係数になおした結果を示したものである。全体的には「平均」の欄の数値とその傾向は変わらないが，観点A12（思考力・判断力）における値が比較的高いのが注目される。なお，評定者ごとの相関係数の値の平均を見ると（表VI－1の最下欄参照）評定者R7の値が特に小さいが，これはこの評定者における分析的評価の評価値の標準偏差が小さい（表V－2参照）ことによるものであろう。

同様に，データIIにおいて，総合評価と分析的評価のための観点との間の相関係数を求めた結果が表VI－2である。表VI－2の数値は全般的に表VI－1の数値よりも小さく，相関係数の最小値は－.10，最大値は.72となっている。いくつか負の相関が見られるのは，各観点の評価値の信頼性が低いことによるものであろう。観点別に見て，相関係数の平均値が最も大きい観点はA'12（思考力・判断力）であり，この観点における相関係数の値はデータIでも小さくない。また，観点A'14（読後感）の相関係数の値も比較的大きく，その点についてはデータIと同じ傾向であるといえる。但し，データIIでは，観点A'9（表現力）における相関係数の値がデータIとは異なりそれ程大きくない。なお，表VI－2の最右欄のプールされた相関係数の各値を見ると，「平均」の欄の数値と比べて特に観点A'6（課題のとらえ方）における数値が高くなっている。また，データIIにおいては，データIよりも，評定者ごとの相関係数の平均値（表VI－1および表VI－2の最下欄参照）の差異が大きいのが特徴である。

2. 総合評価と分析的評価との重相関

総合評価を基準変数，分析的評価のための各観点を説明変数（データIでは15変数，データIIでは17変数）として各評定者ごとに重回帰分析を行なった。しかし，その結果得られた各評定者ごとの各観点についての偏回帰係数，または標準偏回帰係数にはしばしば負の値が見られ，解釈不可能であった。これは，前にも述べたように，各観点ごとの評価値の信頼性が低いためであると思われる。ただ，ここでは参考のために，重回帰分析の結果得られた重相関係数を表VI－3に示しておく。表VI－3からも明らかなように，重相関係数および調整済重相関係数ともに，データIIの方がおしなべて低く，重決定係数の値はいずれの場合も0.4を越えない。

3. 総合評価と分析的評価の第1主成分との関係

表VI—4および表VI—5は，データIおよびデータIIの各々において，評定者ごとに分析的評価の観点間の相関行列を，主成分分析することによって得られた第1主成分の固有ベクトルの要素の値と固有値の大きさを示したものである。表に示された固有ベクトルの要素の値は，第1主成分スコアを算出する際の，いいかえれば分散が最大となるように観点を重みづけ合計する際の，各観点に対する重みの大きさを示すものである。表VI—4の最右欄の「平均」の数値を見ることによって，データIにおいては，観点A14（読後感），A9（表現力），A12（思考力・判断力）およびA15（親近感）に対する重みが比較的大きいことがわかる。また，同様に，表VI—5から，データIIにおいては，観点A'11（論理性・一貫性），A'12，A'14に対する重みが比較的大きいことがわかる。

次に，分析的評価から求められた，評定者ごとの第1主成分スコアの時間的安定性を見たのが表VI—6である。これは1回目の分析的評価から得られた第1主成分スコアと2回目の分析的評価を1回目の測定で得られた固有ベクトルの要素（表VI—4と表VI—5を参照）で重みづけして合計したものとの間の相関をとったものである。いいかえれば，1回目の重みをそのまま用いて2回目のスコアを算出し，1回目のスコアとの相関係数の値を計算したもので，いわば分析的評価の第1主成分スコアの再検査信頼性と見なせるものである。表VI—6において，カッコ内に示されている数値は，総合評価の再検査信頼性の値である（表IV—7参照）。表からも明らかなように，総合評価の再検査信頼性の値の方が全般的にやや高いが，それほど大きな差異はない。

VII 討議

以上のデータ解析から得た結論を簡単にまとめると，以下のようになろう。

結論1　「総合評価の評定者間の一貫性は極めて低い。」

本研究では評価のための基準の統一などは一切行なわれず，各評定者はいわば極めて自由な状況下で評定者ごとに小論文の順序も異なるようにした上で小論文を評定するように求められた。したがって，この結論は小論文の評価に関して評定者間の一貫性を高めることが不可能であることを主張しているわけではない。ちなみに，石井（1981）によれば，立教大学の小論文試験のデータについては予想以上に評定者間の一致度が高かったということである。確かに，採点基準を明確にし，その基準に関する評定者達の理解を徹底させ，いくつかの小論文をサンプルとして実際に評価させて評定者の間でその結果について議論させる等の工夫をこらせば，評定者間の一致度を高めることは十分に可能であると考えられる。しかし，評定者間の一致度を高めるように工夫すること自体が適切であるかどうかは，測定の目的や出題内容等にかかわることである。

結論2　「総合評価の安定性は評定者によってかなり異なり，評定者内一貫性は高いとはいえない。」

この結論は再検査信頼性の値が評定者によってかなり異なること，および再検査信頼性の値がある程度以上の場合でも，1回目と2回目の評価値の平均値や標準偏差に差異が見られたことによって得られたものであるが，1回目の評価と2回目の評価との間の時間差がわずか1～2週間で，評価者の記憶がかなり明確に残っているであろうことを配慮すると，この結論はやや控え目に過ぎるかも知れない。この結論は，評定者間の一貫性の欠如を述べた先の結論1よりも重要である。それは評定者内一貫性が高くなければ，すなわち信頼性が高くなければ，評定者間一貫性も高くなり得ないということによるばかりでなく，小論文の評価には評定者間に差異があって当然であるという立場に立つとしても評定者内一貫性の欠如はとうてい是認することができないからである。評定者内一貫性を高めるためには，各評定者ごとに評価のための明確な基準を確立させること，評価すべき小論文の順序による系列的効果を生じないような工夫をすること，評価の日時の差異による影響を最小限にすること（たとえば，全ての小論文を同一日以内に評価してしまうこと），評価の疲労が蓄積されないようにすること，評価する際の環境が一定であること，等々の努力が必要であろう。大学入試等の場合には，一般に複数の評定者による評価値を平均（または合計）するという手続きがしばしばとられるが，結論2をより重視するならば，むしろ同一評定者に小論文の順序をランダムにした上で2度評定してもらい，得られた評価値を平均するといった手続きこそとられるべきであるということになろう。

結論3　「分析的評価のための諸観点は，少なくとも内容に関するものと言語力に関するものとに分類できる」

第II章の1の(1)でも述べたように，Remondino (1959) は，約20名の教師から得たデータによって作文の分析的観点として17個の観点を見出したが，彼はまたその観点ごとの評定者の評価値の平均値にもとづいた因子分析の結果も報告している。それによれば彼は，第1因子として「文字の美しさ」，第2因子として「用語力」，第3因子として

「内容と構成」，第4因子として「内容の個人的側面」を得たとしている。これに対し，本研究では，観点間のプールされた相関行列の因子分析によって，諸観点が少なくとも，内容に関するものと，文法や用語等の言語力に関するものとに分類されたが，これは明らかに，Remondino による結果と矛盾しない。すなわち，小論文の評価を分析的に見るならば，少なくとも内容評価と言語力の評価の2つから成っているということができる。

結論4　「分析的評価の第1主成分は，総合評価と同じ程度の信頼性をもつ」

評定者ごとに分散が最大となるように，分析的評価に重みをつけて合計すると，得られた小論文の評価値は，第V章の3で得られた総合評価値に関する結果よりも評定者間の一貫性がやや高くなる傾向があることが示された。それに対して，そのようなやり方で分析的評価を重みづけ合計した場合の評価値の時間的安定性は，第VI章の3の結果からもわかるように，総合評価の時間的安定性と同じ程度，あるいはやや低い。このことから，分析的評価の分散を最大とするような重みづけ合計は，総合評価と同じ程度の信頼性をもつことがわかる。但し，このことは必ずしも分析的評価が総合評価と同じ機能を持ちうることを意味している訳ではない。それは，総合評価と分析的評価との間の重相関係数の値が低いこと（第VI章の2を参照）からも明らかである。

また，分析的評価の合成値の信頼性を評価するにあたっては，分析的評価のための観点が15個以上もあることを配慮する必要がある。15個以上のそれぞれの観点について評価することの労力を別にしたとしても，一般に合成する変数が多ければ，信頼性は高くなることが知られており，その意味でも分析的評価の15以上の観点についての評価値の重みづけ合計が総合評価と同程度の信頼性をもつとしても，両者を対等に見なすことはできない。

結論5　「文字の丁寧さに関する評価は評定者間一貫性が高い」

分析的評価のうちで書かれた文字が丁寧かどうかについての評価は，評定者間相関が他の分析的評価のそれと比べて極めて高い（第V章の2参照）。但し，この文字の丁寧さに関する評価は，他の分析的評価と比べて，総合評価とはそれほど高い相関をもたない。このことは，論文の内容の評価というものを無視して，ただ単に評定者間の一貫性を高めるような評価を評定者に求めるならば，文字の丁寧さに関する判断が評価に大きく影響してしまう危険性があることを示している。

結論6　「総合評価は論文を読んだときの印象と深い関連がある」

総合評価は，分析的評価の中でも特に読後感との相関が高く（第VI章の1参照），Wiseman（1949）のいうところの General Impression の総合評価と深く関連していることをうかがわせる。但し，読後感に関する評価の評定者間相関は，データIで0.08～0.58，データIIでは0.08～0.33であり（第V章の2参照），評定者によっては全く意見の一致が見られない。前にも述べたように（第II章の4参照），Wiseman は，分析的評価が総合評価に比べて著しく時間とコストが掛かる反面，それほど信頼性を高めることが期待できないこと，総合評価が重要であること，および評定者間の一致よりも，評定者内の一貫性をより重視すべきことを主張したが，彼の研究から40年たって得られた我々の結論も，彼の主張を支持する方向にあるというべきであろう。

表IV－1　総合評価の平均値と標準偏差

データI	N	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	全体
平均	185	54.4	47.7	59.6	50.9	49.2	54.5	72.5	62.1	49.3	70.8	47.9	56.3
	165	53.8	47.0	59.7	50.1	48.7	53.8	72.3	61.8	49.2	70.4	47.3	55.9
SD	185	8.8	7.8	12.8	15.8	15.3	7.4	5.0	10.8	12.3	6.3	15.3	14.1
	165	8.0	7.5	12.8	15.7	15.1	7.0	4.7	10.2	11.9	6.3	15.1	14.0

データII	N	R'1	R'2	R'3	R'4	R'5	R'6	R'7	全体
平均	144	60.4	53.4	51.9	48.7	61.7	50.8	36.0	51.8
	124	59.2	53.5	52.1	48.7	61.6	49.7	34.0	51.3
SD	144	7.6	8.6	8.0	4.5	9.0	10.8	16.0	12.5
	124	6.4	8.4	7.5	4.6	8.5	10.0	14.7	12.3

表IV－2(a)　総合評価の評定者間相関行列（データI，N＝165）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
R 1		.36	.49	.45	.53	.52	.49	.57	.53	.37	.54	.49
R 2	.36		.40	.52	.27	.29	.41	.41	.30	.23	.42	.36
R 3	.49	.40		.39	.57	.49	.42	.53	.57	.45	.48	.48
R 4	.45	.52	.39		.31	.39	.42	.48	.39	.25	.48	.41
R 5	.53	.27	.57	.31		.50	.27	.48	.32	.56	.52	.43
R 6	.52	.29	.49	.39	.50		.43	.50	.51	.32	.49	.44
R 7	.49	.41	.42	.42	.27	.43		.45	.42	.25	.48	.40
R 8	.57	.41	.53	.48	.48	.50	.45		.55	.41	.56	.49
R 9	.53	.30	.57	.39	.32	.51	.42	.55		.22	.47	.43
R10	.37	.23	.45	.25	.56	.32	.25	.41	.22		.46	.35
R11	.54	.42	.48	.48	.52	.49	.48	.56	.47	.46		.49
											全体平均	.43

表IV－2(b)　総合評価の評定者間相関行列（データII，N＝124）

	R'1	R'2	R'3	R'4	R'5	R'6	R'7	平均
R'1		.05	.32	.35	.37	.09	.22	.23
R'2	.05		.20	.25	.23	.15	.27	.19
R'3	.32	.20		.42	.37	.14	.13	.26
R'4	.35	.25	.42		.58	.29	.26	.36
R'5	.37	.23	.37	.58		.24	.29	.35
R'6	.09	.15	.14	.29	.24		.19	.18
R'7	.22	.27	.13	.26	.29	.19		.23
							全体平均	.26

表IV－3　総合評価の評定者間相関行列の主成分の固有値と寄与率（データI）

主成分	固有値	寄与率	累積寄与率
1	5.39	0.49	0.49
2	1.10	0.10	0.59
3	0.85	0.08	0.67
4	0.62	0.06	0.72
5	0.57	0.05	0.78
6	0.53	0.05	0.82
7	0.46	0.04	0.87
8	0.43	0.04	0.91
9	0.40	0.04	0.94
10	0.37	0.03	0.98
11	0.26	0.02	1.00

表IV－4　総合評価の評定者間相関行列の主成分負荷（データI）

評定者	1	2
R 1	.77	-.01
R 2	.59	.43
R 3	.76	-.15
R 4	.65	.41
R 5	.70	-.51
R 6	.72	-.06
R 7	.65	.34
R 8	.78	.03
R 9	.69	.17
R10	.58	-.57
R11	.77	-.03

表IV－5　総合評価の評定者間相関行列の主成分の固有値と寄与率（データII）

主成分	固有値	寄与率	累積寄与率
1	2.63	0.38	0.38
2	1.05	0.15	0.53
3	0.88	0.13	0.65
4	0.82	0.12	0.77
5	0.63	0.09	0.86
6	0.58	0.08	0.94
7	0.41	0.06	1.00

表IV－6　総合評価の評定者間相関行列の主成分負荷（データII）

評定者	1	2
R'1	.58	-.49
R'2	.45	.59
R'3	.63	-.31
R'4	.79	-.09
R'5	.77	-.14
R'6	.44	.38
R'7	.52	.43

表IV－7 2回の総合評価の関係

	データ I			データ II				
	R1	R2	R8	R'1	R'2	R'3	R'4	R'5
1回目平均	53.8	47.0	61.8	59.2	53.5	52.1	48.7	61.6
2回目平均	54.0	46.6	56.2**	60.8**	49.6**	51.4	47.8*	55.8**
1回目S.D.	8.0	7.5	10.2	6.4	8.4	7.5	4.6	8.5
2回目S.D.	9.0	6.3	15.9	4.8	6.5	4.7	4.8	11.1
相関係数	0.91	0.81	0.71	0.49	0.40	0.41	0.51	0.81
	(N=165)			(N=124)				

表IV－8 推定された真値との相関（データ I）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
(イ) 平均値	.75	.58	.77	.69	.72	.69	.61	.77	.69	.57	.79	.69
(ロ) 中央値	.76	.42	.75	.56	.65	.62	.53	.68	.67	.51	.72	.62
(ハ) 重みづけ平均	.76	.57	.76	.68	.71	.70	.61	.76	.71	.55	.78	.69
(ニ) 変換重みづけ平均	.75	.58	.77	.69	.72	.69	.61	.77	.69	.57	.79	.69
(ホ) 再検査信頼性	.91	.81						.71				

表IV－9 推定された真値との相関（データ II）

	R'1	R'2	R'3	R'4	R'5	R'6	R'7	平均
(イ) 平均値	.50	.52	.55	.66	.69	.53	.69	.69
(ロ) 中央値	.44	.55	.64	.58	.48	.49	.39	.51
(ハ) 重みづけ平均	.45	.41	.52	.62	.56	.49	.66	.53
(ニ) 変換重みづけ平均	.49	.49	.55	.66	.65	.53	.70	.58
(ホ) 再検査信頼性	.49	.40	.41	.51	.81			

表IV－10 一要因配置のもとでの分散分析表（データ I）

変動要因	偏差平方和	自由度	平均平方
小論文	98152.0	164	598.5
残差	258019.0	1650	156.4
全変動	356171.0		

表IV－11 一要因配置のもとでの分散分析表（データ II）

変動要因	偏差平方和	自由度	平均平方
小論文	22695.0	123	184.5
残差	107813.4	744	144.9
全変動	130508.4		

表IV－12 繰り返しの無い二要因配置のもとでの分散分析表（データ I）

変動要因	偏差平方和	自由度	平均平方
小論文	98152.0	164	598.5
評定者	135672.0	10	13567.2
残差	122347.0	1640	74.6
全変動	356171.0	1814	

表IV－13 繰り返しの無い二要因配置のもとでの分散分析表（データ II）

変動要因	偏差平方和	自由度	平均平方
小論文	22695.0	123	184.5
評定者	59802.7	6	9967.1
残差	48010.7	738	65.1
全変動	130508.4	867	

表IV—14　繰り返しのある二要因配置のもとでの分散分析表（データI）

変動要因	偏差平方和	自由度	平均平方
小論文	53330.8	164	325.2
評定者	24815.2	2	12407.6
交互作用	31994.8	328	97.6
残差	15293.2	495	30.9
全変動	125434.0	989	

表IV—15　繰り返しのある二要因配置のもとでの分散分析表（データII）

変動要因	偏差平方和	自由度	平均平方
小論文	22839.9	123	185.7
評定者	25837.5	4	645.9
交互作用	25102.1	492	51.0
残差	18656.0	620	30.1
全変動	92435.5	1239	

表IV—16　各モデルのもとでの評定者1人当りの測定1回当りの信頼性（r：繰り返し数）

	データI	データII
一要因配置	0.20	0.04
二要因配置（r=1）	0.39	0.21
二要因配置（r=2）	0.37	0.25

表IV—17　各モデルのもとでの測定1回当り信頼性が0.90以上になるために必要な評定者数

	データI	データII
一要因配置	35人	231人
二要因配置（r=1）	14人	35人
二要因配置（r=2）	15人	35人

表V—1　分析的評価の平均（データI）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
A 1	4.2	3.3	3.7	4.7	3.5	5.4	5.9	4.1	5.4	4.4	4.8	4.5
A 2	4.3	4.0	4.0	4.3	3.7	4.3	5.9	4.0	4.3	4.3	4.3	4.3
A 3	4.2	4.4	3.6	3.7	3.7	4.6	5.9	3.8	4.8	3.9	4.4	4.3
A 4	4.3	4.5	3.9	4.1	3.4	4.4	6.0	3.7	5.2	4.5	5.2	4.5
A 5	4.4	4.3	4.2	4.6	3.9	4.7	6.0	3.9	4.1	4.4	4.7	4.5
A 6	4.2	4.4	4.4	3.9	3.8	4.7	6.0	3.9	4.5	4.2	5.4	4.5
A 7	4.3	4.1	4.8	4.0	3.9	4.0	6.0	4.3	4.3	4.0	4.2	4.4
A 8	3.7	3.5	4.1	4.0	3.5	4.0	5.9	3.7	3.9	4.3	4.0	4.1
A 9	4.2	4.3	4.2	3.7	3.6	4.0	5.8	3.8	4.0	4.0	3.9	4.1
A10	4.3	4.1	4.2	3.6	4.0	4.1	6.0	4.1	4.0	3.9	4.1	4.2
A11	4.3	3.7	4.0	3.7	3.7	3.8	5.9	3.6	3.9	3.8	4.0	4.0
A12	4.2	3.6	4.0	3.6	3.6	4.0	5.9	3.9	4.0	3.8	3.9	4.0
A13	3.7	4.1	3.9	3.8	3.8	3.9	5.8	3.5	3.8	3.4	3.7	3.9
A14	4.0	3.7	4.5	3.5	3.7	4.0	5.6	4.0	4.0	3.6	3.9	4.0
A15	3.9	4.1	4.4	3.5	4.1	4.2	5.9	4.0	4.2	3.8	4.2	4.2
平均	4.1	4.0	4.1	3.9	3.7	4.3	5.9	3.9	4.3	4.0	4.3	4.2

表V—2　分析的評価の標準偏差（データI）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
A 1	0.8	0.9	0.6	1.6	0.6	0.9	0.4	1.0	1.0	0.7	1.2	0.9
A 2	0.8	0.6	0.5	1.2	0.5	0.7	0.3	0.9	0.9	0.7	0.9	0.7
A 3	1.0	0.7	1.1	1.4	0.9	1.1	0.4	1.0	1.2	0.8	1.1	1.0
A 4	0.8	0.7	0.5	1.4	0.6	0.9	0.2	0.8	0.8	0.6	1.1	0.8
A 5	0.6	0.7	0.8	1.1	0.3	0.7	0.2	0.8	0.9	0.7	0.9	0.7
A 6	0.8	0.9	0.9	1.4	0.8	0.7	0.2	1.1	1.0	0.7	0.9	0.9
A 7	0.8	0.7	1.0	1.3	1.0	0.8	0.5	1.1	0.9	0.8	1.2	0.9
A 8	0.8	0.9	1.0	1.3	0.9	1.0	0.3	1.1	1.0	0.7	1.3	0.9
A 9	0.9	0.6	0.9	1.1	0.8	1.0	0.5	1.0	0.9	0.9	1.2	0.9
A10	0.6	0.3	0.5	0.9	0.3	0.4	0.2	0.8	0.1	0.7	0.3	0.5
A11	0.8	0.6	0.8	1.2	0.7	0.9	0.3	1.1	0.9	0.7	1.1	0.6
A12	0.7	0.7	0.8	1.0	0.7	0.6	0.3	0.9	0.7	0.7	0.9	0.7
A13	0.7	0.6	0.8	1.0	0.6	0.5	0.5	0.9	0.9	0.7	0.9	0.7
A14	1.0	0.7	0.8	1.1	1.1	0.8	0.6	1.0	0.9	0.7	1.3	0.9
A15	0.9	0.8	1.0	1.0	1.0	0.5	0.5	1.1	0.7	0.8	1.1	0.9
平均	0.8	0.7	0.8	1.2	0.7	0.8	0.4	1.0	0.9	0.7	1.0	0.8

表V—3　分析的評価の平均（データII）

	R'2	R'3	R'4	R'5	R'6	R'7	平均
A' 1	3.6	3.3	3.8	4.3	3.7	3.7	3.6
A' 2	3.1	3.3	4.0	4.4	3.3	3.2	3.5
A' 3	4.1	3.6	4.2	5.0	4.0	4.0	4.1
A' 4 a	3.6	3.7	4.0	5.0	3.4	3.4	3.9
A' 4 b	3.7	3.6	4.0	4.9	3.3	3.6	3.9
A' 5 a	3.8	3.8	4.0	5.0	3.7	3.8	4.0
A' 5 b	3.9	4.0	4.0	4.7	3.5	-	4.0
A' 6	4.0	4.0	4.0	4.9	3.6	3.4	4.0
A' 7	4.3	4.0	4.0	4.9	4.2	3.3	4.1

A' 8 a	3.9	3.4	3.8	4.1	3.5	3.3	3.7
A' 8 b	3.9	3.6	3.9	4.1	3.3	3.4	3.7
A' 9	3.8	3.9	4.0	4.8	3.7	3.4	3.9
A'10 a	3.7	4.0	3.6	4.8	4.0	3.2	3.9
A'10 b	3.6	3.5	3.8	4.8	3.9	3.1	3.8
A'11	3.2	3.3	4.0	4.8	3.2	3.1	3.5
A'12	3.4	3.5	3.5	4.6	3.5	3.0	3.6
A'14	3.5	4.0	3.8	4.6	3.7	2.9	3.7
平均	3.7	3.7	3.9	4.7	3.6	3.4	3.8

表V－4 分析的評価の標準偏差（データII）

	R'2	R'3	R'4	R'5	R'6	R'7	全体
A' 1	0.7	0.6	0.4	1.3	0.6	1.2	1.4
A' 2	0.8	0.5	0.2	1.2	0.6	0.7	0.7
A' 3	0.9	0.6	0.4	1.0	0.5	0.8	0.7
A' 4 a	0.6	0.5	0.0	1.1	0.6	0.7	0.6
A' 4 b	0.6	0.5	0.0	1.1	0.5	0.6	0.6
A' 5 a	0.5	0.5	0.2	1.3	0.6	0.6	0.6
A' 5 b	0.8	0.1	0.1	1.1	0.8	-	0.6
A' 6	0.2	0.0	0.7	0.8	1.0	0.8	0.6
A' 7	0.7	0.4	0.8	0.7	0.9	0.7	0.7
A' 8 a	1.0	0.8	0.8	1.2	0.7	0.7	0.9
A' 8 b	0.7	0.7	0.7	0.9	0.8	0.9	0.8
A' 9	0.4	0.4	0.3	0.9	0.6	0.7	0.6
A'10 a	0.6	0.4	0.8	0.8	0.4	0.7	0.6
A'10 b	0.5	0.5	0.8	0.8	0.4	0.6	0.6
A'11	0.8	0.5	0.7	1.0	0.7	0.8	0.8
A'12	0.6	0.6	0.8	0.9	0.7	0.6	0.7
A'14	0.6	0.3	0.6	0.7	0.7	0.6	0.6
全体	0.6	0.5	0.5	1.0	0.7	0.7	0.7

表V－5 観点ごとの評価値の相関係数の分布

データI	相関係数の分布範囲	平均	データII	相関係数の分布範囲	平均
A 1	0.03～0.57	0.36	A' 1	0.22～0.50	0.36
A 2	0.01～0.46	0.27	A' 2	0.04～0.34	0.20
A 3	0.28～0.77	0.58	A' 3	0.24～0.53	0.45
A 4	-0.02～0.41	0.19	A' 4 a	0.00～0.20	0.09
			A' 4 b	0.00～0.32	0.13
A 5	-0.10～0.48	0.24	A' 5 a	0.27～0.82	0.61
			A' 5 b	-0.26～0.27	0.06
A 6	-0.01～0.53	0.27	A' 6	-0.02～0.32	0.12
A 7	0.06～0.58	0.33	A' 7	0.04～0.35	0.19
A 8	0.15～0.52	0.32	A' 8 a	0.18～0.55	0.34
			A' 8 b	0.01～0.32	0.23
A 9	0.08～0.53	0.32	A' 9	-0.10～0.32	0.16
A10	0.00～0.53	0.29	A'10 a	0.07～0.49	0.31
			A'10 b	0.09～0.44	0.20
A11	-0.02～0.43	0.21	A'11	0.07～0.35	0.22
A12	0.04～0.47	0.23	A'12	0.15～0.37	0.26
A13	0.03～0.49	0.26			
A14	0.08～0.58	0.32	A'14	0.08～0.33	0.22
A15	0.02～0.50	0.25			

表Ⅴ－6　観点の主成分の寄与率の分布

	評定者	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均	プール
データⅠ	第1主成分寄与率	50	33	49	54	35	30	26	32	55	54	40	42	43
	第2主成分寄与率	8	10	8	9	11	12	10	13	9	11	11	10	9
	第3主成分寄与率	9	9	6	7	8	9	9	9	7	5	9	8	6
	第3主成分までの累積寄与率	66	52	64	70	54	51	45	54	70	70	60	60	59
	固有値1.0以上の主成分の数	3	3	2	3	5	4	5	4	3	2	4	3.5	2

	評定者	R'2	R'3	R'4	R'5	R'6	R'7 *	平均	プール
データⅡ	第1主成分寄与率	30	15	45	48	39	43	36	38
	第2主成分寄与率	9	11	10	10	9	9	10	9
	第3主成分寄与率	8	9	8	7	8	7	8	7
	第3主成分までの累積寄与率	46	42	64	65	55	59	55	54
	固有値1.0以上の主成分の数	6	8	4	3	5	4	5	4

＊はA'5 bを除く

表Ⅴ－7　第1主成分間相関（データⅠ）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均	総合平均
R 1		.55	.50	.55	.54	.51	.45	.56	.58	.45	.61	.53	.49
R 2	.55		.43	.47	.38	.48	.44	.35	.48	.33	.41	.43	.36
R 3	.50	.43		.46	.49	.54	.30	.52	.46	.45	.50	.47	.48
R 4	.55	.47	.46		.43	.56	.38	.52	.54	.22	.55	.47	.41
R 5	.54	.38	.49	.43		.55	.20	.55	.52	.48	.68	.48	.43
R 6	.51	.48	.54	.56	.55		.34	.62	.58	.45	.59	.52	.44
R 7	.45	.44	.30	.38	.20	.34		.32	.29	.20	.28	.32	.40
R 8	.56	.35	.52	.52	.55	.62	.32		.46	.35	.62	.49	.49
R 9	.58	.48	.46	.54	.52	.58	.29	.46		.29	.55	.48	.48
R10	.45	.33	.45	.22	.48	.45	.20	.35	.29		.40	.36	.35
R11	.61	.41	.50	.55	.68	.59	.28	.62	.55	.40		.52	.49

表Ⅴ－8　第1主成分間相関（データⅡ）

	R'2	R'3	R'4	R'5	R'6	R'7	平均	総合平均
R'2		.38	.41	.43	.22	.46	.38	.19
R'3	.38		.36	.43	.32	.39	.38	.26
R'4	.41	.36		.44	.21	.38	.36	.36
R'5	.43	.43	.44		.34	.52	.43	.35
R'6	.22	.32	.21	.34		.45	.31	.18
R'7	.46	.39	.38	.52	.45		.44	.23

表Ⅴ－9 観点間のプールされた相関行列（データⅠ）

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	平均
A 1		.39	.34	.40	.21	.18	.09	.21	.26	.16	.16	.19	.23	.20	.20	.23
A 2	.39		.40	.46	.41	.32	.37	.38	.55	.41	.34	.47	.35	.43	.39	.41
A 3	.34	.40		.34	.33	.28	.19	.37	.40	.25	.28	.31	.34	.33	.30	.32
A 4	.40	.46	.34		.44	.23	.22	.38	.42	.26	.28	.32	.33	.31	.29	.33
A 5	.21	.41	.33	.44		.34	.39	.45	.50	.28	.42	.41	.39	.43	.39	.39
A 6	.18	.32	.28	.23	.34		.35	.38	.33	.28	.35	.41	.50	.46	.48	.35
A 7	.09	.37	.19	.22	.39	.35		.40	.45	.34	.34	.52	.27	.65	.56	.37
A 8	.21	.38	.37	.38	.45	.38	.40		.49	.30	.51	.44	.47	.45	.44	.41
A 9	.26	.55	.40	.42	.50	.33	.45	.49		.47	.45	.53	.40	.53	.48	.45
A10	.16	.41	.25	.26	.28	.28	.34	.30	.47		.35	.43	.28	.35	.32	.32
A11	.16	.34	.28	.28	.42	.35	.34	.51	.45	.35		.52	.51	.44	.41	.38
A12	.19	.47	.31	.32	.41	.41	.52	.44	.53	.43	.52		.48	.62	.56	.44
A13	.23	.35	.34	.33	.39	.50	.27	.47	.40	.28	.51	.48		.51	.53	.40
A14	.20	.43	.33	.31	.43	.46	.65	.45	.53	.35	.44	.62	.51		.77	.46
A15	.20	.39	.30	.29	.39	.48	.56	.44	.48	.32	.41	.56	.53	.77		.44
															全体平均	.38

表Ⅴ—10 観点間のプールされた相関行列（データII）

	A'1	A'2	A'3	A'4 a	A'4 b	A'5 a	A'5 b	A'6	A'7	A'8 a	A'8 b	A'9	A'10 a	A'10 b	A'11	A'12	A'14	平均
A' 1		.31	.24	.19	.12	.07	.19	.14	.06	.13	.15	.22	.14	.15	.18	.17	.14	.16
A' 2	.31		.18	.39	.26	.23	.38	.29	.20	.22	.24	.41	.28	.31	.40	.32	.30	.30
A' 3	.24	.18		.15	.12	.08	.14	.08	.08	.15	.16	.16	.15	.11	.14	.11	.10	.13
A' 4 a	.19	.39	.15		.32	.20	.26	.30	.20	.14	.18	.30	.24	.28	.27	.26	.28	.25
A' 4 b	.12	.26	.12	.32		.20	.27	.22	.22	.26	.29	.35	.21	.18	.26	.24	.22	.23
A' 5 a	.07	.23	.08	.20	.20		.41	.26	.15	.12	.18	.35	.29	.23	.29	.30	.24	.23
A' 5 b	.19	.38	.14	.26	.27	.41		.37	.33	.21	.26	.48	.30	.26	.35	.37	.42	.31
A' 6	.14	.29	.08	.30	.22	.26	.37		.62	.30	.46	.43	.49	.50	.56	.58	.57	.39
A' 7	.06	.20	.08	.20	.22	.15	.33	.62		.25	.41	.35	.48	.44	.48	.55	.58	.34
A' 8 a	.13	.22	.15	.14	.26	.12	.21	.30	.25		.62	.31	.31	.32	.36	.32	.26	.27
A' 8 b	.15	.24	.16	.18	.29	.18	.26	.46	.41	.62		.44	.41	.43	.49	.47	.42	.35
A' 9	.22	.41	.16	.30	.35	.35	.48	.43	.35	.31	.44		.46	.43	.52	.44	.43	.38
A'10 a	.14	.28	.15	.24	.21	.29	.30	.49	.48	.31	.41	.46		.67	.54	.62	.57	.39
A'10 b	.15	.31	.11	.28	.18	.23	.26	.50	.44	.32	.43	.43	.67		.60	.59	.47	.37
A'11	.18	.40	.14	.27	.26	.29	.35	.56	.48	.36	.49	.52	.54	.60		.67	.56	.42
A'12	.17	.32	.11	.26	.24	.30	.37	.58	.55	.32	.47	.44	.62	.59	.67		.68	.42
A'14	.14	.30	.10	.28	.22	.24	.42	.57	.58	.26	.42	.43	.57	.47	.56	.68		.39
																	全体平均	.31

表Ⅴ—11　観点間のプールされた相関行列の因子分析（データⅠ）

		第Ⅰ因子	第Ⅱ因子	共通性
A14	読後感	.82	.22	.73
A15	親近感	.78	.21	.64
A 7	発想	.70	.12	.50
A12	思考力	.67	.33	.55
A13	一人よがり	.53	.39	.43
A 6	課題のとらえ方	.52	.26	.33
A11	論理性	.51	.37	.39
A 8	文の構成	.48	.44	.43
A10	知識	.38	.35	.27
A 4	文法	.19	.62	.42
A 2	用語力	.35	.61	.39
A 9	表現力	.50	.54	.54
A 1	誤字・脱字	.05	.52	.28
A 3	文字	.23	.52	.32
A 5	文体	.41	.48	.40
	因子寄与	4.02	2.71	

表Ⅴ—12　観点間のプールされた相関行列の因子分析（データⅡ）

		第Ⅰ因子	第Ⅱ因子	第Ⅲ因子	共通性
A'12	思考力	.77	.24	.17	.68
A'14	読後感	.72	.25	.09	.60
A'10 a	知識	.69	.22	.18	.56
A' 7	発想	.68	.11	.13	.49
A' 6	課題のとらえ方	.68	.24	.17	.55
A'10 b	知識	.66	.21	.22	.52
A'11	論理性	.65	.33	.25	.60
A' 2	用語力	.20	.58	.10	.39
A' 5 b	文体	.32	.55	.02	.40
A' 9	表現力	.40	.54	.22	.50
A' 4 a	文法	.20	.47	.04	.26
A' 5 a	文体	.24	.40	-.02	.22
A' 4 b	文法	.14	.40	.22	.23
A' 1	用字	.04	.37	.12	.15
A' 3	文字	.02	.27	.16	.10
A' 8 b	文の構成	.41	.20	.64	.62
A' 8 a	文の構成	.23	.18	.63	.49
	因子寄与	4.02	2.16	1.18	

表Ⅵ—1　総合評価と分析的評価との相関（データⅠ）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均	プール
A 1	.47	.21	.13	.29	.12	.02	.25	.17	.25	.28	.24	.22	.20
A 2	.62	.39	.42	.47	.29	.43	.34	.42	.63	.31	.37	.43	.39
A 3	.62	.34	.40	.39	.29	.32	.22	.38	.38	.25	.40	.36	.30
A 4	.49	.24	.25	.40	.17	.30	.16	.20	.48	.20	.27	.29	.29
A 5	.42	.27	.37	.47	.20	.35	.13	.33	.62	.34	.47	.36	.40
A 6	.49	.61	.43	.66	.55	.27	.25	.27	.60	.36	.40	.44	.48
A 7	.45	.52	.47	.39	.70	.47	.18	.30	.68	.40	.56	.47	.56
A 8	.55	.32	.46	.55	.48	.52	.22	.43	.58	.40	.48	.45	.44
A 9	.63	.50	.51	.54	.49	.45	.31	.37	.65	.43	.57	.50	.48
A10	.46	.21	.44	.49	.14	.49	.13	.40	.22	.34	.14	.31	.32
A11	.43	.46	.32	.51	.29	.53	.15	.31	.57	.35	.39	.39	.41
A12	.49	.29	.38	.54	.60	.46	.25	.40	.64	.35	.56	.45	.56
A13	.44	.43	.33	.52	.20	.39	.28	.37	.54	.35	.51	.40	.53
A14	.65	.53	.48	.61	.67	.48	.39	.32	.61	.34	.62	.52	.77
A15	.66	.55	.44	.66	.64	.46	.44	.34	.52	.34	.60	.51	.53
平均	.52	.39	.39	.50	.39	.40	.25	.33	.53	.34	.44	.41	.44

表VI－2 総合評価と分析的評価との相関（データII）

	R'2	R'3	R'4	R'5	R'6	R'7	平均	プール
A' 1	.07	-.10	.01	.28	-.02	.26	.08	.16
A' 2	.13	.10	.00	.42	.18	.29	.19	.23
A' 3	.15	.00	.26	.09	.07	.25	.14	.14
A' 4 a	.16	-.10	.00	.39	.20	.34	.17	.23
A' 4 b	.06	-.01	.00	.25	.07	.15	.09	.12
A' 5 a	.10	.14	.14	.40	.20	.18	.19	.21
A' 5 b	.08	-.02	.17	.55	.34	-	.22	.23
A' 6	-.05	-.03	.48	.65	.46	.45	.33	.40
A' 7	.18	.13	.47	.61	.34	.36	.35	.33
A' 8 a	.06	.00	.43	.32	.03	.20	.17	.15
A' 8 b	.09	.20	.46	.48	.34	.24	.30	.28
A' 9	-.01	.04	.22	.52	.12	.29	.20	.24
A'10 a	.27	.26	.45	.62	.23	.33	.36	.34
A'10 b	.33	.14	.36	.60	.12	.33	.13	.30
A'11	.20	.27	.40	.66	.27	.36	.36	.35
A'12	.23	.41	.46	.72	.43	.30	.43	.39
A'14	.19	.07	.46	.66	.39	.37	.36	.36
平均	.13	.09	.28	.48	.22	.29	.25	.26

表VI－3 総合評価と分析的評価間の重相関

	評定者	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
データI	重相関係数	.62	.63	.41	.56	.64	.64	.34	.43	.61	.27	.57	.52
	調整済み係数	.58	.59	.35	.51	.60	.61	.28	.37	.57	.20	.53	.47

	評定者	R'2	R'3	R'4	R'5	R'6	R'7	平均
データII	重相関係数	.22	.29	.39	.63	.34	.30	.36
	調整済み係数	.09	.17	.30	.57	.23	.20	.26

表VI－4 第1主成分の固有ベクトルと固有値（データI）

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	平均
A 1	.20	.13	.18	.17	.11	.03	.26	.12	.13	.20	.13	.15
A 2	.29	.28	.22	.26	.22	.26	.27	.33	.28	.22	.25	.26
A 3	.24	.27	.23	.19	.19	.24	.13	.26	.19	.18	.19	.21
A 4	.25	.19	.23	.23	.13	.24	.31	.20	.22	.22	.22	.22
A 5	.24	.22	.26	.26	.14	.24	.12	.24	.29	.28	.26	.23
A 6	.26	.24	.26	.26	.29	.17	.23	.10	.27	.26	.16	.23
A 7	.19	.30	.26	.21	.31	.20	.05	.27	.29	.28	.31	.24
A 8	.26	.22	.29	.26	.30	.26	.24	.25	.31	.27	.28	.27
A 9	.30	.33	.28	.28	.30	.32	.32	.30	.28	.29	.32	.30
A10	.20	.11	.25	.25	.12	.26	.24	.32	.05	.27	.14	.20
A11	.24	.20	.26	.27	.25	.32	.19	.27	.28	.30	.24	.26
A12	.26	.24	.27	.29	.35	.33	.30	.34	.29	.29	.32	.30
A13	.25	.30	.27	.28	.20	.27	.35	.22	.28	.24	.28	.27
A14	.32	.34	.30	.31	.36	.30	.34	.28	.30	.29	.33	.32
A15	.33	.34	.29	.32	.36	.29	.30	.23	.26	.27	.32	.30
固有値	7.54	4.97	7.41	8.14	5.22	4.44	3.88	4.77	8.25	8.15	6.01	

表VI－5　第1主成分の固有ベクトルと固有値（データII）

	R'2	R'3	R'4	R'5	R'6	R'7	平均
A' 1	.13	.04	.09	.15	.04	.08	.09
A' 2	.21	.13	.08	.22	.24	.22	.18
A' 3	.09	-.02	.23	.08	-.01	.13	.08
A' 4 a	.02	.02	.00	.19	.24	.25	.12
A' 4 b	.14	.20	.00	.18	.22	.17	.15
A' 5 a	.20	.19	.03	.21	.11	.11	.14
A' 5 b	.26	-.13	.06	.26	.30	—	.15
A' 6	.17	-.11	.32	.30	.31	.28	.21
A' 7	.26	.13	.30	.27	.26	.28	.25
A' 8 a	.21	.08	.29	.18	.18	.21	.19
A' 8 b	.25	.29	.32	.23	.27	.27	.27
A' 9	.26	.21	.18	.28	.26	.31	.25
A'10 a	.31	.31	.34	.27	.24	.31	.30
A'10 b	.25	.33	.33	.29	.18	.31	.28
A'11	.33	.46	.33	.29	.30	.32	.34
A'12	.33	.46	.34	.31	.33	.29	.34
A'14	.39	.31	.25	.27	.33	.28	.31
固有値	5.10	2.47	6.77	8.22	6.57	6.82	

表VI－6　分析的評価の第1主成分の再検査信頼性

データI	R1	R2	R8		平均
（総合評価	.76	.81	.77		.78
間相関）	(.91)	(.81)	(.71)		(.81)
データII	R'2	R'3	R'4	R'5	平均
（総合評価	.38	.39	.48	.69	.49
間相関）	(.40)	(.41)	(.51)	(.81)	(.53)

参考文献

Benton, S. L., & Kiewra, K.A. 1986 Measuring the organizational aspects of writing ability. Journal of Educational Measurement, 23 (4), 377-386

Blok, H. 1985 Estimating the reliability, validity and invalidity of essay ratings. Journal of Educational Measurement, 22 (1), 41-52

Breland, H. M. & Gaynor, J. L. 1979 A comparison of direct and indirect assessment of writing skill. Journal of Educational Measurement, 16 (2), 119-128

Brennan, R. L. 1983 Elements of generalizability theory. Iowa City : ACT Publications

Chase, C. I. 1968 The impact of some obvious variables on essay test scores. Journal of Educational Measurement, 5 (4), 315-318

Chase, C. I. 1979 The impact of achievement expectations and handwriting quality on scoring essay tests. Journal of Educational Measurement, 16 (1), 39-42

Chase, C. I. 1983 Essay test scores and reading difficulty. Journal of Educational Measurement, 20 (3), 293-297

Chase, C. I. 1986 Essay test scoring : interaction of relevant variables. Journal of Eductional Measurement, 23 (1), 33-41

Coffman, W. E. 1966 On the validity of essay tests of achievement. Journal of Educational Measurement, 3 (2), 151-156

Coffman, W. E. & Kurfman, D. A. 1968 A comparison of two methods of reading essay examinations. American Educational Research Journal, 5 (1), 99-107

Cooper, P. L. 1984 The assessment of writing ability : a review of research. GRE Board Research Report GREB No. 82-15R ETS Research Report 84-12

Cronback, L. J., Rajaratnam, N. & Gleser, G. C. 1963 Theory of generalizability : a liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163

Cronback, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972 The dependability of behavioral measurements. New York : Wiley.

Daly, J. A. & Dickson-Markman, F. 1982 Contrast effects in evaluating essays. Journal of Educational Measurement, 19 (4), 309-316

De Gruijter, D. N. M. 1980 The essay examination. In L. J. Th. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds) Psychometrics for educational debates, 245-262 : John Wiley & Sons Ltd.

Freedman, S. W. 1979 How characteristics of student essays influence teacher's evaluations. Journal of Educational Psychology, 71 (3), 328-338

Godshalk, F., Swineford, F. & Coffman, W. 1966 The measurement of writing ability. College Board Research Monographs, No. 6

Grobe, C. 1981 Syntactic maturity, mechanics and vocabulary as pridictors of quality ratings. Research in the Teaching of English, 15 (1), 75-85

Hales, L. W. & Tokar, E. 1975 The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117

Harris, W. H. 1977 Teacher response to student writing : a study of the response pattern of high school English teachers to determine the basis for teacher judgement of student writing. Research in the Teaching of English, 11, 175-185

Hiller, J. H., Marcotte, D. R. & Martin, T. 1969 Opinionation, vagueness and specificity-distinctions : essay traits measured by computer. American Educational Research Journal, 6 (2), 271-285

Hogan, T. P. & Misher, C. 1980 Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 17 (3), 219-227

Hughes, D. C., Keeling, B. & Tuck, B. F. 1980 The influence of context position and scoring method on essay scoring. Journal of Educational Measurement,

17 (2), 131-135
Hughes, D. C., Keeling, B. & Tuck, B. F. 1983a Effects of achievement expectations and handwriting quality on scoring essays. Journal of Educational Measurement, 20 (1), 65-70
Hughes, D. C., Keeling B. & Tuck, B. F. 1983b The effects of instructions to scorers intended to reduce context effects in essay scoring. Educational and Psychological Measurement, 43, 1047-1050
池田 央 1973 心理学研究法8 テストII. 東京大学出版会
石井 厳 1981 「論文試験とその評価」について 行動計量学, 8 (1), 22-29
国立大学入学者選抜研究連絡協議会 1983 実技検査・面接・小論文 大学入試研究の動向, 1, 21-22
国立大学入学者選抜研究連絡協議会 1984 実技検査・面接・小論文 大学入試研究の動向, 2, 49
国立大学入学者選抜研究連絡協議会 1985 実技検査・面接・小論文 大学入試研究の動向, 3, 46-48
国立大学入学者選抜研究連絡協議会 1986 実技検査・面接・小論文 大学入試研究の動向, 4, 49-50
国立大学入学者選抜研究連絡協議会 1987 実技検査・面接・小論文 大学入試研究の動向, 5, 62-64
Lord, F. M. & Novick, M. R. 1968 Statistical theories of mental test scores. Reading, Mass : Addison Wesley.
Marshall, J. C. & Powers, J. M. 1969 Writing neatness, composition errors and essay grades. Journal of Educational Measurement, 6 (2), 97-101
McColly, W. 1970 What does educational research say about the judging of writing ability ? Journal of Educational Research, 64 (4), 148-156
Meyer, G. 1939 The Choice of questions on essay examinations. Journal of Educational Psychology, 30 (3), 161-171
Peel, E. A. & Armstrong, H. G. 1956 The use of essays in selection at 11 Plus : The predictive power of the English composition in the 11 Plus examination. British Journal of Educational Psychology, 26, 163-171
Remondino, C. 1959 A factorical analysis of the evaluation of scholastic compositions in the mother tongue. British Journal of Educational Psychology, 30, 242-251
佐藤喜一 1987 63年受験用 傾向と対策19 国公立大2次・私立大の小論文 旺文社
Shanahan, T. & Lomax, R. G. 1986 An analysis and comparison of theoretical models of the reading-writing relationship. Journal of Educational Psychology, 78 (2), 116-123
Stewart, M. F. & Grobe, C. H. 1979 Syntactic matunity, mechanics of writing and teachers quality ratings. Research in the Teaching of English, 13, 207-215
Thorndike, R. L. 1982 Applied psychometrics, Boston : Honghton Mifflin
Van der Kamp, L. J. Th. & Mellenbergh, G. J. 1976 Agreement between raters. Educational and Psychological Measurement, 36, 311-317
Veal, L. R. 1966 Measuring writing improvement during an NDEA English Institute. Journal of Educational Measurement, 3 (4), 303-308
Willing, M. H. 1926 Individual diagnosis in written composition. Journal of Educational Research, 13 (2), 77-89
Wiseman, S. 1949 The marking of English composition in grammar school selection. British Journal of Educational Psychology, 19, 200-209
Wiseman, S. 1956 The use of essays in selection at 11 Plus : reliability and validity. British Journal of Educational Psychology, 26, 172-179
Wiseman, S. & Wringley, J. 1958 Essay-reliability : the effect of choice of essay-title. Educational and Psychological Measurement, 18 (1), 129-138

〈付 記〉

本研究は教育情報科学研究室の研究プロジェクトの1つである小論文研究の一環として行なわれたもので，このプロジェクトの参加者は筆者達の他に，芝祐順教授，平直樹，喜岡恵子，孫媛の人々がおり，本研究もいろいろな側面において，これらの人々の示唆や協力を得ている。また，非常に面倒な評価作業を行なって下さった方々にはとくに感謝の意を表したい。なお，本研究に要した費用は，昭和61年度財団法人カシオ科学振興財団研究助成「小論文の評価方法に関する研究」（研究代表者 芝祐順）および昭和63年度文部省科学研究費補助金，一般研究C「小論文テストによるデータの解析」（課題番号63510123 研究代表者 渡部洋）によって補助された。