日本語の文章に対する分析的評価の信頼性に関しての検証

田 島 ま す み

〈目　次〉

1．はじめに

言語能力の評価は，近年のコミュニケーション能力を重視する傾向から，口頭の発話や書かれた文章など実際の言語的産出物を直接評価するパフォーマンステストの重要性が高まりつつある．しかしながら，言語的パフォーマンスの直接評価は，特に評定者の判断で質やレベルを認定する場合，評定者に関する信頼性の問題を避けては通れない．

今年，2002年から実施されている日本留学試験（以下，日留試）の「日本語」において大きな改定があった．文章を書かせる記述問題は，今後出題される課題のタイプの拡充が図られるほか，採点基準が見直された．従来の基準では，文法的能力と論理的能力各３点の合計６点満点の採点で，これは他の日本語問題の合計が400点満点であるのに比べ，非常に低い配点であり，点数の差があまり出ないことから，能力判定の参考にするには問題があることが指摘されていた（村上他，2003）．これに対して，より多くの評価項目を立てて各項目の点数を合計する分析的評価基準の提案もなされていたが（小室，2004），今回の改定では分析的評価基準は採用されず，二つあった文法的能力・論理的能力という項目も消滅してレベル認定のための簡潔な説明がついた総合的評価基準に改定された．配点に関しては満点が50点となって，記述問題以外の日本語問題の総合点400点に対する比率は高くなっている．

なぜ分析的評価は採用されなかったのか．一般的には，単一スコアをつける総合的評価より複数項目の点数を合算する分析的評価のほうが信頼性において優れていると言われる．しかしながら，採点に時間や手間がかかることから大規模テストでは現実的に採用しにくいということはあるだろう．また，大規模テストの見本ともいえるTOEFLのライティング・テストは総合的評価基準を採用しており，その影響は否定できないものがある．改定の詳細を知る立場にはないので，推測の域を出ないが，分析的評価の利点である信頼性の高さになんらかの疑問が呈されたこともあったかもしれない．

2．分析的評価と総合的評価

分析的評価の欠点として，坂井（2005）は，外国語のテスティングを扱う文献の整理から，評価のための適切な下位項目の設定の困難さ，時間がかかること，場合によっては集計された評定順位が必ずしも全体的な印象と一致しないことがあるという３点を挙げている．最後に挙げられた全体的な印象と異なるおそれがあるという点は，その前の下位項目の設定の難しさと関わってくる．文章全体の評価をするには，文章全般にわたって取りこぼしなく分析評価できる下位項目を明確に設定しなければならない．しかし，実際の評価の場面では一般的によく使われている複数の項目が立てられ，その根拠となっているのは単なる経験であったり，ほかで使われていた評価基準に過ぎなかったりするのである．それがはたして全体を網羅するものであるのだろうか．一部分をことさらに取り立てて評価するような下位項目設定をすれば，「全体の印象と違う」という採点結果も起こりえることだろう．

さらに言語テストの場合，ある間違いがあった場合，それがどの項目の間違いであるのか判然としないという場合もある．『日本語テストハンドブック』（1991）には，一例として，学習者が「ねこ」を「れこ」と書いた場合，それが語彙の誤りなのか文字・表記の誤りであるのかは文章を見ただけではわからないという例を挙げている．この場合，評定者は，もし下位項目で「語彙」と「文字・表記」の項目があれば，どちらで減点あるいはレベルを下げるといったことを行えばいいのか迷う．どちらか片方を下げるか，両方を下げるかで評定に違いが出る．さらに，もし「語彙」と「文字・表記」で配点が異なればどちらか片方を下げる場合であっても，どちらを選択するかでまた違いが出る．

この例に限らず，たとえば「正確さ」と「適切さ」という，一見明快に対立して分かりやすいカテゴリーと思われるような項目であったとしても，実際の文章にあっては，意味・用法の理解が不正確で間違った使い方をしてい

るのか，適切な運用ルールへの違反で間違っているのか，判断できないといったことは起ってくる．そして，どの項目に相当するのかが不明瞭であれば，評定者によって採点結果が違ってくるという事態も起こりうるのである．分析的にしたからといって，単純に評価の信頼性は高まるものなのだろうか．

田中・長阪（2004）は，総合的評価法の理論的背景として「書かれたテキスト全体はその部分を全て足したものより大きい（Educational Testing Service：ETS）」という考えを紹介している．総合的評価の利点は単に時間がかからないということばかりに目が向きがちだが，むしろこの考え方を強調すべきだろう．総合的評価の採用は「負担軽減」悪く言えば「手抜き」のように考えられることが多いが，もっと根本のところで分析的評価とは違いがあるのである．前掲の論文で，田中・長阪（2004）は以下のような使い分けを述べている．

> ESL ライティング研究に評価表が使用される場合，ライティング能力を様々な側面から分析する目的では…（中略）…分析的評価方法が用いられ，ライティング全体を問題にしたり，特定の要素がライティング能力とどのように関係したりするのかを解明する目的では総合的評価方法が用いられることが多いと言える．（p.225）

本研究では，分析的評価と総合的評価の違いをあらためて検討することを試みる．信頼性において優位とされる分析的評価は，はたして実際にそうなのだろうか．信頼性の中でも，特に，複数の評定者がいた場合の評定者間信頼性に問題を絞り，実際の評価結果を比較して検証する．

3．研究方法概要

3．1．　資料

検証のための資料には，大学学部の一般教養科目「生命科学」で学期末に実施された記述試験の答案コーパスを利用した．試験問題は「薬剤耐性菌について世界の現状を述べよ」というもので，所定の解答用紙に手書きで記述を求めたものである．授業担当者が採点時に点数や下線などの書き込みをしているため，その影響が出ないよう文章評価には電子化したファイルを印刷したものを使用した．

このコーパスは田島他（2009）が行った別の研究に用いた資料で，その際，総合的評価で3段階の評定をつけたものである．その後，同様に総合的評価で6段階の評定を行った．いずれの場合も3人の評定者が評価した．コーパス全体は40人分の答案からなるが，今回は総合的評価で特に3人の評定が割れたもの10人分を選んで分析的評価の対象とした．

選び方の詳細を述べておく．まず，6段階の総合的評価で3人のつけた評定がそれぞれ6，3，4や3，2，4のように，全員が異なる点をつけたものは，答案40枚中14枚あった．そのうち，5枚は3段階評価では3人の評定者がすべて同じ点数をつけていたものであった．5枚の中で1枚だけ，3段階評価は一致したが6段階では4，1，5と大きく評価が分かれたものがあったのでそれは評価対象として残すこととし，残りの4枚を除外して，最終的に10枚を総合的評価で評定が大きく分かれたものとして抽出した．これら10枚の答案の書き手は全員，学部1年に在籍する日本語母語話者である．

本来，厳密な比較をするのであれば，総合的評価と同様，40人分のコーパス全体を分析的評価でも採点すべきであるが，時間や労力の制約からまずは小規模な調査を行って結果を出すことを優先した．ゆえに本研究は予備調査的な性質を帯びているものである．今回抽出した10枚の答案は総合的評価で

は評定者の評価が大きく分かれている．よって評定者間信頼性が特に低い答案であり，それらに対して分析的な評価をすることで評定者間信頼性が高くなるのかというのが今回の調査の研究課題である．

3．2．　採点基準

総合的評価と分析的評価に用いた採点基準表は，日留試の日本語記述問題の採点基準やこれまでの文章評価に関する研究などで用いられたものを参考に筆者が作成した．日本語では一般的に普及している評価のための統一基準というようなものはない．個々の試験や研究など状況に応じて基準が作成されて使用されているのが現状である．最近の文献を参考に，適切なものとなるよう注意して作成を行った．

3．2．1．　総合的評価基準

表1が総合的評価のための基準表である．当初は日留試の新基準をそのまま使うことも考えたが，その基準では今回資料とした文章の性質に合わない．その点を修正することで，今回の資料に合う基準表を作成した．

主な修正点は2点ある．まず，文章の種類に合わせた基準に変更した．日留試では，書き手が自分の意見をしっかりとした根拠とともに明確に主張できているかを主な観点としている．詳細は稿末に掲載した採点基準で確認できるが，主眼は「書き手の主張」である．しかし，今回の文章は意見を主張するものではなく，「現状を述べよ」という事実の記述である．それに沿う形で基準の説明を修正した．

もう1点は第一言語での産出であることへの調整である．日留試の採点基準は日本語学習者の書く文章を対象としている．母語話者が書いた文章を評価する場合，上位群に対しては天井効果が出てしまって弁別の機能が鈍くなる一方，下位群となる文章は該当数が少なくなってしまうことが予想される．今回の資料は日本語母語話者が大学の記述試験の答案として書いた文章である．日留試の採点基準では，レベルで6段階の評定がなされるが，特に下のレベルの基準説明は日留試よりもやや上のレベルとなるように変更した．

表１　総合的評価採点基準表

得点	基準
6	具体的な内容が秩序立てられ明確に述べられている．かつ，効果的な構成と洗練された表現が認められる．
5	具体的な内容が明確に述べられている．かつ，効果的な構成と適切な表現が認められる．
4	具体的な内容がおおむね明確に述べられている．かつ，妥当な構成を持ち，表現に情報伝達上の支障が認められない．
3	内容の具体性，記述の明確さにやや劣る点が認められる．あるいは，構成，表現に不適切な点が認められる．
2	内容の具体性や量に不足が認められる．あるいは，記述の明確さにおいて劣っている．構成や表現に複数の不適切な点が認められる．
1	内容の具体性や量が乏しい．あるいは，記述が明確ではない．構成に大きな問題点が認められるか，構成意識がない．表現に不適切な点が目立つ．

３．２．２．　分析的評価基準

分析的評価はどのような項目を立てるかということが重要になる．本研究では，田中他（1998）が因子分析の手法で抽出した作文評価の基本構造である４因子を評価項目の基礎とすることにした．４因子とは「正確さ」「構成・形式」「内容」「豊かさ」である．このうち,「構成・形式」を一括して「構成」に,「豊かさ」は内容的な豊かさと日本語の表現から見た豊かさが混同されないよう「表現」とすることで，より明確なカテゴリーとしたうえで評価項目とした．表２が本研究の分析的評価の観点である．

表２　分析的評価の４項目

項目	
内容	課題に対する答えとなる内容が，具体的，かつ明確に記述できているか
構成	答案として適切で明確な構成で記述されているか
表現	適切で多様な表現により十分な記述ができているか
正確さ	文法，表記，語彙，句読法，文体などの面において正確な日本語になっているか

項目の順番はさほど重要とは思われないが，基本4因子では「正確さ」「構成・形式」「内容」「豊かさ」となっていたものを若干変更した．念のため言及しておくと，これは，本研究の資料が母語話者の書いた文章であることを考慮したためである．4因子が抽出された研究（田中他：1998）では日本語学習者の書いた作文が対象であった．「正確さ」が第一の因子となっていることは，第二言語での産出を評価するのであれば妥当な順番である．しかしながら，本研究の資料は母語話者の書いた文章であり，「正確さ」に関して学習者ほど個人差は出ないだろうという予想のもと，「正確さ」は最終項目とした．一方，母語話者の文章であればより個人差があるのではないかと予想された「内容」を第一の項目とした．ただし，特定の項目に配点を多くする重みづけは採用せず，4項目ともすべて6点が最高の6段階評価とした．また，採点時の項目の順番も特に指示しなかった．

これらの項目に対し，それぞれの基準を作成した（表3～表6）．作成にあたっては，文章評価を扱った比較的最近の研究論文（小室他:2004，坂井:2005，田中他：2009，中尾：2009など）から得られた細かい採点基準や基準説明を参考にし，各項目の評定が総合的評価と同じ6段階になるよう設定した．各項目を採点したうえで，それらの合計点を当該答案の得点とした．

表3　「内容」に関しての基準

得点	基準
6	具体的な内容が秩序立てられ明確に述べられている．詳しく説明され，量的にも十分である．
5	具体的な内容が明確に述べられている．特に際立ってはいないが，十分な説明がある．
4	具体的な内容がおおむね明確に述べられている．部分的に不十分さを感じても，全体的な内容として課題に対しての十分な説明になっている．
3	内容の具体性，記述の明確さにやや劣る点が認められる．説明不足の部分や記述の偏り，内容のわかりにくい部分が見られる．
2	内容の具体性や量に不足が認められる．あるいは，記述の明確さにおいて劣っているなど，問題点がある．
1	内容の具体性や量が乏しい．あるいは，記述が明確ではなく，内容がわかりにくい．

表４　「構成」に関しての基準

得点	基準
6	適切な段落分けがなされ，段落間のつながりも整合性があり，バランスの良い構成になっている．書き出しや終わりの部分が適切である．
5	適切な段落分けがなされているが，行頭の１字あけなどの形式面での不備が見られる．段落間のつながりや全体のまとまりに問題はないか，あっても軽微なものである．書き出しや終わりの部分がやや不自然であったりする．
4	段落分けがなされているが，形式面での不備か，意味内容とのずれなど，不適切な部分がある．書き出しや終わりの部分が１文しかなかったり，本論部分の段落分けが粗かったりする．
3	段落分けがなされているが，不適切な部分があって，全体のまとまりに影響している．段落間でつながりの悪い箇所がある．あるいは，段落分けはなされていないが，文章の流れに問題はなくまとまった記述にはなっている．
2	段落分けがなされていなかったり，形式段落がつけられていても意味のまとまりに対応していなかったりする．段落間のつながりが悪い．文の流れや全体のまとまりに問題がある．
1	構成意識がない．形式的にも（形式段落）意味的にも（意味内容のまとまり）段落に相当するものがない．文と文のつながりが悪い部分が目立ち，全体がまとまっていない．

表５　「表現」に関しての基準

得点	基準
6	論述試験の答案として適切な表現・語彙が用いられて，十分に説明されている．語彙や表現の多様性が秀でている．
5	論述試験の答案として適切な表現・語彙が用いられているが，不適切なものが１，２箇所混じっている．多様な語彙や表現が使われている．
4	書き言葉の表現・語彙が用いられているが，話し言葉や縮約形など不適切なものが若干混じっている．語彙や表現は特に多様ではないが，劣ってはいない．
3	書き言葉の表現・語彙が用いられているが，話し言葉や縮約形など不適切なものが単純な間違い程度ではなく混じっている．語彙や表現にやや乏しさや単調さが認められる．
2	表現・語彙に不適切なものがあり，乏しさや単調さが認められる．漢字が使われるべきところで使われていないなど，漢字力の問題が認められる．
1	表現･語彙が乏しく，記述が不十分になっている．全体の文章が稚拙である．漢字の使用が少ないなど漢字力が劣っている．

表６　「正確さ」に関しての基準

得点	基準
6	誤りがなく，意味が明瞭に伝わる文章である．
5	誤りが１，２箇所あるが，軽微なものである．意味がわかりにくい部分はない．
4	誤りが数箇所あるが，軽微なものである．意味がわかりにくい部分はないか，あっても全体に影響はない．
3	誤りが数箇所あったり，意味のわからない部分があったりして，全体の正確さに影響がある．
2	誤りが数箇所あったり，意味のわからない部分があったりして，全体の正確さに問題がある．
1	誤りが目立ったり，意味がわからない部分があったりして，不正確な印象がある．

３．３．　評定者

総合的評価，分析的評価，ともに同じ３人の評定者が行った．３人とも大学レベルでの日本語教育に携わる教員である．以下，評定者Ａ・Ｂ・Ｃと記述する．なお，総合的評価は2010年４月，分析的評価は同年９月に行っている．同一評定者ではあるが，分析的評価時に総合的評価の記憶の影響はほとんどないと考えていいだろう．

４．結果と考察

４．１．　採点結果と評定者間信頼性

分析的評価による評定は表７のようになった．４項目，各６段階なので，合計点は24点である．表を見る限り，評価が分かれている印象は否めない．すべての文章で評価は割れ，同一の点数が３人のうち２人に出て１人だけ点数が違うということもない．３人はすべて異なる点数となっている．

表７　分析的評価得点（合計点）

ID	評定者 A	評定者 B	評定者 C
1	17	16	14
2	15	11	13
3	16	15	13
4	7	6	9
5	19	14	16
6	13	11	9
7	13	11	16
8	17	11	18
9	17	13	12
10	14	13	9
平均値	14.80	12.10	12.90

だが，総合的評価との比較のため，この点数を４で割り，総合的評価のスケールに合わせてみると，評価はそれほどずれない．表８が，その表である．各評定者の下に並ぶ数値は，分析的評価の合計点を４で割り，小数第１位で四捨五入したものである．平均１は小数点以下２桁まで示したもの，平均２は少数第１位を四捨五入した点数である．

表８　分析的評価得点（合計点 /4）

ID	評定者 A	評定者 B	評定者 C	平均1	平均2
1	4	4	4	3.92	4
2	4	3	3	3.25	3
3	4	4	3	3.67	4
4	2	2	2	1.83	2
5	5	4	4	4.08	4
6	3	3	2	2.75	3
7	3	3	4	3.33	3
8	4	3	5	3.83	4
9	4	3	3	3.50	4
10	4	3	2	3.00	3
平均	3.70	3.03	3.23	3.32	3

比較のため総合的評価を同様の表にしたものが表９である．両表を比べると分析的評価のほうが明らかに３人の評定者の点数が割れていない．３人が異なった点数となったのはIDナンバー８と10の２枚のみで，ほかは３人一致が２枚，２人が同じ点数で１人のみ１点違いの点数となったものが６枚であった．

さらに，評定者間信頼性を表す数値も計算した．Spearman-Brownの公式で求めた値は，総合的評価で0.27，分析的評価で0.89であり，分析的評価で著しく高かった．総合的評価で評価の割れたものを抽出して今回の評価対象としているので，総合的評価の信頼係数が低いのは当然と言えば当然のことであるが，評価方法を分析的にしただけで相関係数が0.89というかなり高い値となったことは予想外と言ってもいいぐらいの結果であった．

表９　総合的評価得点

ID	評定者 A	評定者 B	評定者 C	平均1	平均2
1	6	4	3	4.33	4
2	6	4	3	4.33	4
3	5	3	4	4.00	4
4	3	4	2	3.00	3
5	6	4	3	4.33	4
6	5	3	1	3.00	3
7	4	3	2	3.00	3
8	5	4	3	4.00	4
9	6	4	5	5.00	5
10	4	5	1	3.33	3
平均	5.00	3.80	2.70	3.83	4

参考にピアソンの相関係数の値も報告すると，総合的評価では，評定者A・B間が0.00，B・C間が-0.08，C・A間が0.59であったのに対し，分析的評価では評定者A・B間が0.83，B・C間が-0.31，C・A間が0.62であった．無相関検定の結果では，分析的評価の評定者A・B間でのみ１％の有意水準で無相関の帰無仮説が棄却されている．ほかの相関係数ではすべて５％の有意水

準でも無相関が否定されなかった.

これらの結果は，分析的評価は総合的評価に比べて信頼性が高いと言われる一般的な言説に対し，評定者間信頼性に関する部分で一定の根拠を与えるものである．二つの表の比較において特に目立つ違いは，評定者 A と C の結果である．総合的評価で評定者 A は全体的に見て高めの点数を出しているが，分析的に評価した場合はすべての答案において総合的評価の点数よりも低い点数が出ている(注)．平均で比べれば総合的評価の5.00に対し，分析的評価では3.70である．また総合的評価の場合に平均で最も低い点数を出していた評定者 C は分析的評価になった場合，2 枚のみ総合的評価点が分析的評価点よりも高かっただけで，6 枚で分析的評価が総合的評価の点数を上回り，残りの 2 枚は同一の点数であった.

分析的評価の効果とでもいうべき観点からまとめれば，分析的な評価方法は，総合的評価で高めの点数を出す傾向にある評定者に対しては点数を抑制する一方，低めの点数を出す評定者に対しては高めの点数を出させるという作用が認められた結果になっている．方法の項でも前述したが，分析的評価を実施したのは総合的評価の約 5 カ月後であり，評定者は総合的評価で出した点数は分析的評価時には見ていない．ゆえに，評定者 A が総合的評価では高めの結果だったので今回の分析的評価では低めにつけた，あるいは，評定者 C はその逆で今回，前に自分が低めの点数であったから高めにつけた，というような影響はあまり考慮しなくていいだろう．本研究の結果は，分析的評価は総合的評価よりも評定者間の評価の開きを抑制する効果があったことを示している.

4．2．　分析的評価下位項目での評定者間信頼性

総合的評価と分析的評価の比較をしたうえで，分析的評価の四つの評価項目におけるそれぞれの採点結果に関しても考察を加えておく．本研究で採用した項目は，「内容」「構成」「表現」「正確さ」である．それぞれの採点結果を表10から表13で示す.

評定者間信頼性を表すSpearman-Brownの公式を使った数値は，「内容」が0.87,「構成」が0.38,「表現」が0.59,「正確さ」が0.88であった．「内容」「正確さ」に関しては評価が一致しやすく，「構成」「表現」でやや評価が分かれるという結果である．文章評価に関する直感からいえば，形式的側面は比較的一致しやすく，意味的な要素が関わる場合評価が割れるのではないかといった予想が当然出てくる．今回，意味そのものである「内容」で評価が一致し，形式的側面とみなせる「構成」で評価が割れたという結果は意外にも思われる．しかし，本研究の資料が試験問題の回答であり，「薬剤耐性菌について世界の現状を述べよ」という，ある程度決まった内容の記述であったことを考えれば，「内容」の評価でそれほど違いが出なかったのは納得がいく．

注意を払うべきは「構成」である．文章を評価する時にはよく使われる項目であり，一見，明快なものであるような印象を与える．しかし，実際には「構成」は，単純にとらえることのできない，なかなか複雑な事象である．形式的要素と意味的要素が双方ともに関わってくるからで，この二つを「構成」というひとくくりの枠で評価するのはなかなかに困難なことである．

形式的要素のみに限定すれば，長い文章が適度な長さの段落に分けられ，段落分けの約束に従った形式段落がしっかりつけられているかどうかを判定すればいい．いわば表記上の，段落に関してのルールが守られているかという観点である．段落が全くついていない文章は最低点となり，形式段落はついていて改行はしてあるものの段落冒頭の１字あけができていないものは，最低点よりも点数は高くなるだろう．

しかし，「構成」は形だけのものではない．意味的要素を考慮するとまた違う側面を評価することになる．具体的に挙げられる点としては，文章が意味内容のまとまりで段落分けされているかということと，全体が序論･本論･結論といった一つのまとまりを持った文章として明快な展開を持っているかということがある．たとえば，上記の形式段落の例を使ってさらに付け足すと，全く形式段落が付いていない文章ではあるが序論・本論・結論といった構成はできているものと，形式段落はついているが意味内容に対応しない段

落をつけていたものとを比べた場合，どちらが高い評価になるのか，あるいははたして優劣がつけられるのか，という問題である．

今回の採点基準にもこのような難しさが反映している．表４に「構成」に関しての採点基準を示しているが，作成時には苦心して複雑な要素を入れ込んだ．整理すれば，形式的な段落分け，段落間のつながり，全体のまとまりや文の流れ，書き出しや終わりの意識という，四つのさらなる下位分類を設定することが可能である．評定者の採点結果が割れた原因は複数の要素が入り込んでいる判定しにくい採点基準と考えられる．本研究では作文評価の基本４因子を評価項目の基礎としたために「構成」を１項目とした．しかし，評定結果のばらつきを見る限り，今後，分析的評価をする際には「構成」を１項目としてまとめることが適切であるかという問題は検討が必要であろう．

表10　分析的評価「内容」に関しての得点

ID	評定者 A	評定者 B	評定者 C
1	3	3	3
2	3	2	3
3	5	4	3
4	1	2	2
5	6	3	5
6	3	2	2
7	2	2	1
8	4	3	4
9	5	4	4
10	3	4	1
平均	3.5	2.9	2.8

表11　分析的評価「構成」に関しての得点

ID	評定者 A	評定者 B	評定者 C
1	5	3	2
2	2	2	2
3	3	3	2
4	1	2	2
5	4	2	3
6	3	2	1
7	2	1	4
8	4	1	5
9	5	2	4
10	1	1	1
平均	3.0	1.9	2.6

表12　分析的評価「表現」に関しての得点

ID	評定者 A	評定者 B	評定者 C
1	4	4	4
2	5	2	4
3	3	3	4
4	3	1	2
5	5	4	4
6	3	3	3
7	4	3	5
8	4	2	5
9	4	2	2
10	5	2	4
平均	4.0	2.6	3.7

表13　分析的評価「正確さ」に関しての得点

ID	評定者 A	評定者 B	評定者 C
1	5	6	5
2	5	5	4
3	5	5	4
4	2	1	3
5	4	5	4
6	4	4	3
7	5	5	6
8	5	5	4
9	3	5	2
10	5	6	3
平均	4.3	4.7	3.8

5．おわりに

本研究の研究動機としては，分析的評価への一般的な信頼・評価に対して筆者の個人的な疑念があった．大規模テストの文脈でない場合，文章の評価は分析的にするものという観念が一般的であり，日本語教育に携わる専門家であってもそのように考えている．それに対して何らかの反証が得られないものだろうかという疑問が研究の出発点であった．根底には「全体は部分の合計ではない」という思想がある．

しかし出てきた結果は，むしろ一般的な言説を強化するものとなった．もちろん小規模の検証であって制約は多分にある．前述したように，総合的評価をした資料から抜き出したものだけを分析的評価の対象としたことで，2種類の評価方法の比較としては条件がそろっていない．完全な比較ではないものの，本研究の結果は分析的評価の評定者間信頼性に対する数値的な裏付けとして提示したい．

信頼性の高さは示された．一方で，分析的評価の短所とされる，採点に時間のかかる点や採点基準の精度の影響も確認された．総合的評価で判断に迷

う部分が分析的評価では分割されて短時間で判定ができるのではないかということも予想されたが，その効果はあまり認められなかった．また，明確な基準でなければ評定が割れてしまうのは評価項目「構成」のところで述べたとおりである．

重要なことは，総合的・分析的評価，双方の特質を理解していることだと思われる．評価が行われる状況に合わせて，その特質に応じた選択をするというのが妥当な結論ということになろう．そうであるとすると，文章評価基準はその時々に応じて作成するしかないということになり，一律の評価基準は意味がないということになる．しかしながら，日本語のライティングを評価する際，信頼できる統一基準がほしい，あれば便利だというのも実際の日本語教育に携わる者からの現実的かつ切実な意見である．状況に応じた修正を簡単に施せる応用可能な基本基準のようなものが作成されるとよいのではないだろうか．

（注）

ID ナンバー10について注記しておく．表 8 ・ 9 では，総合的評価・分析的評価ともに 4 点となっているが，総合的評価の合計点は14点であり， 4 で割った点は3.5である．表 8 で 4 点となっているのは四捨五入で整数表示をしたことによるものであり，実際には分析的評価点のほうが総合的評価点より低くなっている．

参考文献

1）小室輝代・三谷閑子・村上京子（2004）「日本留学試験『記述問題』の評価基準の提案とその信頼性」『言語と文化』 5 , 55-70
2）衣川隆生（2005）「主観的・統合的な作文評価結果と相関関係を持つ分析的・客観的な量的指標の抽出」『筑波大学留学生センター日本語教育論集』20, 35-44
3）国立国語研究所編（2006）『世界の言語テスト』 くろしお出版
4）坂井美恵子（2005）「学部教員による留学生の作文評価―総合的評価の分析―」『大分大学留学生センター紀要』 2 , 19-30
5）田島ますみ・深田淳・佐藤尚子・玉岡賀津雄（2009）「語彙指標数値と文章主観評価の関係――日本人大学生による 2 種類の書き言葉コーパスを使った

実証研究――」『中央学院大学人間・自然論叢』29，57-77
6）田中真理（2005）「日本語教育におけるライティング評価」国立国語研究所編『日本語教育年鑑2005年版』42-52
7）田中真理・坪根由香里・初鹿野阿れ（1998）「第二言語としての日本語における作文評価基準――日本語教師と一般日本人の比較――」『日本語教育』96，1 -12
8）田中真理・長阪朱美（2004）「日本語と英語を目標言語とするライティング評価基準の展望：第二言語としての日本語のライティング評価基準作成に向けて」『第二言語としての日本語の習得研究』7 , 214-253
9）田中真理・長阪朱美・成田高宏・菅井英明（2009）「第二言語としての日本語ライティング評価ワークショップ――評価基準の検討――」『世界の日本語教育』19, 157-176
10）中尾桂子（2009）「語彙の統計量と総合評価の関係――作文評価の基準特定にむけて――」『大妻女子大学紀要―文系―』41, 129-146
11）日本語教育学会編（1991）『日本語テストハンドブック』 大修館書店
12）三谷閑子・村上京子・小室輝代（2004）「作文の評価手順が評価に及ぼす影響について――analytic scoring の採点に関して――」『言語と文化』5，1 -16
13）村上京子・小室輝代・三谷閑子（2003）「日本留学試験『記述問題』における採点基準の見直し」『名古屋大学日本語・日本文化論集』11, 107-124
14）Sasaki, M. & Hirose, K.（1999）Development of analytic rating scale for Japanese L1 writing. Language Testing 16, 457-478
15）日本留学試験　http://www.jasso.go.jp/eju/index.html （2010年 9 月19日）
16）TOEFL　http://www.ets.org/toefl （2010年 9 月19日）

参考資料

日本留学試験　日本語記述問題　新採点基準

得点	基準
50点	（レベル S） 課題にそって，書き手の主張が，説得力のある根拠とともに明確に述べられている．かつ，効果的な構成と洗練された表現が認められる．
45点 40点	（レベル A） 課題にそって，書き手の主張が，妥当な根拠とともに明確に述べられている．かつ，効果的な構成と適切な表現が認められる．
35点 30点	（レベル B） 課題にほぼそって，書き手の主張が，おおむね妥当な根拠とともに述べられている．かつ，妥当な構成を持ち，表現に情報伝達上の支障が認められない．
25点 20点	（レベル C） 課題を無視せず，書き手の主張が，根拠とともに述べられている．しかし，その根拠の妥当性，構成，表現などに不適切な点が認められる．
10点	（レベル D） 書き手の主張や構成が認められない．あるいは，主張や構成が認められても，課題との関連性が薄い．また，表現にかなり不適切な点が認められる．
0点	（NA） 採点がなされるための条件を満たさない．

※　レベル A，B，C については，同一水準内で上位の者と下位の者を区別して得点を表示する．