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—

三谷閑子・村上京子・小室輝代

キーワード　*作文評価、採点手順、analytic scoring、項目の独立性、評価の妥当性*

0. はじめに

2002年から実施され始めた日本留学試験には、従来の日本語能力試験にはなかった「記述問題」が導入されている。日本の大学への留学を希望する受験生が、大学の教養課程や専門課程で要求される文章を書く力をどの程度持っているかを測ることは重要であり、このような記述問題が導入されるようになったことには非常に意義があると言える。しかし、その評価方法には、評価基準を含めなお検討を要する点があると考えられる。

そこで、現在、評価結果をより活かせるような新しい評価基準作り[1]を試みているが、その信頼性を検討していく過程で、評価者間の採点手順（評価ストラテジー）の違いが評価の結果に影響を及ぼしているのではないかという疑問が出てきた。本研究は、この疑問に基づき、作文の評価を行うときに評価者の採点手順の違いが評価にどのように影響を及ぼしているかを実証的に調べ、最終的には評価の信頼性と妥当性、及び評価者トレーニングへの示唆を得ることを目的とする。

評価ストラテジー[2]は、評価を決定するまでの評価者の思考と行動過程、作文の読み方、判断を下す時期など様々なものを含む。作文の評価方法にはおおまかに分けてholistic scoringとanalytic scoringの２つ[3]があるが、本研究では特にanalytic scoringの場合の採点過程[4]に焦点を当てる。

1. 先行研究と本研究の位置づけ

作文評価者の評価ストラテジーは比較的新しい研究分野である。特に、実証的な研究は1990年代に始まったがその数はまだそれほど多くはない。これまで

に行われたもののなかにはCumming（1990）、Vaughan（1991）、 Huot（1993）、Pula & Huot（1993）によるholistic scorerの評価ストラテジーの研究があり、評価者がそれぞれ異なるストラテジーを使うことが指摘されている。また、最近の動向としては、Wolfe & Ranney（1996）、Wolfe（1997）が評価者の熟達度別にholistic評価ストラテジーの違いを分析する研究を進めている。しかし、analytic scorerの評価ストラテジーについての研究はほとんどなく、これまでに筆者が調べた範囲では、DeRemer（1998）、Lumley（2002）だけだと思われる。日本語教育においてもまだこの分野の研究は進められていない。

DeRemer（1998）は、評価者トレーニングの指導にあたったこともある経験豊かなanalytic scorerの評価作業の過程をthink aloud法によって調査した。そして、評価者が１つの観点について評価するときの評価作業のやり方には、general impression scoring、text-based evaluation、rubric-based evaluationの３つ[5]があったと報告している。DeRemerは、これら３つのやり方は非常に異なるので、評価者がこの中のどれを使うかが評価の妥当性に影響してくるのではないかと指摘する。また、調査では評価者間の一致度が低かったために、今後評価者トレーニングの内容を検討する必要があると述べている（p.26）。この研究からは、評価の観点と基準を具体的に詳細に設定した、言い換えると、holistic scoringと比べて評価者の自由をかなり制限したanalytic scoringの場合にも評価者間に異なるストラテジーが存在し、それが評価に影響を与える可能性があるのではないかと考えられる。

このような研究背景を踏まえ、本研究ではanalytic scoringの際の採点手順を評価者のストラテジーの１つとして捉え、具体的に３つの採点手順を設定し、手順を設定することが評価にどう影響を及ぼすか、また、それぞれの手順の違いが評価に影響を及ぼすかどうかを調べる。さらに、そのような条件下で評価者間にどのようなストラテジーの違いがあるかも見る。先に述べたように、これまでanalytic scorerの評価ストラテジーに焦点をあてた研究はほとんどないため、本研究の視点には意義があると考えられる。

2．研究課題

Ⅰ．analytic scoringの際、評価者間で評価ストラテジーにどのような共通点と相違点があるか。

Ⅱ．採点手順と評価の間にどのような関係があるか。
採点手順と評価の関係は以下の３つの観点から考察する。

① 採点手順によって評価、評価の一致度、採点時間は異なるか。
② 採点手順の違いは各評価項目の独立性に影響を与えるか。

Ⅲ．評価に与える影響において評価者の個人差と条件差のどちらが大きいか。

本研究では、以上の研究課題を２度の調査を通して考察する。調査１ではⅠを扱い、調査２ではⅡとⅢを扱う。これ以降の本稿の構成は以下の通りである。

＜論文の構成＞

- 日本留学試験記述問題の採点基準の見直しについて
- 調査について
- 調査１の結果と評価者の内的ストラテジーについての考察(研究課題Ⅰ)
- 調査２の結果と分析
- 調査２の結果に基づく採点手順と評価の関係についての考察（研究課題Ⅱ）
- 条件差と評価者の個人差が評価に及ぼす影響（研究課題Ⅲ）
- 結論と示唆
- 今後の課題

3．日本留学試験「記述問題」の採点基準の見直しについて

2002年から始まった日本留学試験「記述問題」の評価基準は、評価結果やそれを指導に活かしていく上でいくつかの問題点がある。その中でも特に問題であると考えられるのは、評価の観点項目が少ないために学習者の書く力を十分に測定できていない点や、学習者間にあまり差がつかない採点結果が出ている点である。

このような問題点を改善するために、名古屋大学では新たな採点基準を作成する試みが行われている。新基準では７つの評価観点項目から学習者の書く力を測定する。採点基準の見直しの詳細については、村上その他（投稿中）と小室その他（投稿中）を参照していただきたい。

4. 調査

本研究では調査を２度行った。以下に調査の順番に従って手続きを説明する。

4－1. 調査の手続きと評価者

4－1－1. 調査1とデータ

調査１では、主に評価者の内省に注目し、評価者間で採点ストラテジーに違いがあるかどうかを調べた。この調査は研究課題Ⅰに対応するものである。データは、外国語使用についての意見文（平成14年度日本留学試験「記述問題」）[6]20人分を用いた。

調査の手続きとしては、データを２等分し、10枚を項目別採点方式、残りの10枚を個人別採点方式で採点した。項目別採点方式とは１つの項目ごとに全員分採点していくやり方で、個人別採点方式は１つの作文ごとに全項目を採点していくやり方である。そして、採点後に評価者が集まって採点中に気づいたことを話し合い、それをテープに録音した。

4－1－2. 調査2とデータ

調査２では採点手順と評価の関係に焦点を当てる調査を行った。これは研究課題ⅡとⅢに対応するものである。データは外国語使用についての意見文（平成14年度日本留学試験「記述問題」）の66人分を用いた。この中には調査１で使った作文は含まれていない。

調査２ではデータを３等分し、22枚づつをそれぞれ、項目別採点方式、個人別採点方式、この２つを合わせた折衷方式の手順で採点した。折衷方式とは、７つの項目のうち「正確さ」、「文体」、「語彙の多様性」を項目別採点方式で採点し、残りの「文のわかりにくさ」、「文間」、「段落間」、「内容」は個人別採点方式で行うやり方である。

この調査では、評価者間の評価の一致度、採点時間、さらに評価観点項目の項目間相関を調べた。

4－1－3. 評価者

評価者については、どちらの調査も同じ３人の評価者が採点をした。３人はいずれも日本語教育に関わっている。採点を行った時と場所は評価者によって異なる。

5．調査１の結果と評価者の内省に基づく評価者の内的ストラテジーについての考察

調査１では、項目別方式と個別方式で採点した後、評価者が自分の採点の仕方や気づいたことを話し合い、テープに録音した。以下は調査前に評価者が用いていた採点手順と調査で設定した採点手順別に評価者の内省をまとめたものである。

5－1．調査前の採点方式

調査以前に各自が自由なやり方で採点していたときは、３人の評価者のうち２人が個人別採点方式、１人が項目別採点方式を使っていたことがわかった。また、個人別採点方式を採っていた２人の間では、採点していく項目の順番が異なっていた。項目の順番は、１人は正確さに関わる項目から始めて内容に関わる項目へ、もう１人は、最初に気づいた間違いや問題点に印だけつけた後、内容に関わる項目から正確さに関わる項目の順に採点していた。

5－2．調査時の採点方式についての内省

1）項目別採点方式

項目別採点方式について評価者全員に共通していたのは次の２点である。まず、３人とも今回の調査で指定したような１つの項目ごとに全員分の作文を採点していく項目別採点方式はやや採点しづらく感じていた。それは、先に述べたように、調査以前に各自が自由なやり方で採点していたときには、２人が個人別採点方式をとっていたが、その際、２人とも文章の意味に関わる「文のわかりにくさ」、「文間」、「段落間」、「内容」といった互いに関連のある項目はまとめてつけていたからである。項目別に採点していた残りの１人も、すべての項目を１つ１つ別々に見るのではなく、意味に関わる関連項目はやはり一緒につけていた。次に、３人とも作文を読む回数がかなり多く、１つの作文につき７回から多い場合にはその倍くらい読んでいた。また、そのために採点に時間がかかっていた。

一方、３人の間で異なった見解があったのは採点していく項目の順番である。２人は調査で設定した順番で問題がなかったとし、１人はややややりにくかったとしていた。

このような共通点と相違点以外には、「内容」を採点するときに前の作文との相対的な比較で点数をつけてしまうことがあった、採点の判断に揺れが少な

かった点がよかった、誤用をどの項目で扱うかが個人別方式のときとは異なった、などの指摘があった。

2）個人別採点方式

個人別採点方式では、全員が最初に全体を読みながら間違っている箇所に印をつけ、正しい文に直すことから始めている。しかし、その後の手順は評価者によって異なっていた。評価者Aの場合は、次に「正確さ」を見、その後は「文体の統一」と「語彙の多様性」、「文のわかりにくさ」と「文間」をそれぞれ一緒に採点してから、最後に「段落間」と「内容」の順番で採点する。評価者Bの場合には、次に「内容」を採点し、3回から4回程度読んで内容の採点結果の妥当性を確認しつつ、さらに「わかりにくさ」、「段落間」、「文間」、「文体の統一」、「語彙の多様性」を一緒につけ、最後に「正確さ」を採点している。評価者Cは、次に「文体の統一」を採点し、それから再読して「内容」と「段落間」を一緒に採点する。その後、「正確さ」、「語彙の多様性」、「文のわかりにくさ」、「文間」の順に採点していくやり方をとっている。このように、採点する項目の順番は3人の間で違いが見られた。

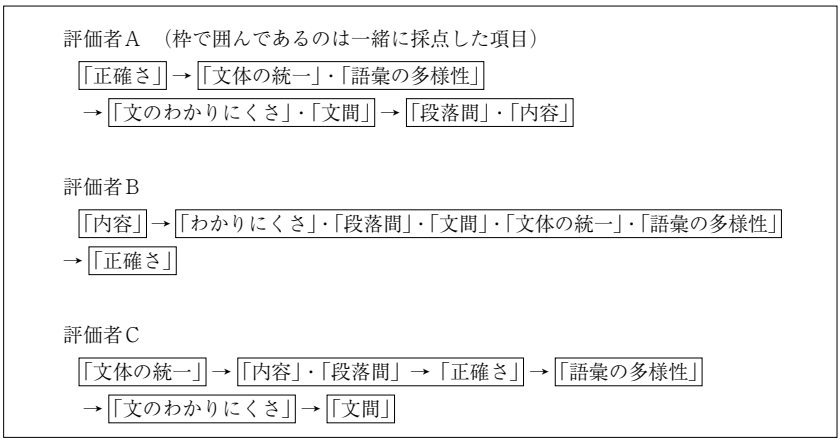

図1　個人別採点方式　評価者別の採点の流れ

評価者間で共通していたのは、採点していく際にいくつかの項目を同時に見て採点していることである。また、作文を読む回数は1つの作文につき3、4回から6回程度で、項目別採点方式に比べると少なかった。

5 ―3. 考察

このように、項目別採点方式、個人別採点方式いずれの場合にも評価者間に共通点と相違点があることが確認された。評価者間で顕著な共通の特徴としては、意味に関わる項目、あるいは正確さに関わる項目を複数まとめて採点する傾向が見られた。個人別採点の場合、何を一緒にするかには若干違いがあったほか、採点する項目の順番にかなり違いがあった。相違点が多く見られたのが個人別採点方式の方であったことは、個人別採点方式の方が項目別採点方式に比べると評価者の自由度が大きいことを考えれば、当然の結果ともいえる。いずれにせよ、この内省からはanalytic scoringの場合に採点手順の種類を設定しても、採点手順によっては評価者の内的ストラテジーにかなり違いがあることがわかった。

調査１で明らかになった評価者間の共通の特徴や採点時間を考慮して、調査２では項目別採点方式と個人別採点方式を合わせた、言うならば折衷採点方式を新たに加え、合計３つの採点方式について今度は各採点方式と評価結果の関係に焦点を当て、調査を行った。

6. 調査２の結果と分析

6 ―1. 採点手順別　評価の一致度、採点時間

表1. 採点手順別　各項目[7]評価の一致度[8]

各項目：１）正確さ　２）文体の統一　３）語彙の多様性
４）文のわかりにくさ　５）文間接続　６）段落間接続　７）内容

	正確さ	文体	語彙	わかり	文間	段落間	内容	合計
項目別	0.86	0.96	0.76	0.77	0.55	0.48	0.81	0.88
個人別	0.87	0.99	0.71	0.69	0.48	0.66	0.68	0.93
折衷	0.78	0.96	0.64	0.70	0.68	0.69	0.61	0.89

表１に見られるように、信頼係数は項目別方式が0.88、個人別方式が0.93、折衷方式が0.89となり、採点方式によって評価の一致度に大きな違いは見られなかった。また、項目別の評価の一致度を見た場合、どの採点手順においても形に関わる項目である「正確さ」と「文体」は一致度が高く、意味に関わる項目である「文のわかりにくさ」、「文間」、「段落間」、「内容」はそれよりは低くなるという共通した特徴が見られた。このことから、評価者間の考え方、判断の

仕方の違いは特に意味に関わる項目に大きく現れることがわかる。

表2．採点手順別　採点時間（分）
（作文１つあたりの平均と作文22人分の合計）

採点手順	採点時間	評価者A	評価者B	評価者C
項目別	平均	17.7	8.2	13.4
	合計	371.0	180.0	295.0
個人別	平均	10.8	4.9	10.4
	合計	227.0	107.0	230.0
折衷	平均	9.5	7.3	10.7
	合計	209.0	160.0	235.0

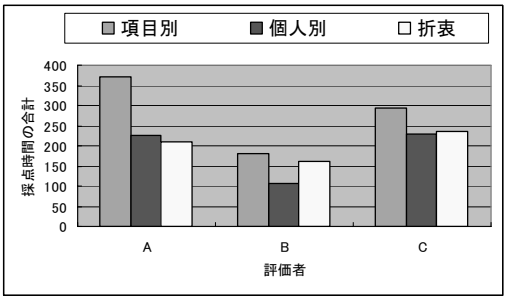

図２　評価者内手順別採点時間の比較

採点時間を測定した結果、３つの採点方式のうち最も採点に時間がかかったのは項目別方式であった。個人別方式と折衷方式を比べると、評価者Ｂは前者の方が後者よりも短時間で終わったが、評価者ＡとＣの場合は両者の間にあまり差がなかった。これは、折衷方式は個別方式よりも採点に時間がかかるだろうという調査前の予測に反する結果であった。この原因は、評価者がこの調査で折衷方式によって採点を行うときには、すでにこのテーマの作文の評価にかなり慣れていたために採点が早くなかったからではないかと考えられる。

６－２．採点手順別の項目間相関

ここでは、採点手順の違いと評価項目の独立性の関係を分析するために項目間の相関を報告する。

表3．①項目別採点方式　項目間相関

項目名	正確さ	文体	語彙	わかり	文間	段落間	内容
正確さ	1.00						
文体	0.11	1.00					
語彙	0.49	−0.03	1.00				
わかり	0.16	0.18	0.18	1.00			
文間	−0.05	−0.14	0.11	0.26	1.00		
段落間	0.21	−0.14	0.25	0.37	0.37	1.00	
内容	0.39	−0.26	0.35	0.31	0.37	0.31	1.00

項目別採点方式では、「正確さ」と「多様性」間にやや強い相関（＝0.49）が見られるが、それ以外は0.4よりも低い数値であるため、全体としては項目間の相関は弱いと言える。これに対して個人別採点方式（表４）の場合は、「正確さ」と「わかりにくさ」間にかなり強い相関（＝0.71）が見られるほか、「正確さ」と「内容」、「文のわかりにくさ」と「内容」、「段落間」と「内容」、「多様性」と「内容」間に0.5以上の値が見られ、全体的に項目間の相関が強いと言える。

表4．②個人別採点方式　項目間相関

項目名	正確さ	文体	語彙	わかり	文間	段落間	内容
正確さ	1.00						
文体	0.31	1.00					
語彙	0.35	0.41	1.00				
わかり	0.71	0.21	0.28	1.00			
文間	0.47	0.11	0.16	0.50	1.00		
段落間	0.08	0.19	0.24	0.28	0.24	1.00	
内容	0.54	0.37	0.56	0.55	0.31	0.53	1.00

表5. ③折衷方式　項目間相関

項目名	正確さ	文体	語彙	わかり	文間	段落間	内容
正確さ	1.00						
文体	0.42	1.00					
語彙	0.00	0.24	1.00				
わかり	0.46	0.32	−0.08	1.00			
文間	0.42	0.45	0.11	0.29	1.00		
段落間	0.42	0.39	−0.02	0.29	0.48	1.00	
内容	0.18	0.41	0.10	0.36	0.47	0.25	1.00

折衷方式（表５）では、項目別別採点方式と同様項目間に強い相関はないが、0.4以上の値を持つものが８つ（「正確さ」と「文体」、「正確さ」と「文のわかりにくさ」、「正確さ」と「文間」、「正確さ」と「段落間」「正確さ」と「段落間」、「文体」と「内容」、「文間」と「段落間」、「文間」と「内容」）あることに着目するならば、項目別方式と比べて項目間の相関がやや強いと言える。

このように３つの採点方式を比較した結果、個人別方式は項目間相関が強く、項目別方式と折衷方式は項目間相関が弱いことがわかった。これは言い換えるならば、項目別方式と折衷方式を採用した場合には各項目の独立性が強くなり、ある項目の評価が別の項目の評価の影響を受けないということである。

6 −3. 調査２の結果に基づく研究課題Ⅱについての考察

本稿では研究課題Ⅱ（採点手順と評価の間にどのような関係があるか）の観点として次の３つを挙げた。

① 採点手順によって評価、評価の一致度、採点時間は異なるか。異なる場合、どのように違うか。
② 採点手順の違いは各評価項目の独立性に影響を与えるか。
③ 評価において評価者の個人差と条件差のどちらが大きいか。

調査２の結果、観点①と②について明らかになったことは以下の４点である。

・３つの採点方式のうち最も採点時間が短いのは個人別方式である。
・採点方式によって評価の一致度に大きな違いは見られなかった。
・項目間相関は、３つの採点方式のなかで項目別採点方式が最も各項目の独立性が高い。
・どの採点方式の場合も、形に関わる項目（「正確さ」、「文体」）の一致度は高く、それに比べると意味に関わる項目（「わかりにくさ」、「文間」、「段落

間」、「内容」）の一致度は低かった。
これらのことから、評価の妥当性を高めるためには項目別採点方式を採用するのがよいが、実際に大量の作文を採点していく場合は、項目別採点方式は個人別採点方式よりも時間がかかり効率が悪いと言える。

このように、項目別採点方式と個別採点方式にはそれぞれ長所と短所があるため、この調査では、これら２つの手順を折衷した採点方式による評価と採点時間がどうなるかも調べた。その結果、折衷した採点方式の場合は、評価の一致度は個人別採点方式の場合とあまりかわらないこと、項目別採点方式よりは短時間で採点が済むことがわかった。従って、どうしても時間的効率を優先する必要がある場合には、個人別採点方式あるいは折衷方式を採用するのがよいのではないかと考えられる。

7. 条件と個人差

指定した採点手順を条件と考えるならば、今回の調査では評価の結果に条件はあまり影響を及ぼしていないように見える。なぜなら、７つの項目評価の合計と評価の一致度は採点手順間であまり違いがないからである。しかし、項目別の評価と評価の一致度を見るとやや揺れが見られ、評価に影響を及ぼしているのは条件というよりも、むしろ評価者間の考え方や判断の仕方の個人差ではないかと考えられる。例えば、項目別方式、個人別方式いずれの場合も意味に関わる項目の「文のわかりにくさ」「文間」「段落間」「内容」では評価者間にやや揺れが見られ、一致度はそれほど高くない。

ここでは、評価に影響を及ぼす点において個人差と条件差のどちらが大きいかを、項目別採点方式と個人別採点方式のデータを対象に二元配置の分散分析の手法を使って検証した結果を報告する。

7－1. 結果

条件間（条件差）、評価者間（個人差）について次のような有意傾向と有意差が見られた。

採点手順間（条件差）

「文体の統一」	有意傾向	（F=2.8, p<0.1）
「文のわかりにくさ」	有意差あり	（F=7.3, p<0.01）
「内容」	有意差あり	（F=5.0, p<0.05）
7項目全体	有意傾向	（F=3.8, p<0.1）

評価者間（個人差）

「語彙の多様性」	有意差あり	（F=3.8, p<0.05）
「文のわかりにくさ」	有意差あり	（F=6.1, p<0.01）
「内容」	有意差あり	（F-4.3, p<0.05）

7－2. 結果の分析と考察

以上の結果を見ると、条件差と個人差両方が評価結果に有意な差を生み出している。しかし、両者を比較した場合、全体的に見ると条件差の方が個人差よりもやや強く評価に影響を及ぼしているのではないかと考えられる。従って、作文を評価する際には採点手順を慎重に選ぶ必要があると言える。

また、「文のわかりにくさ」、「内容」で評価者間の個人差が評価に有意な差異を生み出していることは、これらの項目では解釈、判断に主観が伴うために揺れが出やすく、今後さらに検討していく必要がある項目であることを示唆していると考えられる。

8. まとめと結論

8－1. まとめ

ここでは、本研究の目的、研究課題と方法をもう一度振り返り、調査1と2の結果わかったことをまとめる。

2002年度から実施されている日本留学試験には、従来の日本語能力検定試験にはなかった小論文が導入されているが、その評価方法には、評価基準の妥当性・信頼性を含め今後さらに検討していくべき問題点があると考えられる。そこで、新しい評価基準作りを試みるプロジェクトを始めたが、基準の信頼性を検討していく過程で、評価者間の採点手順（評価ストラテジー）の違いが評価の結果に影響を及ぼしているのではないかという疑問が出てきた。本研究は、

この疑問に基づき、作文の評価を行うときに評価者の採点手順の違いが評価にどのように影響を及ぼしているかをanalytic scoringの場合に限定して実証的に調べ、そこから評価の信頼性と評価者トレーニングへの示唆を探ることを目的として出発した。

具体的な研究課題としては、Ⅰ．analytic scoringの際、評価者間で評価ストラテジーにどのような共通点と相違点があるか、Ⅱ．採点手順と評価の間にどのような関係があるか、Ⅲ．評価に与える影響において評価者の個人差と条件差のどちらが大きいか、の３つを設定した。研究方法は、日本語学習者の書いた意見文を指定された採点手順に従って３人の評価者が採点するというものである。具体的な調査の流れとしては、まず、調査１でanalytic scoringで採点手順を決めてもまだ評価者間で内的ストラテジーの違いが見られるかどうかを評価者の内省を通して確認し、次に、採点手順と評価結果との関係を調べる調査２を行った。

その結果、調査１では項目別採点方式、個人別採点方式いずれの場合にも評価者間に共通点と相違点があることが確認され、特に個人別採点の方では採点する項目の順番にかなり大きな個人差があることがわかった。先に見たように、holistic scoringについての先行研究では、すでに評価者間で評価ストラテジーが異なることが指摘されているが、この調査により、評価者の自由度が制限されるanalytic scoringの場合であっても、やはり評価者間にストラテジーの違いがあることが確認されたと言える。また、これ以外には、評価者間で共通の特徴として、意味に関わる項目、あるいは正確さに関わる項目を複数まとめて採点する傾向があることや、項目別採点方式の方が個人別採点方式よりも時間がかかることも確認された。

調査２は、調査１の結果に基づき、項目別採点方式と個人別採点方式を合わせた折衷採点方式を新たに加え、合計３つの採点方式について評価結果と採点時間の関係に焦点を当てて行った調査である。項目別採点方式と個人別採点方式を比べた結果、まず、採点時間については、項目別採点方式の方が個人別採点方式よりも大幅に時間がかかるという調査１で得られた結果が再確認された。このことから、実際に大量の作文の評価を行う場合は、個人別採点方式の方が時間的効率が良く、より実際的採点方式であると言える。次に、項目間相関については、項目別採点方式の方が各項目の独立性が強く、個人別採点方式の場合は、項目間にやや強い相関が見られることがわかった。各項目の独立性が強いということは、採点の際、ある項目が他の項目の評価の影響を受けることがないということであるから、項目別採点方式を採用した場合の方が個人別採点方式よりも評価の妥当性が高いと言える。

調査２では、このような項目別方式と個人別方式それぞれの長所と短所を考慮し、２つの方式を折衷した新たな採点方式を考えた。折衷採点方式とは、具体的には、評価基準で設定した７つの項目を関連項目ごとにまとめ、そのままとまりごとに全員分の作文を採点していくというやり方である。調査者は、この方式で採点すると項目の独立性は項目別採点方式の場合よりもやや弱くなるが、採点時間は短縮されるのではないかと予測した。その結果、予測通り、項目間相関は項目別採点方式よりやや弱いが個人別採点方式よりも強く、採点時間は短くなっていることがわかった。

３つの採点方式と評価者の一致度については、３つの間にはあまり差はなく（α係数　項目別＝0.88、個人別＝0.93、折衷＝0.89）、比較的高い一致度が得られた。従って、採点手順と評価結果の間には予想していたような強い関係は無かったと言える。

８－2．結論と示唆

本研究の調査の結果から、採点手順の違いは評価の信頼性に強い影響を及ぼしていなかったが、評価の妥当性と時間的効率の観点からは、採点手順の種類を慎重に選んだほうがよいと言える。この点で、本研究で試みた折衷採点方式は実際に評価を行っていくときに有効な採点方式の一つであろう。また、analytic scoringの場合であっても、やはり評価者間にストラテジーの違いがあることが確認された。これは、analytic scoringの場合もholistic scoringと同様に評価者トレーニングが必要であることを示唆している。

９．今後の課題

今回の調査結果と同じ結果が得られるかどうかを今後さらに検証していく必要がある。また、今回、評価者の内省は評価者が覚えていることを話し合うという方法をとったが、それでは詳細が正確にわからないため、今後はthink aloud法を使った調査も行う必要がある。

注

1　名古屋大学　平成15、16年度科研費研究「日本留学試験における記述問題の実施方法と分析観点に関する実証的研究―記述問題の問題形式・量およ

び評価基準の適正さについて」

2　評価ストラテジーを評価プロセスと呼んで扱っている論文（Lumley 2002）もあるが、本稿では評価ストラテジーとして扱う。これまでの研究では、評価プロセスと評価ストラテジーを区別するのか、あるいは評価ストラテジーを評価プロセスと同義に扱うか、まだ明確に定義付けられていないようである。

3　holistic scoringとは、作文の全体的印象を基に単一の得点で作文の評価をする評価方法のことである。Educational Testing Serviceが行うTOEFLのライティング・テストやBritish Councilが行うELTSはこの方法を採用している。この評価方法に対して、analytic scoringは作文を複数の観点から採点し、最後に個々の観点の得点を合計して評価する。

4　本研究では、採点過程を評価者の評価に至るまでのその行動過程として考える。そして、採点手順は行動過程の一部として捉える。

5　DeRemerはこれらのやり方を次のように説明している。
1）general impression scoringでは、作文を再読したり、基準の説明の意味を考えながらと基準と作文を照応することなくすぐに点数をつける。 2）text-based evaluationでは、採点する前に基準の説明を吟味して読むが、採点を始めるともう基準の説明を振り返って読み返すことをしない。 3）rubric-based evaluationでは、評価者は自分の判断を基準の説明に合わせるために何度も基準を読み返す。

6　平成14年11月（第２回）に実施された記述問題。試験問題の指示は次の通りである。
「①外国語に行って、その国の人といっしょに仕事をする場合、ある人は、＜Ａ＞その国の言葉はできるだけ上手に話せたほうがいい、と言います。またある人は、＜Ｂ＞その国の言葉はそれほど上手ではなくても、仕事に必要な程度だけ話せればいい、と言います。あなたは＜Ａ＞と＜Ｂ＞どちらの意見に賛成しますか。＜Ａ＞か＜Ｂ＞かどちらかの立場にたって、その賛成理由を書いてください。(句読点を含み400字程度)」

7　各項目とは、1）正確さ　2）文体の統一　3）語彙の多様性　4）文のわかりにくさ　5）文間接続　6）段落間接続　7）内容、の７つである。これら７項目は、表1，3，4，5で共通である。表では、枠幅を揃えるために「文のわかりにくさ」は「わかり」のように項目名を簡略化している。

8　クロンバックの α 係数を用いた。

＜参考文献＞

・小室輝代他（投稿中）「日本留学試験記述問題の新基準の提案とその信頼性」『言葉と文化』５号

・村上京子他（投稿中）「日本留学試験「記述問題」における採点基準の見直し」名古屋大学留学生センター紀要

・Cumming, A.（1990）Expertise in evaluating second language compositions. *Language Testing* 7,31-51.

・DeRemer, M.（1998）Writing assessment: raters'elaboration of the fating task. *Assessing Writing* 5（1）. 7-29

・Huot, B.（1993）The influence of holistic scoring procedures on reading and rating student writing. In M. Williamson & B.Huot（Eds.）, *Validating holistic scoring for writing assessment*（pp.206-236）. Cresskill, NJ:Hampton Press Inc.

・Lumley, T.（2002）Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing* 19（3）. 246-276.

・Pula, J., & Huot, B.（1993）A model of background influences on holistic raters. In M. Williamson & B.Huot（Eds.）, *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*（pp.237-265）. Cresskill, NJ: Hampton Press Inc.

・Vaughan.（1991）Holistic assessment: what goes on in the rater's mind? In Hamp-Lyons, L..（Ed）,*Assessing second language writing in academic contexts.*（pp.111-125）. Westport, NJ: Ablex,

・Wolfe, E.W., & M.（1996）Expertise in essay scoring. In D.C. Edelson &. Domeshek（Eds.）, *Proceedings of ICLS 96*（pp.545-550）. Chalottesville, VA: Association for the Advancement of Computing in Education.

・Wolfe, E.W.（1997）The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4（1）. 83-106.