日本留学試験「記述問題」の評価基準の提案とその信頼性

小室輝代・三谷閑子・村上京子

キーワード　*日本留学試験、記述問題、作文の評価、評価基準、妥当性、信頼性、天井効果*

1　はじめに

2002年に始まった日本留学試験では、従来からある言語知識の測定を中心とした日本語能力試験と異なり、400字程度の「記述問題」が課せられるようになった。これは、大学入学後必要となるacademic writing[1]の基礎的能力を身につけているかどうか問うもので、画期的な改善と言えよう。

しかし、この「記述問題」には評価基準の点で大きな問題が存在している。現在「記述問題」は文法的能力と論理的能力の２つの評価項目があり、各項目３点ずつの合計６点満点で採点が行われている。[2]評価基準の第一の問題はこの２項目のみという評価項目の少なさである。更に基準自体にも問題がある。文法的能力では文法・表記上の適切さよりも執筆者の意図が理解可能かどうかを重視しており、一方の論理的能力では根拠さえ書いてあれば、３点か２点のいずれかで採点されてしまう。[3]

もう一つの問題は、合計でも６点満点という配点の小ささにより得点の分散がしずらいと考えられる。図１は名古屋大学を受験した学習者の「記述問題」での総合点を示したものである。最高点である６点を取ったものの頻度が一番多く、天井効果を生み出している。この得点結果から受け入れ側の教師が個々の学習者の診断的情報を得ることは難しいであろう。

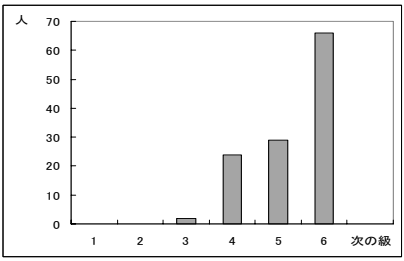

図１：「記述問題」総得点の得点分散

現在「記述問題」に限っては得点のみならず、受験者の書いた作文のコピーが各大学に送られてくる。本稿ではこの送られてきたコピーを再評価するための新しい評価基準の提案を目的とする。評価基準を立てる上では、受け入れ側の教師にとって今後の指導に役立つような情報量の多いものとなることを主眼とし、更に、上記の「記述問題」評価基準の問題点を改善するよう試みた。具体的には、①評価項目の検討と明確な基準作り、②評価基準の妥当性と信頼性、③採点結果の有効性（分散の大きさ）の確保を検討していく。

2　先行研究

大規模な記述試験としては、ＴＯＥＦＬのTest of Written English（ＴＷＥ）やＡＣＴＦＬのWritingテストがある。ＴＷＥのタスクは２者の対立する意見を提示し、どちらか一方の立場から自分の意見を述べさせる形式であり、また、評価段階も１～６点の６段階[4]というようにＴＷＥと日本留学試験の「記述問題」の形式は酷似している。このため、ＴＷＥにも上記で指摘したように凡そのレベル判定はできるが、個々の学習者の診断的情報は得にくいと考えられる。また、ＴＷＥが採用しているholistic scoring[5]は評価者の主観に左右されるという問題もある。

ＡＣＴＦＬのWritingテストの評価[6]には、初級、中級、上級、超上級の各レベルがあり、それぞれのレベルはさらに３つに下位分類され、合計12段階の評価となっている。それぞれの段階ではどのようなことができるか[7]というCan-do scaleが細かく用意されている。しかし、１回のテストでどのようなことができるのかを測るにはタクスを精査する必要があり、出題形式が規定されている「記述問題」を再評価する基準として用いるには問題がある。

日本語の作文評価については、ＴＷＥやＡＣＴＦＬのように公に確立された評価基準はないものの、森田（1981）や菊地（1987）、斉山（1994）のように各機関が独自に採用している基準は見られ、いずれもanalytic scoring[8]を取っている。analytic scoringは、holistic scoringより評価者の主観が入りにくいとされているが、菊地（1987）では100点満点中50点を主観的評価に頼る内容の配点にあてており、評価の信頼性に疑問が残る。また、森田（1981）の提示する評価項目では、内容に関する項目が「全体として言いたいことがよくわかるか」、「表現に豊かさがあるか」、「面白みが感じられるか」、「文章の運びにうまさがあり、全体としてまとまりがあるか」と４つに細分化されているが、これらの項目は評価者の主観に強く左右される恐れがある。

斉山（1994）は評価項目に「内容」と「日本語」の大きな区切りを設け、内容を「明確さ・豊かさ」、日本語を「明確さ・適切さ」に下位分類し、合計４項目で評価を行っている。また、この４分類に類似したものとして田中他（1998）がある。田中他（1998）は22の評価項目を決定し、それを因子分析した結果「正確さ」、「構成・形式」、「内容」、「豊かさ」の４因子を抽出している。田中他（1998）の４因子を構成する下位項目を見ると、それぞれの因子がどの範疇を評価しているかが明白になるが、これは作文評価の基本構造として項目を列挙したのみでレベル判定のための記述はない。

本稿ではこれら先行研究の評価項目を参考にしつつ、以下に述べるような評価項目を作成していった。

3　予備調査

3．1　対象と評価者

予備調査にあたっては、本調査で使用した作文とは別に12の作文を用意した。これらは大学進学のための予備教育を受けている学習者が日本留学試験「記述問題」の模擬試験として書いた作文である。また、評価者は日本語教育に関わる日本人母語話者３名である。

3．2　評価項目作成までの経過

まず、評価項目を立てるうえで、評価者間で検討を重ね、下記の７つの評価項目を決定した。

作文評価基準

１）正確さ：文法・語彙・表記などに間違いはないか。

2）書き言葉表現:「です･ます」文体や縮約形などの話し言葉を使っていないか。
3）語彙・表現の多様性：レポートで使われるような難しい語彙（～性、～化など）や難しい構文･文法･表現を積極的に使っているか。
4）文の首尾一貫性（1文の意味のわかりやすさ）:文としてのねじれがないか。言葉足らずなどで分かりにくい点はないか。
5）文と文の間の結びつきの適切さ:段落内での文間の接続関係は適切か。
6）段落間の意味的整合性（段落構成）:主張と根拠の間の関係や根拠と根拠の間がうまくつながっているか。
7）内容：論理性・説得力があるか。多角的な示唆や配慮がなされているか。

以上の7つの項目ごとに5段階評定を行う。
1：たいへん悪い　2：悪い　3：ふつう　4：良い　5：たいへん良い

上記の評価基準を基に、12の作文を評価したところ、評価者間一致度（Cronbackのα係数）は、内容以外のすべての項目で十分な値を示さなかった。そこで、一致度の低い項目に対し、評価者個々の基準を出し合い、評価基準の見直しを行った。

3. 3　主な見直し点

まず、不一致の大きな原因は各項目で5段階評定を行う際に、どのような質の作文にどの点数を与えるかという基準が評価者間で異なったことにある。そこで、各評定者が自身の基準を出し合い討議した上で、評価者間で合意できる得点ごとの細かな基準を設けた。また、一律5点満点を廃し、項目ごとの重み付けを行った。配点に関しては採点の容易さの他に、項目を大きく「形に関わるもの（項目1～3）」、「意味に関わるもの（項目4～7）」に分け、それらがほぼ同様の重みになるよう配慮した。

項目ごとの基準作りで問題となったのは以下の点である。

① 「正確さ」において、誤用のレベルをどのように規定するか。
② 誤用の原因は文法、語彙、文のねじれ、言葉足らずなど重複して起こることが考えられるが、複数の項目から減点するのか。
③ 同じ形式の誤用が繰り返し現れた場合、そのつど減点するのか。
④ 項目4、5、6の相違をどのように定義づけるか。

⑤ 「内容」における説得力をどのように規定するか。

①については、中上級において習得が進むにつれ減少する誤りと中上級においてもなかなか習得が進まず、残ってしまう誤りを基準に区別した。②については、評価者の判断を一致させるため、協議を重ね、その判断を評価基準に盛り込むよう努力した。また、③については、同じフレーズの誤りが繰り返された場合、誤りの回数に関わりなく１回と数えることとした。④については、基本的に項目４が一文内（「。」で区切られる範囲内）、項目５は段落内の文と文のつながり、項目６は段落と段落の続き具合を見ることとした。

⑤については、最も意見が分かれた。まず、根拠の有無から根拠が無い場合１点、有れば２点以上、そして根拠に説得力が無い場合を３点、有れば４点または５点と評価することにした。しかし、４点と５点の違いをどうすべきかが問題となった。academic writingでは多角的な視野が必要であり、反対意見に対する配慮がなければ最高点を与えられないのではとの提案もあったが、反対意見を想定することが難しいテーマ[9]や、その有無が必ずしも説得力を左右しない場合もあった。そこで、最終的には反対意見に対する配慮がなくとも説得力を基準に判断することとした。

3．4　新評価基準

評価者間で幾度か採点と基準の検討を重ねた結果、先の７つの項目に以下のような評価基準を設けた。

新基準（見直し）
項目１）正確さ
５点：ほとんど誤りがない（軽微な誤り[10]が１、２箇所程度見られる）
４点：中上級レベルの誤りが見られる（軽微な誤りが３箇所以上ある）
３点：誤りが見られるが、初級レベル、重篤な誤りが１～２箇所
２点：初級レベル、重篤な誤りが３～４箇所
１点：初級レベル、重篤な誤りが５箇所以上

項目２） 文体
３点：文体が統一されている
２点：文体が統一されていない（１、２文混じっている）
１点：文体が統一されていない（３文以上混じっている）

項目３） 表現の多様性（以下「語彙」）

５点：話し言葉[11]が全く使用されておらず、かつ、レポート・論文に使われるようなアカデミックな語彙や難しい文型が積極的に使用されている

４点：話し言葉は全く使用されていないが、レポート・論文の表現としてはやや不適切なところがある

３点：縮約形、終助詞、助詞の省略、俗語、[12]感情表現[13]の使用が１、２箇所見られる

２点：縮約形、終助詞、助詞の省略、俗語、感情表現の使用が３箇所以上見られる

１点：縮約形、終助詞、助詞の省略、俗語、感情表現の使用が頻出する

項目４） 文のわかりにくさ[14]（以下「文」）

３点：特にわかりにくい文がない

２点：わかりにくい文が１、２箇所ある

１点：わかりにくい文が頻繁にある

項目５）文と文の間の結びつき（以下「文間」）[15]

３点：接続関係[16]が適切であり、問題がない

２点：接続関係の悪い箇所が１箇所含まれる

１点：接続関係の悪い箇所が２箇所以上含まれる

項目６）文章構成、全体構成（以下「段落間」）

３点：段落構成・意味関係が適切であり、問題がない

２点：つながりの悪い箇所がある

１点：段落構成がなっていない、意味の整合性がとれていない
段落意識がない

項目７） 内容

５点：根拠に論理的整合性[17]があり、説得的である[18]

４点：根拠に説得力が見られるが、十分ではない

３点：一貫してはいるが根拠が説得的[19]ではない

２点：最初に述べたことと、結論が一致していない（一貫性がない）[20]

１点：意見の根拠が示されていない（意見ばかり述べている）

上記の基準を用い、本調査で使用する作文をテーマごとに２枚ずつ採点し、

両テーマともこの評価基準がうまく機能することを確認した。

4　本調査の対象及び手順

本調査では、名古屋大学を受験した学習者の平成14年度第２回日本留学試験「記述問題」[21]121作文を使用した。テーマによる内訳は、外国語使用に関するテーマ（以下「テーマ１」・「外国語」）が88、農薬使用に関するテーマ（以下「テーマ２」・「農薬」）が33である。テーマ１、テーマ２とも予備調査と同じ３名の評価者が上記の基準を用い、採点を行った。[22]

5　結果と考察

5. 1　評価項目の妥当性

表１は、121作文を３名の評価者で採点し、３名がつけた得点の平均点を各作文の最終結果として評価基準の項目間相関を表したものである。表１が示すとおり、「正確さ」と「文」、「正確さ」と「文間」の間で相関が見られる他は、各項目間相関は低く、それぞれの項目が独立したものを測っていると言える。よって、７つの評価項目は個別の観点から評価しているものであると考えられる。

表１：項目間相関

	正確さ	文体	語彙	文	文間	段落間	内容
正確さ	1.00						
文体	0.30	1.00					
語彙	0.29	0.28	1.00				
文	**0.46**	0.17	0.20	1.00			
文間	**0.42**	0.16	0.01	0.35	1.00		
段落間	0.24	0.18	0.16	0.30	0.29	1.00	
内容	0.39	0.22	0.33	0.29	0.33	0.24	1.00

5. 2　評価基準の信頼性とトレーニング効果

次に、テーマ毎の平均点[23]と評価者間一致度は表２のようになった。

表２：テーマごとの平均点と評価者間一致度

	受験者数	平均（ＳＤ）	評価者間一致度（α係数）
テーマ１	88名	19.82（2.63）	0.91
テーマ２	33名	19.67（2.61）	0.90
全体	121名	19.78（2.62）	0.90

表２からテーマ１、２とも平均、標準偏差（ＳＤ）両項目でほぼ同じ結果となり、今回の課題間には難易度の差がなかったものと想定できる。また、両テーマとも評価者間一致度はα＝0.9以上となり、各評価者が同じ基準で評価ができたことを示している。

表３：項目ごとの評価者間一致度

	正確さ	文体	語彙	文	文間	段落間	内容	合計
テーマ１	0.81	0.96	0.75	0.74	0.58	0.71	0.67	0.91
テーマ２	0.82	0.94	0.81	0.76	0.64	0.84	0.73	0.90
全体	0.81	0.96	0.77	0.75	0.60	0.75	0.68	0.90

次に、表３において評価項目ごとに細かく一致度を見てみると「文間」と「内容」の一致度がやや低いことがわかる（文間　α＝0.60、内容　α＝0.68）。しかし、全体の合計では、テーマ１、２ともα＝0.9となり、評価者間で高い相関を示している。したがって、本稿の評価基準の信頼性は満足できるものであると考えられる。

次に、評価者別の採点結果を示す。表４～６は評価者別に評価項目の平均点を示したものであり、それをグラフ化したものが図２～４である（Ａ、Ｂ、Ｃは評価者を示す）。

表４：テーマ１「外国語」の評価者別項目平均点

	正確さ		文体		語彙		文		文間		段落		内容		合計	
	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD
A	2.43	0.99	2.75	0.59	3.30	0.69	2.48	0.60	2.36	0.71	2.69	0.49	4.08	0.76	20.09	2.75
B	2.65	1.15	2.77	0.54	3.24	0.87	2.06	0.76	2.64	0.59	2.69	0.61	3.68	0.70	19.76	3.10
C	2.38	1.07	2.75	0.53	3.60	0.72	1.95	0.77	2.30	0.73	2.69	0.49	3.94	0.83	19.60	2.75
平均	2.48	0.91	2.76	0.53	3.38	0.62	2.16	0.58	2.43	0.50	2.69	0.42	3.90	0.59	19.82	2.63

表５：テーマ２「農薬」の評価者別項目平均点

	正確さ		文体		語彙		文		文間		段落		内容		合計	
	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD
A	2.85	0.86	2.85	0.36	2.82	0.72	2.03	0.67	2.55	0.50	2.61	0.60	3.61	0.81	19.30	2.59
B	3.00	1.26	2.82	0.46	3.12	1.07	1.85	0.70	2.64	0.48	2.76	0.49	3.64	0.73	19.82	3.26
C	2.42	1.05	2.85	0.43	3.61	0.74	1.85	0.61	2.48	0.56	2.73	0.51	3.94	0.92	19.88	2.67
平均	2.76	0.91	2.84	0.39	3.18	0.73	1.91	0.55	2.56	0.39	2.70	0.47	3.73	0.66	19.67	2.61

表６：合計の評価者別項目平均点

	正確さ		文体		語彙		文		文間		段落		内容		合計	
	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD	平均	SD
A	2.55	0.97	2.78	0.54	3.17	0.73	2.36	0.65	2.41	0.66	2.67	0.52	3.95	0.80	19.88	2.73
B	2.74	1.19	2.79	0.52	3.21	0.93	2.00	0.75	2.64	0.56	2.71	0.58	3.67	0.71	19.78	3.14
C	2.39	1.06	2.78	0.51	3.60	0.72	1.93	0.73	2.35	0.69	2.70	0.49	3.94	0.86	19.68	2.73
平均	2.56	0.92	2.78	0.50	3.33	0.66	2.09	0.58	2.47	0.48	2.69	0.43	3.85	0.62	19.78	2.62

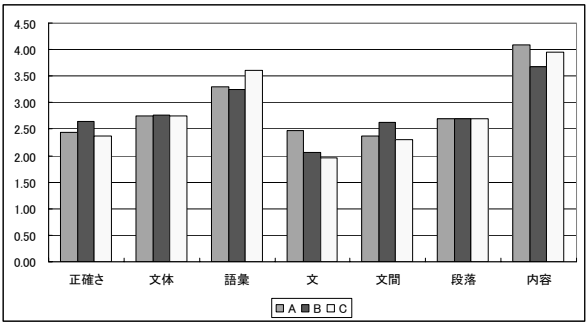

図２：テーマ１「外国語」の評価者別項目平均点

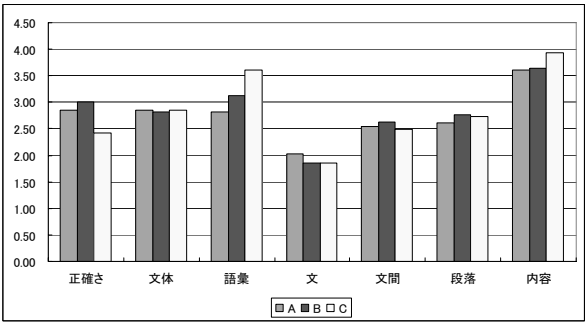

図３：テーマ２「農薬」の評価者別項目平均点

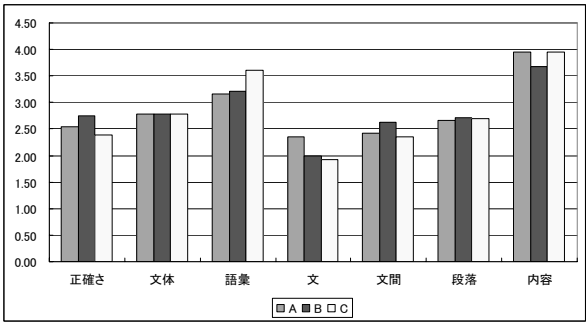

図４：合計の評価者別項目平均点

図２～４からは各評価者の評価の傾向がわかる。例えば、評価者Ｃは「語彙」、「内容」では他の評価者よりも甘い評価をしているが、「正確さ」では逆に辛い評価をしている。また、評価者Ｂは「文間」で他の評価者より甘い評価をしている一方、「内容」では、やや辛い評価となっている。このように評価者間で多少評価にばらつきが見られるが、３名の評価者は概ね同じ評価傾向を持っていると言っても良いだろう。これは、３名が評価基準の作成に当り、何度も検討を重ねた結果、それがトレーニング効果として評価基準に対する共通の認識を深めたと考えられる。

5．3　採点結果の有効性

次に、受験者の合計得点[24]の分布がどのようになっているかを、図５、６で示した。上記の表２で示したようにテーマ１、２とも課題間に難易度の差がなかったが、図５、６からも学習者の分散がテーマ１、２で同じような傾向を示していることがわかる。

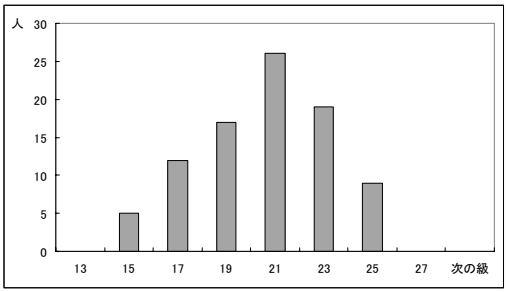

図５：テーマ１「外国語」の得点分布

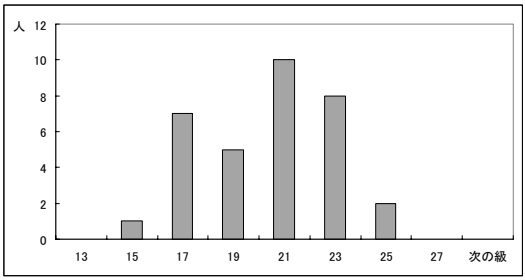

図６：テーマ２「農薬」の得点分布

日本留学試験の「記述問題」で採用されている評価基準では天井効果をもたらし、本当にできる学習者を見分けることができなかったが、本稿の評価基準を使って採点した結果では、両テーマとも最高得点と最低点を取った学習者の頻度が最も低く、学習者のレベル分けがうまくできることが証明された。

6　今後の展望

総合点における信頼性は概ね満足できる結果となったが、評価項目ごとの評価者間一致度（表３）で見たように、「文間」と「内容」の評価にはまだ信頼性の問題が残る。一致度を下げた原因としては、まず、「文間」では評価が１～３点の３段階しかないため、得点分散が少なく、評価者間の多少の得点のずれが大きく一致度を下げる結果になったと考えられる。また、「文間」が段落内の接続関係を主に見るため、文と文の間に接続詞が省略された場合、それを著しい論理の整合性の欠如と見るかどうか評価者間にずれが現れたことも原因の一つである。「内容」では根拠の説得力を判断するのに評価者の主観が大きく働いてしまったことが原因であろう。

「意味に関わるもの（項目４～７）」は「形に関わるもの（項目１～３）」より評価者の判断に揺れが生じやすいと予測できる。しかし、今回の評価においては上記の「文間」と「内容」以外は評価者間一致度が0.75以上とある程度満足のいく結果が得られた。これは、本調査を行う前に評価基準に対する協議を重ね、実際に評価を何度か行ったことがトレーニング効果となって評価者間一致度を高めたと予測できる。今後は低い一致度に留まった「文間」、「内容」の項目の一致度をより高めるための検討を進めていく必要がある。

また、本稿では日本留学試験の「記述問題」の評価を主眼に基準の開発を行っ

たため、説明文など他のジャンルの評価には、特に「内容」の項目で改良が必要である。この点も今後の課題として取り組んでいかなければならない。

注

1 本稿ではacademic writingを「大学での勉学生活で要求される書く能力」と定義する。

2 日本国際教育協会による「記述問題」の評価基準は以下に示すとおりである。

⑴文法的能力（0 ～ 3 点）

○ 個々の文についても、文章全体についても、執筆者の意図が明快に理解可能であるもの（文法・表記上の軽微な誤りや文体上やや不自然な点は許容する。）……3 点

○ 文法・表記上明らかに適切でない点を含むが、文章全体から執筆者の意図は明快に理解可能であるもの……2 点

○ 文法・表記上明らかに適切でない点がかなり目立つが、文章全体から執筆者の意図を想像することは可能であるもの……1 点

○ 意味不明の文が多く、文章全体から執筆者の意図を理解することが不可能又は極めて困難なもの……0 点

⑵論理的能力（0 ～ 3 点）

○ 主張に根拠が示されており、かつ、主張と根拠との間に十分な論理的関係があり、矛盾が認められないもの……3 点

○ 主張に根拠が示されており、概ね論理的な関係が認められるが、一部に論理的矛盾や非整合性も存在するもの……2 点

○ 主張は示されているが、その根拠が示されていない、又は、根拠が示されていても、論理性・客観性を著しく欠いているもの……1 点

○ 筆者自身の主張が示されていない、又は、何を主張したか曖昧であるもの……0 点

3 その他、この基準では語彙の適切さや文体の不統一を評価できない、初級レベルの誤りと中上級レベルの誤りに何らの区別もされていないという問題や、接続語等の誤りは論理的関係の有無で見るのか、文法の不適切さで見るのかがはっきりしないという問題がある。

4 TOEFL test of written English guide

http://www.essayline.com/upload/twetopics/tweguid.pdf

5 holistic scoringは作文を読んだときの全体的印象から評定する評価方法であり、短時間に採点できる利点がある一方、評価者の自由裁量の範囲を広げる欠点もある。

6 Preliminary ACTFL proficiency guidelines / Writing revised 2001
http://www.actfl.org/public/articles/writingguidelines.pdf

7 例えば、超上級ではformal、informal両方の場面に対応することができ、様々なトピックや抽象度の要約や大意などが書けなければならない。また、様々な文構造や文型、語彙なども駆使でき、ディスコースにより、適切なスタイルや形式が選択できなければならない。

8 analytic scoringは項目別に細かく観点を設定し、それぞれについての評点を加え合わせる評価方法である。評価者の自由裁量の余地は抑えられるが、項目ごとに何回も評定しなければならないので、評価者に大きな負担がかかる。また、項目の重み付けをどのようにするかという問題もある。

9 例えば、本調査で使用したテーマ１では外国語が「上手に話せたほうがいい」という意見の対案である「下手でもいい」という意見は想定しにくい。

10 「軽微な誤り」、「重篤な誤り」の基準は以下に示す通りである。

- ＊ 重篤な誤り：動詞の活用・格助詞・初級文型の間違い、語彙の間違いのうち意味が推測できないもの、日本で使用されない漢字(アルファベットで書かれたものは軽微な誤りとする)
- ＊ 「が」と「は」の間違い、難しい文型（使役文、受身文、授受表現等）を積極的に使用したことによる間違いは軽微な誤りとする
- ＊ 原稿用紙の使い方の間違い：句点の打ち方の誤りは重篤な誤り、マスの使い方の誤りは軽微な誤りとする
- ＊ 表記の問題（清濁/特殊音）：動詞の活用内での誤りや誤解が生じる可能性のあるものは重篤な誤り、それ以外は軽微な誤りとする
- ＊ アスペクトの間違いは軽微な誤りとする

11 「だから」、て形接続は話し言葉とは見なさない。「したがって」、連用接続を使用した場合は、アカデミックな表現とする。その他ひらがなの多用、体言止めは話し言葉的表現とし、減点対象とする。

12 「俗語」とは「やっぱ・すごい・ちょっと・やつ・おれ・あいつ・メチャクチャだ」などのような表現を指す。

13 感情表現とは「すごい！・いいなあ・がんばるぞ」などのような表現を指す。引用で感情表現を用いた場合（「すごいと思う」など）は減点対象としない。

14 前後の文脈を考えると、言葉足らず、または、接続語句に問題がある場合、句点内に接続の悪さがあれば項目4で減点する。2文以上に分かれている時は、従属節にしたほうがよい、または、接続助詞が欠落している場合でも項目5で減点する。

15 段落の最初の接続語の誤りは項目6）で減点する。

16 接続表現に限らず指示詞や前後の意味的つながりの自然さも含める。一文一文は正しくても論理的飛躍がある場合は減点対象とする。(「一文一文の正しさ」を検討する場合、前後の文脈を考えたうえで判断する。文脈を曲解すれば、文法的に正しい文と判断できるようなものは含めない。(この場合項目4「文のわかりにくさ」で減点する))

17 「論理的整合性」があるとは、結論とそれ以前の文の流れの間に論理的一貫性があることを示す。

18 論理的整合性があり、且つ説得力に富むものであっても、問題の趣旨に沿って書かれていない場合、最高を3点とする。

19 「説得的」とは、根拠が主張を支えているかどうか、根拠の説明に納得できるかどうかを問題とする。

20 説得力が見られても、最初に述べていることと、結論が一致していない場合、2点とする。

21 受験者に与えられた問題は以下のとおりである。

① 外国に行って、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〉の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)。

② 野菜や穀物などを育てる時に、害虫を殺すための農薬を使うことがあります。ある人は、〈A〉農薬は人間の体によくないから使わないほうがいい、と言います。またある人は、〈B〉農薬を使わなければ十分な収穫が得られないので、少量なら使ってもいい、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〉の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)。

22 評価手順は各評価者で、多少のずれ（どの項目から評価をするか、項目ごとに評価するのか、学習者ごとに評価するのか）があったが、評価手順の

違いによる評価者間一致度の重大な影響は見られなかった。詳しくは三谷他（投稿中）を参照。

23　3名の評価者がそれぞれに出した合計点（評価項目ごとの得点を合計したもの）の平均をその作文の最終得点とし、121作文の平均点を出した。

24　合計得点は3人の評価者の平均点である。（満点は27点）

参考文献

Jacobs, H. et（1981）　*Testing ESL composition: a practical approach* Newbury House

Coffman, W. E.（1966）　On the validity of essay tests of achievement Educational measurement Vol.3, No.2 pp.151-156

Hamp-Lyons, L.（1990）　Second language writing: assessment issues Kroll, B. Ed. *Second language writing* Cambridge pp.69-87

Jordan, R. R.（1997）　*English for academic purposes* Cambridge

菊地康人（1987）「作文の評価方法についての一私案」『日本語教育』63号 pp.87-104

斉山弥生（1994）「大学に学ぶ留学生のための作文評価試案」『産能短期大学紀要』第27号　pp.67-77

田中真理　他（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号　pp.1-12

成田高宏（2003）「第二言語としての日本語作文に対する評価の実態調査―文法形式面と内容構成面をめぐって―」『作文教育改善のためのデータベース・ツール活用』国立国語研究所　pp.79-88

三谷閑子他（投稿中）「作文の評価手順が評価に及ぼす影響　―analytic scoringの採点に関して」『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」『日本語・日本文化論集』vol.11

森田富美子（1981）「作文の評価」『日本語教育』43号　pp.17-33