コーパス分析とラッシュ・モデルを用いたライティング・テストでの困難度比較

茨城県／筑波大学大学院在籍　長橋 雅俊

概要　本研究は，作文テストで与えるトピックの違いから評価への影響を調べ，教育現場での公正な作文評価がどこまで可能か検証する。

予備調査では，極めて熟達したESL学習者によるTOEFL Test of Written English（TWE）の練習作文を，コーパス分析し，作文の長さ，語彙的特徴を測定した。この結果から対象のトピックを選び，本調査で日本人学習者のパフォーマンスを調べる。本調査１では予備調査でのコーパス分析を引き継ぎ，トピックごとに書かれた語彙的特徴を比較した。また６段階の全体的評価で採点し，得点に深刻な差がないか調べた。本調査２では，同一の学習者に２回テストを実施し，どの程度採点結果が一貫するのか調べた。

結果，異なるトピックによる作文は，高度な語彙の使用頻度に違いをもたらした。一方，全体的評価の平均点には差がなく，トピックの違いがパフォーマンスに与える影響は小さいと言える。ただし評価者の採点基準は常に一定とは限らず，厳しさの違いが確認された。この違いが得点に誤差をもたらす可能性から，現場教師のパフォーマンス評価には独断に陥らないための採点手続きが求められるだろう。

1　はじめに

これまで日本の英語教育では，総合的な技能としての指導がなされず，語彙・文法などの分割された知識の習得，文脈とは切り離された例文の反復学習が大部分を占めていた。鈴木（2002）によると，それまでの学習の基本原理には「分割可能性」（decomposability）と「非文脈化」（decontextualization）の考えが根底にあり，個々の要素を組み合わせればコミュニケーション能力は身につくと考えられてきた。しかし，文部省（1999）は新学習指導要領で「実践的コミュニケーション能力の育成」を掲げ，今日では場面に応じた会話の実践や，スピーチなどの創作活動といった指導の見直し，模索が続いている。

ライティングについても，コミュニケーションへの関心・態度を配慮した指導が見直され，和文英訳からまとまった文章を書かせる指導へ推し進めようと，繰り返し提案はされてきた（金谷, 1993）。実際，留学を志願する学生を対象としたTOEFL，IELTSなどに追随し，国内の商用言語テストでも，英検，G-TEC for STUDENTSなどが直接ライティング技能を評価するようになった。今や文法や構文理解を確認する物差しにとどまらず，技能としてのライティング活動が教室で求められているといってよいであろう。

しかし言語パフォーマンスを扱う場合，評価には時間的コストや課題の選定に労力を必要とし，現場教師が独自テストを実施し，生徒へ的確なフィードバック（例えば文法の添削，内容についてのコメントなど）を与えるまでには困難が伴う。言い換えるなら，ライティング・テストは，(a) 長い時間で１つの課題に取り組むため，数多くの言語サンプルを引き出せず，(b) 一度使用したトピックは，受験者の「慣れ」が影響するため，妥当な言語サンプルを引き出せない点が挙げられる。

2 先行研究

2.1 ライティング・テストとサンプリング

Hamp-Lyons（1991）によると，ライティング・パフォーマンスを測るテストは，以下の５つの特徴を含んでおり，日本の大学入試や高校の定期試験で課される和文英訳とは明らかに異なる。

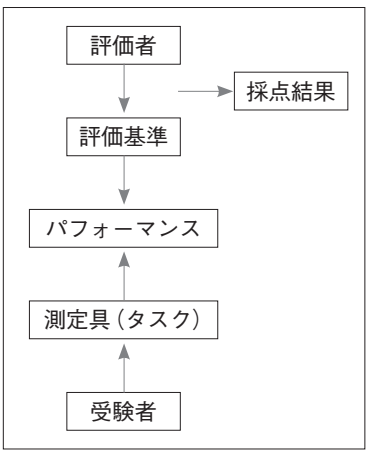
⑴ 受験者は長文（少なくとも100語）のテキストを書くことが期待される
⑵ 提示された指示の応答には十分な時間が与えられる
⑶ 書かれたテキストは１人ないし複数の評価者が読んで評価する
⑷ 評価者はサンプル・テキストや解説などで共通の基準を与えられ，評価する
⑸ 評価者の判定結果は数値で表す

Weigle（2002）はこれらの特徴について，⑴で要求されているテクストの産出量，そして⑵の「指示」が事前に受験者へ知らされないことから，自らの意見を目標言語で即興的にまとめるテストと補足している。ライティング・テストには，こうした言語能力や知識面で，高度な能力を要求するため，日本の教育現場で敬遠されがちだったと考えられる。しかし，Hughes（1989）は作文能力をテストする最善の方法は，実際に書かせることであると述べており，こうした観点の教育的評価が欠けていたのも事実である。また，先述の学習指導要領を再度取り上げるまでもなく，コミュニケーション能力の必要性が意識され始めたのは近年からであり，いずれにせよ具体的な育成方法は模索していく必要がある。

技能を直接サンプリングし，パフォーマンスを評価する難しさの原因は，日本国内の教育事情に限った問題だけではない。言語テスティング理論の立場から，McNamara（1996）は次のようなモデルを示し，テスト環境に取り巻く要因を指摘している（図１）。

この図でも明らかなように，実際の直接ライティング・テストでは文字どおり，じかに言語パフォーマンスの測定・評価を行っているわけではない。パフォーマンスとは受験者に与えられたタスクを

▶ 図１：パフォーマンス評価の特性（McNamara, 1996）



きっかけに引き出されるもので，常に受験者の典型的な能力を映し出すとは限らない。また評価者には主観が入り込むため，基準に応じた訓練が必要である（Weir, 1990）。

他にも Hamp-Lyons and Mathias（1994）は，このモデルにない要因の結び付きから評価の影響を指摘している。彼らは専門の評価者にアンケートを実施し，タスクの困難度を判定させた。そこから実際の作文採点を比べたところ，評価者に難しいと判定されたタスクの作文は平均得点が高かった。つまり，難しいと判断されるタスクは寛大に評価されがちとなり，パフォーマンス評価には，タスクの困難度と評価者判断との間にも相互作用があると考えられる。

2.2 タスクの特徴とトピック

ライティング・テストで提示される課題は，タスク（task）と呼ばれるのが一般的であり，そのさまざまな特徴を含んでいる。Weigle（2002）によると，タスクはプロンプト（prompt）よりとらえ方が広く，後者はタスクが受験者へ与える指示のみを言及している。TOEFL を主催する Educational Testing Service（ETS）は，TWE に関するリサーチ・レポートを公表しており，プロンプトの伝わりやすさも検証している。Golub-Smith, Reese and Steinhaus（1993）によると，当時よく使用していた TOEFL TWE のタスクを暗示型（Implicit）と明示型（Explicit）に分け，さらに図表タスク（Chart-graph）を加えた３つのタイプから評価得点を比べた。結果

として，明示型タスクの平均得点がわずかながら高く，以後に開発されるタスクは明示的な表現に限られるとともに，図表タスクは見られなくなった。また，Kroll and Reid（1994）の報告より，文化的に特化した言葉遣いは，多様な国籍・文化を背景としたTOEFLの受験者に誤解を招く恐れがあると指摘し，指示文の表現にも注意が払われるようになった。

指示文の他にも，受験者のパフォーマンスや評価に影響を及ぼすといわれるタスクの特徴はある。その代表的なものに談話モード（discourse mode）の違いが挙げられる。一般に，時間や空間的な順序立てで書き記す物語文（narrative）や描写タスク（descriptive task）は比較的易しいといわれ，初級の学習者によく用いられる。一方，書き手が情報を評価・吟味したり，読み手を納得させる論理的なつながりを求める論証・説得型のタスク（argumentative / persuasive task）は難しい。Crowhurst（1980）は，物語文と論証文を比較しており，後者は統語的に複雑で，作文評価にも影響すると述べている。また，Du, Brown and Rogers（1997）は，合衆国で5年生と7年生の生徒を対象に，物語文と説得型及び情報提供型（informative）の作文3タイプを書かせて評価得点を比較した。結果，物語タスクで最も平均得点が高く，個人の知識に左右される情報提供型タスクは最も低いという結果だった。

タスクの談話モードは，段落の展開や論理的構成のパターンを方向付ける。その一方，トピックは作文の内容に大きくかかわるため，特に外国語として英語を学んでいる学習者には，熟達度や教育目的に応じた慎重な選択が必要である（Hyland, 2003）。ETS（2005）によると，TOEFL TWEでは受験者に不公平が生じないよう，新たに考案したトピックは試行され，採用後も事後検証を行っている。Carlson, Bridgeman, Camp, and Waanders（1985）は，当時のTWEで同じタスク・タイプ（比較対照，及び図表タスク）に属すトピック同士であれば，平均得点がほとんど変わらないと報告している。一方，Reid（1990）はCarlsonらが研究で用いた一連の作文データを併せて，ETSコーパスの構築による追調査を行っている。それによると比べたタスク・タイプの間には，総語数と語彙的な複雑さ（単語の長さ，内容語の割合）に違いがあると報告した。

しかし，以上で紹介したETSリサーチ・レポートを見る限り，データの比較がタスク・タイプに集約され，個々のトピックの違いまでは明らかにしていない。また当時の調査には日本人受験者の報告はなく，中国語，アラビア語，スペイン語母語の受験者と英語ネイティブスピーカーの協力者による作文を基盤としている。つまり，外国語として英語を学ぶ日本の初級学習者に適用できる確証はない。限られた語彙知識を背景とする日本人学習者への利用には検証の余地があるだろう。

2.3　ライティング・パフォーマンスの測定と評価

近年では情報処理技術の進歩により，コーパス分析を用いた作文特徴の測定が容易となり，ライティングのみならず第2言語習得の研究で主流になりつつある。先に述べたReid（1990）の語彙分析では，総語数，綴りの長さといった機械的な測定方法が主で，語彙の分類方法（内容語，代名詞）も限られていたが，Laufer and Nation（1995）は，作文中で綴りの異なる語彙を整理可能とした。また，使用頻度が低く，レベルの高い語彙をリスト化することで，語彙の豊かさ（lexical variation）や洗練度（lexical sophistication）といった語彙知識の発達調査にもコーパスを応用している。

日本国内でも教育目的でコーパス利用が増えており，研究成果を上げている。例えば，大学英語教育学会（JACET）基本語改訂委員会（2003）は，日本の英語教育向けに語彙目録（通称JACET 8000）を開発し，習熟段階に配慮した語彙レベルを明らかにしている。投野（2007）は，中学・高校生の英作文およそ10,000件（総語数は約670,000 words）からJEFLLコーパスを編纂し，初級学習者の語彙や構文の発達過程や文法誤りの参考データとしてウェブ上に公開している。また，投野（2003）は教室のライティング指導にも有用な示唆を与えており，過去の生徒作文をミニ・コーパス化することで，トピックに必要なテーマ語彙の把握と補充資料の作成に役立つと提案している。

以上で述べたように，ライティング・パフォーマンスの解明にはコーパス分析による客観的な測定が増えつつあるが，Hamp-Lyons（1991）が示した評価者による作文採点が廃れたということでもない。むしろ，パフォーマンスの測定が研究向けの手法であるのに対し，評定尺度（rating scale）や「A, B, C...」といった記号採点（letter grading）は教育現

場向けである（Ellis, 2005）。

評価者を介入した作文採点の利点は，その実用性だけではない。特に，TOEFL TWE でも採用されている全体的評価は，実際に読み手の反応を見るため，妥当性が高いといわれている（Weigle, 2002）。しかし，先述のとおり評価者の主観性から批判の的にもなりやすく（2.1参照），いかに評価手続きで信頼ある採点結果を導くかが鍵となる。

2.4 項目応答理論（ラッシュ・モデル）

作文に使われる評定尺度とは，得点同士の開きが必ずしも等間隔にはならないため順序尺度でもある。このため，いわゆる古典的テスト理論からは，異なるタスクや評価者間での採点結果を対等な尺度で比較することは困難だった。一方，数学者の Rasch はロジスティック・モデルを応用することで，順序尺度を線形的な測定値と同等に扱う方法を考案し，その簡便さと実用性の高さから言語テストに関する多くの研究者に利用されてきた（大友, 1996）。

現在では，ソフトウェア FACETS（Linacre, 2006）から大規模なテストのラッシュ・モデル分析が可能であり，代表的な分析結果に，統計値 logit，fit 値が報告される。前者はテスト得点への影響力の強さ（受験者，タスク，採点者など）を統一した尺度から解釈可能にし，後者は実際の採点結果から予測モデルを構築したとき，実測と予測との間にどのくらいの逸脱があるのかを示す（本調査のデータ分析・結果にて詳述）。

Weigle（1998）はこれらの測定値を用いて，作文評価者の熟練度による採点の妥当性及び訓練の効果を実証している。また，国内でも2000年以降，スピーキングやライティングなどの採点セッションの信頼性検証に用いられ，数々の診断的情報や，教育現場での有用な示唆を提供している（秋山, 2002; 占部, 2007など）。

3 研究の目的

本研究では，ライティング・テストの実施にあたり，以下の 2 点に留意したトピックの選定を目的とする。

(1) より幅広い学習者が，ライティングを通じて自らの考え・意見を述べる機会を与えられること
(2) 学習の進度などに応じて，異なる実施時期でも公正な評価得点の比較ができること

なお，本研究での調査資料には，TOEFL TWE を用いている。大きな理由としては，(a) 185種類ものトピックや関連教材からリソースが豊富で，(b) 受験者の規模から，ある程度タスクの標準化に期待されることの 2 点が挙げられる。ただし，TOEFL TWE の受験者層は，米国やカナダへの留学準備段階にある学習者が対象となる。よって本研究では，語彙知識が発達途上の学習者や，エッセイ・ライティングの初心者にも適用可能かを検討に含む。

4 予備調査

4.1 目的

後述の本調査への準備段階として，予備調査では以下のリサーチ・クエスチョン（RQ）を掲げた。

RQ1：コーパス分析の観点から，初歩の受験者へ論証型の作文テストを実施するには，どのタスク・タイプがふさわしいか
RQ2：トピックの違いにより，書かれたエッセイのコーパス分析には，どのような変化が見られるか

4.2 方法

4.2.1 資料

市販されている TOEFL TWE のサンプル・エッセイ集（ToeflEssay.com, 2004）を使用した。ToeflEssay 社は TOEFL 受験予定者向けの学習支援サイトを提供し，独自に模擬作文テストを行っている。今回使用した書籍には，サイト利用者による Score 6 と Score 5 の練習作文が，それぞれ450篇，805篇収録されている。

4.2.2 データ入力

まず，収録されたサンプル・エッセイをコーパス・ツールで分析可能な電子テキストに変換した。そして TOEFL TWE の185に及ぶトピックをタスク・タイプへ分類するにあたり，Lougheed（2004）

を参考に整理した。それによると，TWE ではタスクの指示文やエッセイのまとめ方に特徴があり，以下 4 つのタスク・タイプに大別される。

(1) Making an argument（MA）：
仮想の状況が与えられ，何らかの意思決定とその理由を論じる
(2) Agreeing or disagreeing（AD）：
1 つの方策や命題について，賛成か反対かを理由付けて論じる
(3) Stating a preference（PR）：
2 つの意見や立場について，長所と短所を比較しながら好ましい方を論じる
(4) Giving an explanation（EX）：
ある現象や人々の行いについて，どうしてそれが起きたり，重要であるのか説明する

電子テキストを同一のタスク・タイプ及びトピック別に整理した後，コーパス分析ツールv8an（清水, 2006）を用いて，作文の語彙的特徴を数値化した。この分析ツールはウェブ上から利用可能であり，専用のホームページから英文テキストを送信することで，即座に測定結果が表示される。測定結果は JACET 8000で定められた 8 段階の語彙レベルに基づいており，最も基本語の1,000語レベルから，最も高度とされる8,000語レベルまで，総語数（tokens）及び異なり語数（types）の計測結果が得られる。

v8an は，異なり語数と総語数との比率（type token ratio）も自動的に計算して表示する。これは客観的なライティング評価でも広く用いられ，表現可能範囲（the range of expressions）の指標とされている（Read, 2000）。しかし，長文には基本語や特定の語彙が繰り返し用いられるため，むしろ内容や展開の豊かな作文で低い値を示す。これに対し，Daller, van Hout, and Treffers-Daller（2003）が紹介した Guiraud Index（type / $\sqrt{}$token）は，総語数の自乗根で比率を補正的に計算し，長さの異なるテキスト同士で比較する問題に対応している。したがって，本研究では Guiraud Index を分析の 1 つに用いる。

また，タスク・タイプやトピックに応じて要求される語彙の複雑さを考慮するため，一定のレベルを超える使用語彙の割合を明らかにする必要がある。Laufer（1994）は，最も頻繁に使われる語彙リスト2,000語に含まれない単語を洗練語とした。また，Willis（1990）によれば，最頻で使われる基本語2,500語から Collins COUBUILD English Course の全テキスト80％が表現可能であるという。そこで本研究では，JACET 8000で2,000語レベル及び3,000語レベルを超える語彙の比率を総語数と異なり語数から計算し，これら複数の語彙洗練度の値を併用して分析する。

4.2.3　データ分析

まず，タスク・タイプ別に，総語数，異なり語数，Guiraud Index，そして語彙洗練度を従属変数とし，一元配置の多変量分散分析を行った。

一方，トピックの中には収録されているエッセイの数が少ないものも含まれていた。したがって，すべての比較に等分散性は期待できないと判断し，ノンパラメトリック検定を用いた。

4.3　結果と考察

4.3.1　タスク・タイプ間の比較

コーパス分析によるサンプル・エッセイの測定値は, 表 1 に示すとおり Lougheed（2004）のタスク・タイプ 4 種類に分けてまとめた。書かれたエッセイの長さにはばらつきこそあるが，総語数の全体平均と標準偏差からおよそ70％の書き手が260〜420 words で作文をまとめていることが推計される（平均：338.63，標準偏差：75.95）。また，論証型のタスクで求められる語彙の豊かさという点では，いずれのトピックにしても異なり語数にして150 words 前後で十分に議論が展開できると考えられる。

多変量分散分析により，これらの測定値でタスク・タイプ間の違いを調べた結果，総語数に有意な差はなく［Tokens: $F(3) = .42$, $p = .742$］，異なり語数に関しても，有意傾向だが明らかな違いとはいえなかった［Types: $F(3) = 2.20$, $p = .086$］。一方，語彙の表現可能範囲を Guiraud Index から比べた場合，違いが見られ［$F(3) = 5.92$, $p = .001$］，洗練語彙率と種類にも，いずれかのタスク・タイプ間で差があることが示された［Lv2 Tk/Tk: $F(3) = 9.80$, $p = .000$; Lv3 Tk/Tk: $F(3) = 5.14$, $p = .002$; Lv2 Tp/Tp: $F(3) = 5.77$, $p = .001$; Lv3 Tp/Tp: $F(3) = 2.71$, $p = .044$］。

そこで，どのタスク・タイプ間の違いなのか特定するため，Tukey の HSD 法で事後検定を行った。

■表1：タスク・タイプごとのコーパス分析結果

		総語数（Tokens）		異なり語数（Types）		Guiraud Index	
	n	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
MA	410	338.70	72.95	155.96	29.07	8.48	.92
AD	367	335.32	72.15	154.44	29.81	8.44	1.03
PR	285	340.51	80.43	151.78	30.41	8.23	1.02
EX	193	341.98	82.50	158.50	29.62	8.59	.94

		語彙洗練度（%）							
		Lv2 Tk / Tk		Lv3 Tk / Tk		Lv2 Tp / Tp		Lv3 Tp / Tp	
	n	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
MA	410	9.13	3.41	6.08	2.88	15.07	4.88	10.25	4.13
AD	367	7.98	2.89	5.37	2.28	14.19	4.57	9.59	3.80
PR	285	8.04	3.49	5.58	2.71	13.68	4.78	9.51	3.84
EX	193	8.47	3.45	5.55	2.64	14.96	5.08	9.99	4.05

（注）Mean = 平均値；SD = 標準偏差；n = 個数；
Lv2 = JACET 8000 で2,000語レベル超の語彙数；Lv3 = JACET 8000 で3,000語レベル超の語彙数；
Tk / Tk = 総語数に対する洗練語の使用頻度の比率；Tp / Tp = 異なり語数に対する洗練語の種類の比率。

まず Guiraud Index に関しては，PR タイプが他のどのタイプよりも低いことがわかった［MA > PR: p = .006, AD > PR: p = .039, PR < EX: p = .001］。語彙洗練度については，2,000語と3,000語レベルを境目とした比較で，互いに類似した組み合わせに差が見られたが，2,000語レベルからの比較で傾向をより強く示していた。総語数の比率から2,000語レベルを超える洗練語の使用頻度を比べた場合，MA タイプが AD・PR タイプを上回っており［(Lv2 Tk/Tk) MA > AD: p = .000, MA > PR: p = .000］，3,000語レベルから比べても MA・AD タイプの間に差が確認された［(Lv3 Tk/Tk) MA > AD: p = .001］。また，異なり語数の比率から洗練語種類の豊富さを調べた場合，2,000語レベルからの基準で PR タイプは MA・EX タイプを下回っていたが［(Lv2 Tp/Tp) MA > PR: p = .001, PR < EX: p = .023］，3,000語レベルでは有意傾向を検出するのみだった［(Lv3 Tp/Tp) MA ≧ AD: p = .094, MA ≧ PR: p = .071］。

以上の結果から RQ1 に関して考察すると，(a) タスク・タイプによる作文の長さにほとんど違いはなく，(b) PR タイプは語彙知識に求められる表現の範囲が比較的小さく，また (c) AD・PR タイプは MA タイプほどに高度な語彙使用を要求しないと考えられる。特に，二者択一型のタスクは，受験者が議論の方向付けを選べて書き出しやすいためか，TOEFL 以外の言語テストや専門の研究者でもよく用いられている。また，タスク・タイプによって引き出すライティング・パフォーマンスが変化することが既に明らかにされているため（Reid, 1990），本研究では以降の調査対象を PR タイプに絞って進めることにする。

4.3.2　トピック間の比較

前節の結果を踏まえ，PR タイプのトピックを調べたところ，185のうち39がこのタイプに属している。ただし構築したコーパスから，サンプル・エッセイの数が 5 つに満たないトピックは分析対象から除いた。表 2 は残りのトピック29種類で書かれたエッセイ248篇の要約である。

クラスカル・ウォリスの順序和検定で，全体からの比較を行ったところ，おおむねタスク・タイプと同じ指標に違いが見られた。つまり，総語数と異なり語数には，有意差はなく［(tokens) X^2 = 26.91, p = .523; (types) X^2 = 19.00, p = .898］，語彙洗練度には，すべての指標で有意な差が見られた［(Lv2 Tk/Tk) X^2 = 93.03, p = .000; (Lv3 Tk/Tk) X^2 = 94.66, p = .000; (Lv2 Tp/Tp) X^2 = 57.36, p = .001; (Lv3 Tp/Tp) X^2 = 63.36, p = .000］。ただし，Guiraud Index にはトピック間での統計的な差は見られなかった［(Guiraud Index) X^2 = 27.65, p = .483］。つまり，PR タイプ内を調べた限り，トピックで要求される語彙的な表現の範囲に違いがないことが予想される。た

■ 表 2：PR タイプ作文のコーパス分析結果の記述統計

	総語数（Tokens）		異なり語数（Types）		Guiraud Index	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Max	396.08	138.09	174.77	45.48	9.15	.74
Min	292.00	50.99	133.14	25.16	7.43	1.20
Total	344.39	82.48	152.63	31.03	8.23	1.04

	語彙洗練度（%）							
	Lv2 Tk / Tk		Lv3 Tk / Tk		Lv2 Tp / Tp		Lv3 Tp / Tp	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Max	12.85	4.48	8.74	3.23	17.87	2.55	13.06	2.93
Min	4.42	2.19	2.74	1.13	9.09	3.28	5.91	2.05
Total	8.04	3.48	5.60	2.72	13.81	4.82	9.60	3.91

（注）*N*（全個数）= 248;
Max = 数値が最も高かったトピックの平均; Min = 数値が最も低かったトピックの平均;
その他の略号は表 1 を参照。

だし，エッセイを書く難しさへの要因には，トピックで頻繁に使いがちな語彙のレベルが考えられる。そこでさらなる検証として，個々のトピックに引き出された語彙レベルの比較を試みた。

表 3 は PR タイプから抽出されたトピックと，それらの作文からコーパス分析した測定値の要約である。トピックはサンプル・エッセイの数が十分で，なおかつ測定値の平均順位が上位から下位まで幅広く観測できるよう，総合的に判断して選んだ。

これら 6 種類のトピックを，作文の語彙洗練度からマン・ホイットニーの U 検定で総当たりの比較を行った。その結果，トピック D で書かれたエッセイは3,000語レベルを超える語彙の使用頻度（Lv3 Tk/Tk）がトピック B（$U = 17.00, p = .002$），C（$U = 15.00, p = .014$），F（$U = 14.00, p = .040$）の作文より高く，語彙の種類（Lv3 Tp/Tp）から比べても，B（$U = 30.00, p = .019$），F（$U = 14.00, p = .040$）より豊富な洗練語の組み合わせで書かれていたことが推定された。一方，トピック B によるエッセイは，2,000語レベルを超えた語彙の出現頻度がトピック A のエッセイより低かった（$U = 49.00, p = .042$）。

以上の結果から RQ2 について考察すると，TOEFL TWE で用いられているトピックは，内容面で多岐にわたっていながら，コーパス分析による測定結果で均整が保たれている語彙的特徴と，そうでないものがあることがわかった。まず，各トピックで引き出された作文全体の総語数や異なり語数には，統計的な違いはなかった点が挙げられる。つまり，タスク・タイプやトピックを変更しても文章の長さや論を展開する語彙の種類は，時間制限などの条件が統制されている限り，一定の量に収束されるようである。

一方，作文に用いていた語彙は，トピックから引き出される意見や具体的な事例によって多方面へ特徴付けられる。これらテーマ語彙は，今回の JACET 8000を基準とした学習者の習得時期から見てもさまざまであり，語彙レベルが一様とは限らないことがわかった。

しかし，予備調査の資料は作文の評価得点が極めて高く，熟達した学習者のコーパスを基盤としており，語彙知識が発達途上の学習者のパフォーマンスを予見するには未調査である。以降の本調査では，日本人学習者が，トピックによって作文結果にどの程度影響を受けるのか検証を進めていく。

5 本調査 1

5.1 目的

予備調査の結果を踏まえ，ここでは絞り込まれた 6 つのトピックからライティング・テストを実施し，独自の調査を行う。前節では，異なるトピックから引き出されたエッセイをコーパスによって分析・測定したが，本調査でも RQ2 に基づいた同様の比較を試みる（4.1参照）。

また，トピックの違いが読み手の反応に影響を与

■ 表 3：対象トピックにおける語彙的特徴比較

		総語数（Tokens）			異なり語数（Types）			Guiraud Index		
トピック	*n*	*Mean*	*SD*	*Rank*	*Mean*	*SD*	*Rank*	*Mean*	*SD*	*Rank*
A	13	396.08	138.09	1	174.77	45.48	1	8.78	1.23	2
B	14	376.50	110.60	6	157.93	36.01	6	8.14	.98	17
C	9	334.78	80.65	16	156.67	35.98	9	8.53	.95	5
D	10	342.40	88.27	14	168.80	30.43	2	9.15	.74	1
E	8	341.13	89.15	15	157.50	44.23	8	8.47	1.26	7
F	7	312.00	57.58	27	134.14	24.76	28	7.59	1.03	28

		語彙洗練度（%）					
		Lv2 Tk / Tk			Lv3 Tk / Tk		
トピック	*n*	*Mean*	*SD*	*Rank*	*Mean*	*SD*	*Rank*
A	13	9.33	3.97	8	6.44	3.61	8
B	14	6.71	1.89	22	4.39	1.72	22
C	9	7.21	1.87	19	4.88	1.74	19
D	10	10.18	2.72	5	7.40	2.13	5
E	8	7.69	2.09	15	5.70	1.97	14
F	7	7.38	3.47	18	4.74	1.97	20

		語彙洗練度（%）					
		Lv2 Tp / Tp			Lv3 Tp / Tp		
トピック	*n*	*Mean*	*SD*	*Rank*	*Mean*	*SD*	*Rank*
A	13	16.61	6.65	4	11.13	5.77	8
B	14	13.56	2.63	16	8.92	2.32	18
C	9	13.15	3.86	17	8.89	3.66	19
D	10	17.39	4.98	2	12.48	3.51	2
E	8	14.32	3.96	12	11.17	3.73	7
F	7	12.92	4.33	18	8.64	2.69	20

（注）Topic A～F の指示文については，資料 1 を参照；
Rank = 全 29 トピック中で数値の高い順に表示；
その他の略号は表 1 を参照。

えないか，全体的評価によって調べた。関連するリサーチ・クエスチョンを以下のとおり追加する。

RQ3：トピックの違いにより，書かれたエッセイの評価得点には，どのくらい変化が見られるか

5.2 方法

5.2.1 参加者

受験者には141名の大学 1 年生が参加した（内訳は文系35名，理系65名，医療系41名）。

また評価者には，調査実施者を含む大学院生 8 名（英語教育または英語学を専攻）が参加し，うち 5 名が大学，高校，または英会話学校での教員経験を有していた（平均：3.50年，標準偏差：4.33）。

5.2.2 資料及び手順

まず，事前に熟達度テストを用意し，語彙・文法に基づく単文完成20問（英検準 2 級～準 1 級），並べ替え15問（英検準 2 級と 2 級），文法誤り15問（TOEFL 練習問題）の計50問を40分で課した（テストの信頼性については，資料 2 を参照）。

ライティング・テストは，熟達度テストの採点結果に基づき等質の 6 グループに分け，それぞれに予備調査で選定したトピックを与えた。TOEFL TWE の規程に合わせるため，トピックを印刷した課題用紙と清書用紙，下書き用紙を配布して30分間で実施した。

収集したエッセイは Microsoft Word で転記し，電子テキストとしてデータ保存した。転記の際は，学生の書いた文法や綴りの誤りもそのまま写した。これら電子テキストを予備調査と同様，v8an を用いてコーパス分析し，総語数，異なり語数，洗練語彙の頻度と率を測定した。

また書き手個々の筆致が読み手の印象に影響しないように，電子テキストを印刷した冊子を全体的評価に利用した。評価基準には Criterion Scoring Guide（ETS, n.d.）を用い，採点トレーニングの後 6 段階で採点した（資料3）。採点セッションには，図 2 に示すような部分交差モデルで担当を割り振り，1 篇のエッセイは必ず 5 名の評価者が読むことにした。

5.2.3　データ分析

RQ2 を明らかにする上で，作文中の測定値（総語数，異なり語数，Guiraud Index，語彙洗練度）をグループ間で比較した。予備調査と同様，統計処理には一元配置の多変量分散分析を用いた。

RQ3 に関しては，全体的評価の採点データを FACETS（Linacre, 2006）を用いてラッシュ・モデル分析した。今回の分析では 3 相の構成要素（受験者，トピック，評価者）それぞれに統計値 logit が報告され，受験者であれば能力の高さを，トピックと評価者であれば，難しさと厳しさを測定する。従来, 相関係数（ピアソンの積率相関係数など）では，例えば一方の評価者が別の評価者の倍近い得点を与えるような状況でも，相関図のパターンが線状を描けば高い相関係数を示していた。これでは実態を欠いた採点を見過ごす上，どちらの配点が妥当かさえ明らかにできない。一方，ラッシュ・モデルでは logit を調べることで不均衡な配点を確認し，併せて fit 値を調べることで評価者の一貫性を診断できるところに特徴がある。

5.3　結果

5.3.1　コーパス分析による比較

表 4 はトピック別に 6 グループの受験者が書いたエッセイを，コーパス分析による測定結果でまとめたものである。予備調査の極めて熟達した ESL 学習者コーパスと比べた場合，本調査の学習者の書いた作文は総語数ー異なり語数の平均で，およそ40%〜45%に縮小していた。

多変量分散分析からは，グループ間で総語数と異なり語数のいずれにも統計的に有意な差はなかった［Tokens: $F(5) = .11$, $p = .989$; Types: $F(5) = .56$, $p = .731$］。このことから，テストに用いるトピックを変更しても，他のテスト条件（例：時間制限など）が同じである限り，作文の長さにほとんど影響のないことが改めて確認された。

ただし，語彙の洗練度に関しては，頻度と種類の多さとで結果に不一致が見られた。高度な語彙を使った回数を反映した場合，トピック間に統計的に有意な違いが2,000語・3,000語のいずれの基準からも見られた［Lv2 Tk/Tk: $F(5) = 2.61$, $p = .028$; Lv3 Tk/Tk: $F(5) = 2.73$, $p = .022$］。一方で，高度な語彙を異なり語数からの比率で求めた場合，有意差は認められなかった［Lv2 Tp/Tp: $F(5) = 1.82$, $p = .113$; Lv3 Tp/Tp: $F(5) = 2.02$, $p = .080$］。

ここまでの結果から，予備調査とは部分的に重なっており，トピックが作文の使用語彙を左右する

▶ 図 2：採点セッション・モデル 1

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Rater 1	x	x	x	x	x	x	x					
Rater 2			x	x	x	x	x	x	x			
Rater 3					x	x	x	x	x	x	x	
Rater 4	x						x	x	x	x	x	x
Rater 5	x	x	x						x	x	x	x
Rater 6	x	x	x	x	x						x	x
Adjuster-Rater		x		x		x		x		x		x
Master-Rater	x	x	x	x	x	x	x	x	x	x	x	x

（注）Rater = 評価者；P = 参加者の書いた作文；評価者 1 〜 6 は X 印の付いた作文採点を担当した。
Adjuster-Rater は評価者人数の調整分を担当し，Master-Rater（筆者）はすべての作文を担当した。

■ 表 4：日本人学習者によるトピックごとの語彙的特徴比較

		総語数（Tokens）		異なり語数（Types）		Guiraud Index	
トピック	*n*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
A	24	136.88	53.75	66.00	20.95	5.65	.79
B	24	130.58	44.54	65.75	16.58	5.77	.71
C	24	134.79	54.92	66.13	16.70	5.78	.72
D	23	134.26	53.32	67.70	22.07	5.84	.96
E	23	141.09	43.83	72.74	17.65	6.15	.73
F	23	135.30	44.54	64.91	15.32	5.62	.72

		語彙洗練度（%）							
		Lv2 Tk / Tk		Lv3 Tk / Tk		Lv2 Tp / Tp		Lv3 Tp / Tp	
トピック	*n*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
A	24	4.25	2.41	3.17	2.21	8.43	4.77	6.29	4.28
B	24	3.97	2.26	2.75	2.04	7.87	4.85	5.50	4.49
C	24	4.53	3.78	3.50	3.31	9.17	8.61	7.28	7.95
D	23	6.23	3.09	5.16	2.83	12.04	5.95	10.09	5.90
E	23	6.15	2.99	4.53	2.78	11.54	5.40	8.54	5.12
F	23	4.80	2.67	3.69	2.31	9.85	5.60	7.59	4.65

（注）Topic A～F の指示文については，資料 1 を参照；
略号については表 1 を参照。

ことは断定できる。ただし，洗練語の種類の豊富さで差異が見られなかったのは，トピックによって高度な語彙を引き出すかどうかが，書き手の使用可能な語彙知識，つまり発表語彙のサイズにもかかわっているためと考えられる。

次に，どのトピック組み合わせで洗練語彙の使用頻度に違いがあるのか，Tukey HSD 法で事後検定を行った（表 5 参照）。それによると，トピック D はトピック A，B，C に比べ，2,000語と3,000語どちらの基準で比較した場合でも，作文中でより多くの高度な語彙を引き出していた。同様にトピック E はトピック A，B に対し，2,000語レベルを超える洗練語彙の使用頻度が多く，トピック B に至っては3,000語レベルから比較しても有意な違いであった。したがって，トピック D，E を与えられた受験者は比較的高度な語彙を使いながら書いていた一方で，トピック B を与えられた受験者は平易な語彙で内容を構成していたと考えられる。

■ 表 5：事後検定（Tukey HSD）結果

比較トピック	測定法	p
A < D	Lv2 Tk/Tk	.021
	Lv3 Tk/Tk	.010
B < D	Lv2 Tk/Tk	.009
	Lv3 Tk/Tk	.002
C < D	Lv2 Tk/Tk	.048
	Lv3 Tk/Tk	.031
A < E	Lv2 Tk/Tk	.027
B < E	Lv2 Tk/Tk	.011
	Lv3 Tk/Tk	.021

5.3.2 全体的評価得点の比較（その 1）

では，高度な語彙が特定のトピックに偏って引き出されていると仮定したとき，全体的評価の結果には影響するのだろうか。以下の表 6，7 はトピックごとの難しさ，評価者の厳しさ，そして各得点結果を，FACETS で分析した結果である。

先述のとおり logit とは，評価得点を変動させる強さを表し，例えばトピックなら難しさ，評価者なら厳しさを測定したものである。また，FACETS はテストにかかわる要因（例：受験者，トピック，評価者）の測定結果を別個の表で提示し，実測の平均得点（observed score）と公正化された平均得点（fair average score）を報告する。特に後者の公正化された値は，他の要因の影響を同等と見なして測定するため，対等な比較解釈が可能である。

まず，表6からトピックの難しさについて調べると，logitは平均が0を示している。これは正の値をとると平均より難しく，負の値をとれば易しいことを表す。これに従えば，トピックEとCが両極端に位置し，難しさの差はlogit = .57であった［C - E = .29 - (-.28)］。一方，実際の得点で見ると，6段階評価の平均が3.39点，公正化した値なら3.27点であった。トピックごとの実測得点と公正化された得点も，これら平均値から大きく離れることはなく，仮にトピックEとCを評価得点で比べた場合，0.17点の差が生じるのみであった。したがって，ライティング・テストの全体的評価における，トピックからの得点への影響は極めて小さいとみられる。

一方，今回の評価者による採点セッションが妥当か診断するには，表7を参照する。ラッシュ・モデル分析で評価の信頼性を確かめる場合，fit値とlogitが有効であることは先に述べたとおりである。本研究ではinfit mean square（Infit MS）から評価者の判断の一貫性を調べ，logitと公正化された評価得点から，厳しさに個人差がなかったか検討する。

まず，評価者の一貫性を測るInfit MSは1.00が最も理想的とされ，この値を上回るほど評価得点の分布に歪みがあるとされる。McNamara（1996）は許容できる上限の値を1.30とし，それ以上の値を示す評価者は判断に揺らぎがあると説明している。この基準によれば，今回の評価者8名はすべて許容範囲内であり，おおむね採点セッションはうまく実施できたとみられる。

しかしInfit MSによる診断結果は，作文の優劣の区別が似通っていることを示すのみで，段階得点ごとの配分が評価者間で同じとは限らない。例えば，評価者個々のlogitを参照したとき，評価者4は最も厳しく，評価者2は最も寛大な配点をしていた。これらを公正化した得点に換算すると，平均で0.5点近い差が確認された（3.47-3.01 = 0.46）。このことから単純に推計しても，例えば評価者2が4点を

■ 表6：トピックのラッシュ・モデル分析結果（その1）

トピック	実測			公正化された平均得点	Logit	Infit MS
	得点	個数	平均			
A	392	115	3.41	3.26	.04	.69
B	421	120	3.51	3.34	-.24	.75
C	387	120	3.23	3.19	.29	1.03
D	387	115	3.37	3.26	.05	1.24
E	389	110	3.54	3.36	-.28	.94
F	379	115	3.30	3.23	.14	1.05
Mean	392.5	115.8	3.39	3.27	.00	.95
SD	13.3		.10	.06	.20	.19

■ 表7：評価者のラッシュ・モデル分析結果（その1）

評価者	実測		公正化された平均得点	Logit	Infit MS
	個数	平均			
1	79	3.20	3.15	.43	.74
2	83	3.55	3.47	-.64	1.09
3	79	3.28	3.19	.28	.84
4	80	3.08	3.01	1.00	.84
5	84	3.40	3.30	-.11	1.25
6	84	3.46	3.33	-.19	1.27
Adjuster	67	3.49	3.35	-.27	.74
Master	139	3.53	3.42	-.50	.83
Mean	86.9	3.39	3.28	.00	.95
SD		.20	.14	.50	.20

与えた作文の半数近くを，評価者 4 が 3 点と判断する状況は起こり得ていたことがうかがえる。

5.4 本調査 1 の考察

以上の結果から，コーパス分析と全体的評価による双方の結果を総合して考察すると，ライティング課題で提示される特定のトピックの中には，高度なテーマ語彙を多く引き出すものが存在するが，全体的評価の得点平均に大きな差はないことがわかった。また，語彙洗練度の平均値が高く測定されたトピック D，E の評価得点がともに低いということはなく，トピック E は 6 つの中でも中程度であった。むしろ比較的平易な語彙を多く使っていたとされるトピック B の受験者グループの方が，全体的評価では 2 番目に平均得点が低く，作文中の語彙洗練度と評価者の意思決定には，明らかな関係性は見られなかった。考えられる理由として，2 つのことが推測される。その 1 つは語彙洗練度の測定結果から，そしてもう 1 つは判断の厳しさをトピックごとに調整した可能性が挙げられる。

第 1 の可能性を振り返ると，本研究では JACET 8000から一定レベル以上の語彙の割合を算定し，語彙洗練度の評価に充てていた。そして，トピック間に違いを示していたのは，洗練語彙の頻度（Lv2, Lv3 Tk/Tk）のみで，異なり語数の比率（Lv2, Lv3 Tp/Tp）には違いがなかった。つまり，特定の難しい語彙こそ繰り返し用いてはいたが，種類の豊富さという点では，評価者にとって表現が高尚と認めるほどではなかったことが考えられる。

第 2 の可能性としては，Hamp-Lyons and Mathias（1994）の指摘のとおり，評価者のタスクへの判断が採点に影響した場合が考えられる。本調査では評価者へ割り当てたエッセイのトピックについても，ほぼ均等に分配していた。したがって，同一のトピックで書かれた作文同士を突き合わせ，得点のバランスを調整することも可能であった。とはいえ，本調査ではこれら心的行為の証拠を得ることは目的としておらず，評価者の誰もがこうした方策を採ったか明らかではない。この点に関しては，今後採点プロセスに焦点を向けた別の研究報告が待たれるであろう。

また，もう 1 つ未検証の課題として，同一の学習者が複数回のテストを受験した場合の結果を確かめる必要がある。学習進度に応じてライティング技能を評価する場合，同じ課題の繰り返し提示はトピックの練習効果が混在し，純粋なパフォーマンス評価には適さない。そこで，同等の比較が可能なライティング課題の一考察に向けて，本調査 2 へと検証を進めていく。

6 本調査2

6.1 目的

ここではまず，本調査 1 に掲げた RQ3 の追検証を目的としている（5.1参照）。また学習の進展に応じて，ライティング・テストを行うとすれば，過去の採点結果とどこまで信頼のおける比較が可能だろうか。この点に関し，以下のリサーチ・クエスチョンを追加して調査を進める。

RQ4：同一の書き手に異なるトピックを課したとき，どのくらい一貫した評価得点が得られるのか

なお，ここで「同一の書き手」とは，先と後で実施される 2 回のライティング・テストの間に，一切の学習やリサーチ機会を与えていないことを表す。つまり，英語熟達度と背景知識に変化がないと仮定される参加者から比較調査を行う。

6.2 方法

6.2.1 参加者

受験者には，調査 1 とは別クラスに属す67名の大学 1 年生が参加した（文系34名，理系33名）。

また，調査 1 で参加した評価者 8 名に加え，4 名の現職英語科教員（高校 1，専修学校 1，大学2）にも協力をいただき，最終的には12名中11名の現場経験者で評価者団を構成した（平均：6.73年，標準偏差：8.22）。

6.2.2 資料及び手順

本調査 1 と同じ熟達度テストを実施し，グループ分けを行った。ライティング・テストも，本調査 1 と同じ規程と配布物（課題用紙，清書用紙，下書き用紙）で30分間，1 回目を実施した。そして，数分間の小休止の後，異なるトピックを印刷した課題用紙を各自に配布する以外は，同じ手順で 2 回目のテ

ストを行った。

なお，1 グループ当たりの受験者の人数を考慮し，用いるトピックを 6 つから 4 つに再検討した。本調査 1 で測定された logit から，比較的易しいとされるトピック B（-.28），E（-.24），中程度のトピック A（.04），そして最も logit が高く難しいとされたトピック C（.29）に絞って調査を継続した。表 8 は各グループに 1 回目と 2 回目のテストで与えたトピックである。順序効果を差し引くため，グループ 4 ～ 6 はグループ 1 ～ 3 で提示するトピックと順序を対称に配置した。また，全作文評価を係留して分析するため，全受験者にトピック B と E の両方，またはいずれかが必ず提示されるようにした。

■ 表 8：トピックの組み合わせと提示順序

グループ	*n*	トピック	
		テスト 1	テスト 2
1	11	E	B
2	10	E	C
3	11	B	A
4	12	B	E
5	12	C	E
6	11	A	B

採点セッションに用いる評価基準は，本調査 1 と同じ Criterion Scoring Guide を用いた。また，作文も前の調査に倣い，原文から忠実に Microsoft Word で転記・印刷したものを綴じて配布した。

評価者への作文の割り当てについては，図 3 に示すとおり，前の調査から継続して協力をいただいた評価者と，新たにセッションに参加した評価者とで調整を行った。最終的に，エッセイは 1 篇あたり12名のうち 7 名に必ず読まれる。

6.2.3　データ分析

評価者12名による全体的評価の採点データを，本調査 1 と同様，FACETS を用いて分析した。これにより，トピック間で logit の深刻な差異がないか再確認する。また前の調査と異なる点として，(a) 4 相（受験者×トピック×評価者×テスト回数）のラッシュ・モデルで分析していること，(b) 受験者とトピックとの相互で誤差分析（bias analysis）を行っていることが挙げられる。前者は，受験者全体の傾向として，ライティング・テストの 1 回目と 2 回目とで難易度が変動していないか調べている。後者については，各受験者とトピックとの 2 通りの組み合わせから，測定された logit が統計的に有意な差をもたらしてないか検定結果が報告される。

6.3　結果

6.3.1　全体的評価得点の比較（その 2）

まず，本調査 1 と同じ順序でトピックの難しさ，評価者の厳しさの分布を確認する。表 9 は本調査 2 で調査の対象となったトピック 4 つの難しさ，及び評価得点の平均である。FACETS のラッシュ・モデル分析により，得点平均は実測値と公正化された予測値が併記されている。

今回の測定結果を logit に従うと，トピック A が

▶ 図 3：採点セッション・モデル 2

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Rater 1		x	x	x	x	x						
Rater 2				x	x	x	x	x				
Rater 3						x	x	x	x	x		
Rater 4								x	x	x	x	x
Rater 5	x	x								x	x	x
Rater 6	x	x	x	x								x
Rater 7	x	x	x	x	x	x	x	x	x			
Rater 8				x	x	x	x	x	x	x	x	x
Rater 9	x	x	x				x	x	x	x	x	x
Rater 10	x	x	x	x	x	x				x	x	x
Adjuster-Rater	x		x		x		x		x		x	
Master-Rater	x	x	x	x	x	x	x	x	x	x	x	x

最も易しく，トピックBが最も難しいと測定され，本調査1とは順位に多少の変動があった。しかしながら，これら2つのトピックで平均得点を比較した場合，実測は0.37点差，さらに公正化得点で採点者の厳しさを補正すると0.20点差に縮まった。つまり，本調査2例によるトピックの難しさの違いは，順位の交替が起こり得るほどに僅差であったと言える。

一方，評価者の特性が変化に富んでおり，採点の厳しさと一貫性にばらつきが見られた。表10を参照すると，特に評価者4と評価者10はInfit MSが基準値（Infit MS < 1.30）を超えており，採点の一部に判断の揺らぎがあったと見られる。

採点の厳しさの面でlogitに注目すると，トピックよりも影響の大きい組み合わせが見られた。例えば，最も厳しい評価者4と，最も緩い評価者5を比べた場合，彼らのlogit差は3.17［logit = 1.83 - (-1.34)］で，6段階評価に換算すると，1段階強の配点の格差ができることがわかった（公正化された平均得点：3.80 - 2.50 = 1.30）。増員によって，異なる評価者の特性を観測したことが原因の1つに考えられるが，一方でこの状況を教育現場に当てはめた場合，科目担当の独自採点にゆだねることが，いかに不安定な結果を生むか注意しなくてはならないだろう。

6.3.2 受験者とトピックとの誤差分析

では，以上のライティング評価の実態を踏まえて，テスト実施期間に熟達度の大きな変化がないとすれば，同一の学習者にはどこまで一貫した得点結果が導かれるだろうか。表11はFACETSの誤差分析（bias analysis）をまとめたものである。この分析では，それぞれの作文に与えられた評価者7名の採

■ 表9：トピックのラッシュ・モデル分析結果（その2）

トピック	実測			公正化された平均得点	Logit	Infit MS
	得点	個数	平均			
A	540	154	3.51	3.30	-.35	1.14
B	989	315	3.14	3.10	.23	.99
C	507	154	3.29	3.20	.04	.82
E	984	315	3.12	3.20	.08	.92
Mean	755	234.5	3.22	3.20	.00	.97
SD	231.8		.20	.09	.21	.12

■ 表10：評価者のラッシュ・モデル分析結果（その2）

評価者	実測		公正化された平均得点	Logit	Infit MS
	個数	平均			
1	58	2.90	3.00	.61	.69
2	58	3.40	3.30	-.31	.67
3	58	3.40	3.40	-.59	1.11
4	58	2.50	2.50	1.83	1.68
5	58	3.80	3.80	-1.34	1.10
6	58	3.30	3.30	-.18	1.15
7	98	3.20	3.10	.30	.70
8	100	3.40	3.30	-.37	1.10
9	97	3.20	3.20	.12	.64
10	97	3.20	3.20	.11	1.62
Adjuster	64	3.50	3.50	-.65	.38
Master	134	3.00	3.00	.46	.87
Mean	78.2	3.20	3.21	.00	.98
SD		.30	.30	.76	.39

点結果に基づき，*t* 検定を行っている。表中では，5％水準で有意確率の低いものから順次に示し，2回のテストで与えたトピック，受験者の logit，及び6段階評価に換算した得点（最確得点）を表している。

この結果によると，2回のライティング・テストで実に67名のうち20名の受験者から，有意な得点差が検出された。しかし特定のグループや，トピック組み合わせ，及び提示順序に特徴的な傾向はなかった。つまり，グループ平均で比較した logit の結果（5.3.2，6.3.1参照）と同様，トピックの難しさに大きな違いはなく，それが全体的評価で深刻な不平等には結び付かないと結論できる。ただし，テスト得点の一貫性を考慮した場合，およそ3割の受験者で得点差が生じることは決して望ましい状況とは言えない。では，これらの得点の不一致は，どのようにして起こったのであろうか。

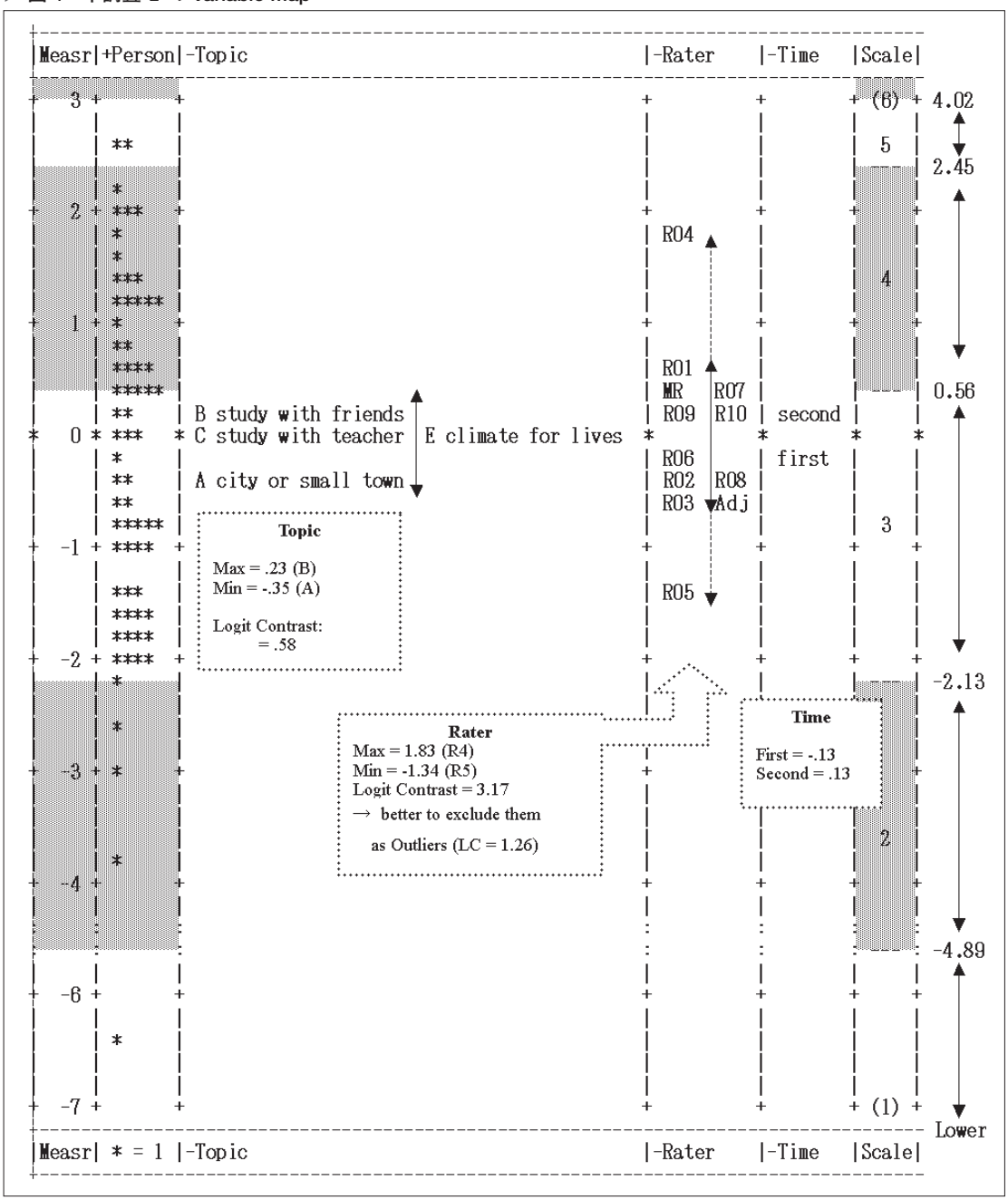
図4は本調査2におけるライティング・テストの採点状況を，FACETS の Variable Map で表したものである。Variable Map はテストにかかわる相（facet：受験者，トピック，評価者など）の logit を上下の位置関係で表し，それぞれの影響力の強さが視覚的に一覧できる。本調査のモデルでは受験者（Person）が左端の列を表し，記号（＊）が上に位置するほど能力の高い受験者が分布していることを示す。一方，左端から数えて2～4番目には，それぞれトピック（Topic），評価者（Rater），テスト回数（Time）の相が表示されている。これらは負の相（negative facets）で定義されており，高い位置にあるトピックほど難しく，評価者なら厳しい評価をする傾向を示している。

また，調査で用いた評価得点で受験者の能力を解釈したい場合，右端の列にある目盛（Scale）がその役割を果たしている。例えば，今回の受験者で logit = 0.00と測定された場合，6段階評価では3点を与えられる確率が最も高く（most probable score：最確得点），同様に logit = 1.00であれば最確得点は4点となる。なお，ここで「確率」という言葉を用いて説明するのは，実測の採点結果が評価者の厳しさなどの影響で，必ずしも logit の示すとお

■ 表11：誤差分析レポート

通し番号	Test 1				Test 2				*t* 検定		
	トピック	実測平均	Logit	最確得点	トピック	実測平均	Logit	最確得点	Logit対比	*t*	*p*
1	B	3.9	1.42	4	E	2.3	-2.30	2	-3.72	-4.51	.001
2	A	3.7	.63	4	B	2.0	-3.08	2	-3.71	-4.38	.001
3	E	3.1	-.18	3	C	2.0	-3.59	2	3.41	3.88	.002
4	B	3.1	-.20	3	A	2.3	-3.34	2	3.14	3.62	.004
5	B	2.3	-2.35	2	A	1.6	-5.44	1	3.09	3.32	.006
6	B	2.9	-.63	3	A	4.4	1.90	4	-2.53	-3.16	.008
7	E	3.1	-.35	3	C	4.3	1.99	4	-2.34	-2.96	.012
8	A	4.1	1.22	4	B	2.7	-1.10	3	-2.32	-2.87	.014
9	C	2.6	-2.03	3	E	3.3	.28	3	2.31	2.71	.019
10	E	2.9	-.85	3	B	3.9	1.36	4	-2.21	-2.70	.019
11	E	2.4	-2.02	3	B	3.4	.27	3	-2.29	-2.70	.020
12	A	4.3	1.60	4	B	4.9	3.55	5	1.95	2.64	.022
13	A	4.6	2.12	4	B	3.1	.08	3	-2.04	-2.58	.024
14	E	2.3	-2.62	2	C	3.1	-.42	3	-2.20	-2.54	.026
15	B	4.1	2.08	4	A	3.7	.22	3	1.87	2.46	.030
16	B	3.0	-.54	3	E	2.3	-2.69	2	-2.14	-2.45	.031
17	B	2.9	-1.06	3	E	3.6	.98	4	2.03	2.45	.031
18	A	3.4	-.26	3	B	4.0	1.63	4	1.89	2.43	.032
19	A	4.7	2.25	4	B	3.3	.37	3	-1.88	-2.40	.034
20	E	2.3	-2.53	2	C	3.1	-.57	3	-1.96	-2.27	.042

▶ 図 4：本調査 2 の Variable Map



りに返ってくるとは限らないからである。

そこで，2 つの作文テストにおける評価得点の不一致について，改めて原因を振り返ると，トピックと評価者の列にある両矢印に注目されたい。これらは logit の最大と最小の幅を示し，トピックや評価者を交替させた場合，実際に受験者に与えられる得点がどのくらい変動するのか，極限の可能性を示している。この図示からも明らかなように，トピックによって受験者の logit が変動する幅はわずかであり，むしろ評価者の厳しさの違いで 2 つの最確得点を行き来する受験者の方が多い。

実際のエッセイを採点する際には，常に複数の評価者が完全に一致した判断を返すとは限らない。端的な例を挙げるなら，「評価者 1 名が 5 点，3 名が

4点，2名が3点」と，評点のいくつか分かれる作文採点も見られた。同じ評点の最確得点に属す受験者群でも，上下に幅広く分布するのは，そのためである。今回の採点結果の不一致で問題となった20名分の作文は，その多くが2～4点の作文だったことから，しきい値［logit = .56（Score 3-4）; logit = -2.13（Score 2-3）］の周辺に位置した受験者の得点だったと考えられる。

6.4 本調査２の考察

改めてRQ3について考察すると，ラッシュ・モデルによる全体的評価の分析からは，トピックから評価得点への影響を懸念する必要はないとみられる。しかしながら，発表語彙知識が限られた学習者にとって，ライティング・パフォーマンスを引き出すことは，繊細な行為であり，テストの実施条件や受験者のコンディションにも左右されるようである。また，全体的評価は評価者の主観や価値判断に依存する部分も多く，採点結果に信頼がおけるかどうかは評価者の技量が重要となってくる。

特にRQ4の検討にあたり，同一の書き手で複数回のライティング・テストを実施した場合，おおむね同等の評価得点を導いていたが，厳密な成績資料として扱うには，より高い精度が求められるだろう。本調査では6段階評価基準を使用し，評価得点が一致したのは67名のうち47名（全体で約70%）にとどまった。その一方，パフォーマンスに明らかな不均衡を示した2名を除くと，残り18名から各2篇の作文には1点ずつ評価得点の違いが見られた。これら採点で現れた差異をいかに処理するかが，教育現場でパフォーマンス評価を普及させていく重要な課題となる。

考えられる解決策の1つとして，第1に評価者の訓練が挙げられる。既に先行研究で実証されているように，訓練の有効性は認められ（Weigle, 1994），日本人の英語科教師を評価者とした場合でも，ある程度の成果が報告されている（占部, 2007）。ただし，Weigle（1998）によると，トレーニングは評価者の判断の一貫性を修正する上で有効だが，評価者個々の厳しさを画一化させるほどではないことも指摘されている。また採点の訓練に多くの時間を割くことができないのが教育現場の実情でもあろう。これらを総合的に考慮した場合，評価者訓練にかける時間はできる範囲で確保しつつ，妥協点を見いだすことが必要と考える。

そこで第2の提案として，生徒作文を複数名で評価できる教員の組織体制作りを掲げる。日本の教育現場では伝統的に，単独の科目担当が評価を担うことから責任が重く，先進的な評価方法を採用することが難しい。しかし今回の評価結果の誤差が，段階得点同士のしきい値周辺で起こっているなら，評価者判断の一貫性を追求するだけでは限界があり，評価基準で作文特徴の分類を精緻にしていく必要もあるだろう。ただし10段階，15段階と得点尺度を増やすことは，判別できる限界を超えた採点作業を強要し（Weir, 2005），採点に使われず機能しない得点尺度も出現するため，好ましい改善策とはいえない。一方，TOEFL TWEでは，実用に適った採点手続きをテスト実施の初期から採っていたことが改めて実感される。ゆえに，以下でGolub-Smith et al.（1993）のリサーチ・レポートをもとに詳述する。

彼らによると，TOEFL TWEの採点手続きは専門的に訓練された評価者2名が採点し，1点差の不一致があった採点結果は平均化，つまり厳しい評価得点の方から0.5点を加えた得点を採用している。もし2点以上の不一致があった場合には，より上級の評価者が採点に加わり，一致した2名の得点，または近い評価者間の平均を採用している。こうした手続きを採ることにより，独断で公正さに欠けた採点を防ぐだけでなく，現行の6段階評価で実質11段階のきめ細かな分類が可能となる。

7 まとめ

豊かな表現力と，グローバル化していく社会に対応した言語能力を獲得するためにも，生徒自らが考えや意見を伝える活動は一層必要となる。しかし，総合的な言語パフォーマンスやコミュニケーション能力を評価するには，公平さという点で教育現場での課題が先送りされたままで，実践に二の足を踏む教員も少なくなかった。本研究では，ライティング指導から評価への問題にあたり，以下3点の教育的示唆を提案する。

(1) 学習者の言語知識に適した課題の提示
(2) 評価者と評価基準の技術的向上
(3) 言語パフォーマンスの指導と評価への環境改善

第1に，タスク選びは教師にとって重要な任務である。タスクの困難度を吟味し，学習者の言語発達に応じた提示順番を整備するのが望ましい。まず，学習者の年齢に相応の内容か，求められるパラグラフ構造の理解は十分か吟味すべきである。トピックの観点からは，使用が予測されるテーマ語彙を習得済みか，把握しておくべきである。テーマ語彙が未習得の学習者には，トピックの提示を見合わせ，または実施前に語彙の補充を行うべきである。可能であれば，過去の生徒作文から学習者コーパスを構築し，トピックが引き出す語彙使用範囲の見積もりを長期的に蓄積することを勧める。また，学校教員間でのコーパス構築技術の普及，及び連携による大規模な学習者コーパスの確保が，今後のライティング指導の一助となる可能性は十分あるだろう。

第2に，テスト実施時期を越えた対等なパフォーマンス評価の比較を検証した結果，タスクの含むトピックの特性より，むしろ教員間の判断によって誤差の生じる可能性が明らかとなった。こうした状況に遭遇したときこそ，教員間が訓練する機会を確保し，評価技術の維持に努めるべきである。具体的には，今回の研究で用いた6段階評価であれば，少なくとも2段階以上の誤差を解消できるくらいまで検討し，学習者の習熟度に合わせて評価基準の精緻な解釈に努めるべきである。また学期ごとの成績評価など，重要な評価には，単独ではなく複数教員で採点セッションを実施し，平均得点と判断の根拠を書き込んで与えることが望ましいだろう。

最後に，現状の日本の教育現場で言語パフォーマンス評価の普及には，教員間で一層の修練が必要である。その一方，近年の商用テストでは実社会に使える技能の育成を前面に押し出しており，学習者の言語能力を直接評価する流れにある。こうしたテストの波及効果からも，教室で学習者のニーズに応じた指導が求められることだろう。今後，より多くの学校で，平素から生徒と教師がコミュニケーション活動を共有し，互いの言語観と意思伝達する態度とを育成できる環境が備わることを願うばかりである。

謝　辞

本研究の機会を与えてくださいました（財）日本英語検定協会の皆様，そして選考委員の先生方に厚く御礼申し上げます。

筑波大学大学院の磐崎弘貞先生には，調査の実施，コーパスの分析，論文執筆に至るまで貴重なご助言を賜りました。いつもお忙しい折にもかかわらず，進捗状況を定期的に報告する機会をいただき，推敲を丁寧にご校閲してくださいました。心より感謝申し上げます。

同じく筑波大学大学院の平井明代先生には，統計分析で有益なアドバイスをいただきました。とりわけ項目応答理論の基礎からデータ処理に至るまで，貴重な勉強の機会をいただき，大変お世話になりました。深く感謝申し上げます。

また，作文採点には多くの院生の皆様と現職の先生方からご協力をいただきました。改めて感謝申し上げます。そして，本調査に際しましては，科目担当の先生方のご厚意と数多くの学生のご協力なくしては実現できませんでした。この場をお借りして，お礼を申し上げます。本当にありがとうございました。

参考文献（＊は引用文献）

＊秋山朝康.（2002）.『スピーキングテストの分析と評価―項目応答理論を使っての研究―』. *STEP BULLETIN*, vol.12, 67-78.

＊Carlson, J.G., Bridgeman, B., Camp, R., & Waanders, J.（1985）. *Relationship of admission test scores to writing performance of native and nonnative speakers of English*（TOEFL Research Report 19）. Princeton, NJ: ETS.

＊Crowhurst, M.（1980）. Syntactic complexity and teachers' quality ratings of narrations and arguments. *Research in the Teaching of English*, 14, 223-231.

＊大学英語教育学会基本語改訂委員会（編）.（2003）.『大学英語教育学会基本語リスト』. 東京：大学英語教育学会.

＊Daller, H., van Hout, R., & Treffers-Daller, J.（2003）. Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24, 197-222.

＊Du, Y., Brown, W.I., & Rogers, C.（1997）. *Raters and single prompt-to-prompt equating using the FACETS model in a writing performance assessment*.（Eric Document Reproduction Service No. ED410 291）.

＊Educational Testing Service（ETS）.（2005）. TOEFL: *Test of written English guide*（5th ed.）. Princeton, NJ: Author.

＊Ellis, R.（Ed.）（2005）. *Planning and task performance in a second language*. Amsterdam and

Philadelphia: John Benjamins.
＊ETS.(n.d.). (2005). Criterion online writing evaluation [Web file]. Retrieved May 18, 2007, from http://www.ets.org/Media/Products/Criterion/topics/topics.htm
＊Golub-Smith, M.L., Reese, C.M., & Steinhaus, K.S.(1993). *Topic and topic type comparability on the Test of Written English*.(TOEFL Research Report 42). Princeton, NJ: ETS.
＊Hamp-Lyons, L.(Ed.).(1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.
＊Hamp-Lyons, L. & Mathias, S.P.(1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3, 49-68.
＊Hughes, A.(1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
＊Hyland, K.(2003). *Second language writing*. Cambridge: Cambridge University Press.
＊金谷憲.(1993).「和文英訳からライティングへ：内容表現のための作文指導」.『英語展望』99号, 24-29.
＊Kroll, B. & Reid, J.(1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing*, 3, 231-255.
＊Laufer, B.(1994). The lexical profile of second language writing: Does it change over time?, *RELC Journal*, 25, 21-33.
＊Laufer, B., & Nation, P.(1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307-322.
＊Linacre, J.M.(2006). *A user's guide to FACETS*. Chicago: MESA Press.
＊Lougheed, L.(Ed.).(2004). *How to prepare for the TOEFL essay: test of English as a foreign language* (2nd ed.). Barron's Educational Series, Inc.
＊McNamara, T.(1996). *Measuring second language performance*. Essex, U.K.: Longman.
＊文部省.(1999).『高等学校学習指導要領解説―外国語編 英語編』. 東京：開隆堂.
＊大友賢二.(1996).『項目応答理論入門』. 東京：大修館書店.
＊Read, J.(2000). *Assessing vocabulary*. Cambridge University Press.
＊Reid, J.(1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178-210). New York: Cambridge University Press.
＊清水伸一.(2006). v8an - revised web edition [Computer program]. Retrieved May 24, 2007, from http://www01.tcp-ip.or.jp/~shin/j8web/j8web.cgi
＊鈴木秀幸.(2002, June).「新指導要領下で評価はどう変わるか」.『英語教育』51号, 8-11.
＊ToeflEssays.com.(2004). Sample Essays for the TOEFL Writing Test (TWE) - E-Book Edition [Computer software]. Retrieved December 23, 2006, from http://www.lulu.com/content/56371
＊投野由紀夫 (2003).「第5章　語彙指導にコーパスを利用する」.(引用元) 望月正道・投野由紀夫・相沢一美.『英語語彙の指導マニュアル』(pp. 145-79). 東京：大修館書店.
＊投野由紀夫 (編).(2007).『日本人中高生一万人の英語コーパス』. 東京：小学館.
＊占部昌蔵.(2007).「項目応答理論を応用した英作文評価者トレーニングの有効性について」. *STEP BULLETIN*, vol.19, 14-22.
＊Weigle, S.C.(1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.
＊Weigle, S.C.(1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.
＊Weigle, S.C.(2002). *Assessing writing*. Cambridge: Cambridge University Press.
＊Weir, C.J.(1990). *Communicative language tesing*. UK: Prentice Hall International Ltd.
＊Weir, C.J.(2005). *Language testing and validation: An evidenced-based approach*. New York: Palgrave Macmillan.
＊Willis, D.(1990). *The Lexical Syllabus: A New Approach to Language Teaching*. London: Collins ELT.

資　料

資料 1：ライティング・テスト使用トピック（日本語訳）

A. 小さな町に住みたがる人もいれば，大きな都市に住みたがる人もいます。あなたはどちらに住みたいと思いますか。その答えの根拠となる，具体的で詳しい理由を入れてください。
B. ひとりで勉強したいと思う生徒もいれば，みんなで勉強したいと思う生徒もいます。あなたはどちらを好みますか。その答えの根拠となる，具体的な理由や事例を入れてください。
C. 先生と一緒に学ぶより独力で学んだ方がよいと考える人もいれば，先生と一緒に学ぶ方が常によいと考える人もいます。あなたはどちらを好みますか。その答えの根拠となる，具体的な理由や事例を入れてください。
D. 映画には，まじめで観客に深く考えさせる作品もあれば，愉快で楽しませる作品もあります。あなたならどちらのタイプの映画を好みますか。その答えの根拠となる，具体的な理由や事例を入れてください。
E. 年中同じ天候の続く場所に住みたがる人もいれば，年間で何度も天気の変わる地域に住みたがる人もいます。あなたはどちらを好みますか。その答えの根拠となる，具体的な理由や事例を入れてください。
F. じっくりと自由時間の活動計画を立てたがる人もいれば，一切計画を立てない人もいます。あなたならどちらを好みますか。その答えの説明となる，具体的な理由や事例を入れてください。

資料 2：熟達度テスト結果

	トピック（*n*）	観測		公正化された平均得点	Logit	Infit MS
		得点	平均			
本調査 1	A（24）	881	.73	.78	-.02	.98
	B（24）	875	.73	.79	.02	1.01
	C（24）	867	.72	.79	.01	.97
	D（23）	843	.73	.79	.03	1.02
	E（23）	838	.73	.79	.02	1.03
	F（23）	843	.73	.79	.00	1.02
本調査 2	E&B（11）	411	.75	.79	.03	.96
	E&C（10）	373	.75	.79	.03	1.02
	B&A（11）	382	.69	.78	-.04	.98
	B&E（12）	407	.68	.78	-.03	.99
	C&E（12）	406	.68	.78	-.04	1.06
	A&B（11）	382	.69	.79	.01	1.00
	Mean	625.7	.72	.79	.00	1.01
	SD	243.1	.03	.00	.03	.03

（注）信頼性係数は FACETS より separation reliability に基づく。

項目：r = .97; 受験者：r = .73。

資料 3：Criterion Scoring Guide（日本語訳）

6　Excellent
・立場をわかりやすく明言しており，読み手に議論が妥当であることを効果的に説得できている。 ・エッセイ全体を通して多くの具体例，及び関連のある詳細な記述によってアイデアが展開している。 ・論理展開が明瞭で効果的に構成されており，焦点が定まっている。 ・多様な構文を使っている。 ・具体的な語彙の選択が多く見られる。 ・文法や綴り・句読法の誤りはほとんど，もしくは全く見られず，誤りがあっても理解に支障がない。
5　Skillful
・立場を明らかにしており，読み手を十分に説得している。 ・いくつかの具体例，及び関連のある詳細な記述によってアイデアが展開している。 ・わかりやすく構成し，情報提示に整然さが見られる。ただし論理展開が不十分である。 ・いくらか多様な構文を使っている。 ・具体的な語彙の選択が見られる。 ・文法や綴り・句読法の誤りが見られるが，理解には支障がない。
4　Sufficient
・立場を明らかにしており，読み手を説得する試みが十分見られる。 ・明確なアイデアを提示しているが，展開が不十分で，具体例に欠ける。 ・わかりやすく一連の情報を提示しており，おおむね，そのつながりも示されている。 ・少し構文の正確さにムラがある。 ・主に基本語彙を使っているが，具体的な語彙の選択も時折見られる。 ・文法や綴り・句読法の誤りが多少見られるが，ほとんど理解には支障がない。
3　Uneven
・いずれの立場か明らかにしているが，不明瞭もしくは展開が不十分である。 ・提示される情報が限られているか不完全であり，情報の列挙あるいは概要程度でしかない。 ・載せている情報のまとまりが悪い。またはつながりのない情報を提示したり，関係の見えないものを扱っている。 ・構文の正確さにムラがある。 ・いくつか不適切な語彙選択が見られる。 ・理解に支障となる文法や綴り・句読法の誤りが，時折含まれる。
2　Insufficient
・明確な立場を示してない，または説得しようとする試みがほとんど見られない。 ・情報が少なく，内容を展開させる努力がほとんど見られない。 ・非常に構成のまとまりが悪い，または短すぎて構成がわからない。 ・構文の正確さがほとんど見られない。 ・多くの個所で不適切な語彙が選択されている。 ・綴りの誤り，単語の脱落，語順の誤りや，文法や句読法の誤りが深刻なために，大部分で非常に理解が難しい。
1　Unsatisfactory
・どちらの立場もとっていないか，理由が示されていないため，擁護・説得しようとする試みがほとんど見られない。 ・解答しようとしているが，課題文を単に言い換えただけだったり，極端に短い。 ・構成面でのまとまりが見られない。 ・構文が全く成立していない。 ・大部分で不適切な語彙が選択されている。 ・綴りの誤り，単語の脱落，語順の誤りや，文法や句読法の誤りが深刻であるために，全体を通して理解することができない。