

指導と評価の一体化をめざした信頼性の高い英作文 評価基準表の作成：多変量一般化可能性理論を用いて

東京都／津田塾大学大学院後期博士課程在籍 大久保 奈緒

申請時：東京都／津田塾大学 演習助手

概要

本研究においては、英作文評価基準表を作成し、その評定項目及び、評定者に関する信頼性の検討を多変量一般化可能性理論や評定者フィードバックを用いて検討した。この評価表は、ジャンル分析研究を参考に作成された。内容、構成、語彙、言語使用の4観点から成立し、各観点到、3から4の下位項目が設置されている。3人の英語母語話者である英語教師が、41人の大学生が書いた英作文を、この評価表を用いて評定した。多変量一般化可能性理論を用いた分析では、信頼性の高い結果が導き出された。しかし、語彙と言語使用の多変量一般化可能性係数、多変量信頼度指数が、内容及び構成に比べ信頼性の低い結果となり、前者2観点については改善が示唆された。また、評定者フィードバックから、内容・構成の採点の際に、評定者が過去の経験から構築された内的基準と本評価表との間で、すり合わせを行っている様子が浮かび上がった。

1 はじめに

2003年に発表された『『英語が使える日本人』の育成のための行動計画』（文部科学省、2003）の中では、実践的コミュニケーションが強調され、英語をコミュニケーションの手段として使用し、4技能の育成を図ることが推進されている。このような流れとともに、2004年には英検において1級に自由英作文が、準1級に記述式問題が導入されるなど、和文英訳や一文単位の英作文に限らない、まとまりのある英文を書く能力が求められる傾向が強まっている。しかし、英文ライティングの評価は評価観点が多岐にわたり、複雑であるため敬遠されがちである。本研究においては、英文ライティングの指導内容を反映し

た英作文評価基準表（以下、評価表）を作成し、その採点項目及び、評定者に関する信頼性の検討を多変量一般化可能性理論や評定者フィードバックを用いて検討する。

2 研究の背景

2.1 英文ライティング指導と理論

英文ライティング指導のために、英文ライティング理論は重要な役割を果たしてきた。例えば、書き手の認知構造に焦点を当てる研究から「下書き」や「推敲」などの過程を重視するプロセスアプローチが発展した（Grabe & Kaplan, 1996）。大学の英作文の授業では、exposition（論説文）、description（記述文）などのジャンルを指導していることが多いことを踏まえ、ジャンル研究をもとにしたライティング理論を本研究では用いる。ジャンル研究をもとにしたライティング理論は大きく3つの流れに分けられる（Hyland, 2002; Johns, 2003）。機能言語学（Systemic Functional Linguistics: SFL）を基盤とする研究、English for specific purpose（ESP）/ English for academic purpose（EAP）研究、the New Rhetoric 研究である。ESP/EAP 研究とは、特定の目的を共有する特定の集団の中で望ましいとされる文章構成を研究する方法である。例えば、ビジネス分野での覚書や報告書の書き方である。単なる記述文や論説文にとどまらない高度な英語能力が必要とされる特殊分野のライティングに焦点を当てている。the New Rhetoric とは、テキストの構造だけでなく、ジャンルと頻繁に起こる状況を結び付けて分析する研究である。この the New Rhetoric は分析手法の意味合いが強く、教育現場への応用へは消極的

な立場をとっている (Johns, 2003)。

一方, SFL を基盤とするジャンル研究では, 文法や単語などの言語形式は文脈に応じて異なるという視点で教育現場に応用されている。SFL によるジャンル研究を教育現場に応用しているオーストラリアのニューサウスウェールズ州の初等中等教育課程では, このようなジャンル中心の言語教育を, 協同学習などを通じて行い, 成功を収めている (Hyland, 2002)。ニューサウスウェールズ州教育省のシラバスでは「文脈 (目的, 聞き手 / 読み手, コミュニケーションや内容を媒介する手段) に応じて言語の異なった使われ方についての知識を高めること」を求めている (Board of Studies NSW, 1998, p.5)。このシラバスでは, ジャンル (論説文, 情報伝達文など) ごとに, そのジャンルで求められている目的, 文章構造, 文法項目などを明確に示している。例えば, 情報伝達文における目的は自然科学や社会環境のトピックに関して情報を示すことである。情報伝達文における構造は, General statement identification ^ Description of Aspect ^ Description of activity で表される。文法項目では, 動作動詞や現在形の使用が望ましいとされている。このようにある特定のジャンルを学ぶ際には, そのジャンルに適切な文法・語彙学習を指導することを推奨している。このため, ジャンルを中心に教えている日本の大学の英作文教育にとって有益な示唆を与えてくれると言える。

しかし, このシラバスにおける評価項目は, 英語を外国語として学習している学生の英作文の評価に適切であるとは言えず, ジャンルの特徴が示された別の評価表が必要である。

2.2 ライティング評価

2.2.1 評価

評価には主に妥当性 (validity), 信頼性 (reliability), 真正性 (authenticity), 実用性 (practicality / feasibility) が必要であり, ライティングの評価においてもこの4点は重要である (Bachman & Palmer, 1996; Hughes, 2003; Brown, 2004)。

まず, 妥当性 (validity) とは, テスト内容がテストの目的を適切に測定しているかどうかである。Grabe and Kaplan (1996) によると, 英作文は多肢選択式などの間接的測定方法に比べ, 実際にまとまりのある英文を書かせるため, ライティング能力を適切に測っていると言える, 妥当性が高いとされる。

次に信頼性 (reliability) とは, 評価結果が一貫しているかどうかである。金谷 (2003) によると, 信頼性とは「同じ能力を持った受験者に対して同じテストを実施した場合に, 常に同じ結果を出すかどうか」(p.69) である。つまり, 測定における誤差を減らし, 測定結果の安定性を求めるものである。

第3に, 真正性 (authenticity) とは, 使用するタスクが測りたい対象を測っているかどうかである。例えば, ビジネスライティングのためのテストで, メモを取るタスクや覚書を作るタスクではなく, 物語を書かせるタスクを行った場合, 測定対象とタスクの間にずれが生じるため, 真正性が低いと言える。

第4に, 実用性 (practicality / feasibility) とは, 時間, 費用, 評定者数, 設備などのテスト実施における物理的条件である。例えば, 作文1枚に対し, 評定者が10人必要とされるテストや拘束時間の長いテストは実用的ではないのである。

2.2.2 ライティング評価における妥当性の検討

ライティング評価では, 多肢選択式などの間接的評価が行われてきたが, 妥当性が疑問視され, 近年直接的評価に移行している (Hamp-Lyons, 1991)。Grabe and Kaplan (1996) は, 「良いライティングテストは生徒に実際にライティング活動をさせること」(p.399) と述べている。

ライティングを直接測る際には, 主に全体的評価 (holistic scoring), 分析的評価 (analytic scoring), 特定要因の評価 (primary trait scoring) の3種類の評価方法が用いられる。全体的評価とは, 1人または複数の評定者が, 評価表もしくは評定者個人の印象をもとに全体評価として1つのスコアをつける方法である。例えば, TOEFL iBT の independent task では, 6段階で全体的評価を行っている。しかし Weigle (2002) は, 全体的評価は1つのスコアしか与えられないため, 分析的評価ほど妥当性が十分ではないと述べている。また, 全体的評価には, 診断的フィードバックのような付加的情報が乏しいと指摘されることが多い (Hamp-Lyons, 1991, 2003)。このような欠点もあるが, 分析的評価に比べ短時間で採点が可能なことから, 実用性が高い評価方法であるとされ, 大規模テストにおいて使用されることが多い。

一方, 分析的評価は, ライティング能力は複数の要因から成立しているとの前提で行われる評価であ

る。Weigle (2002) は、英語学習者が異なった要因(文法、内容、語彙など)に関して、それぞれ伸びていく様子がわかるため、妥当性が高いとしている。ライティングにおける分析的評価としては、Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) の ESL composition profile が最もよく知られている。

特定要因の評価とは、タスクごとに、特定の状況や読み手を想定し、採点基準を示した評価である。Weigle (2002) は教室活動において、生徒が何をできるようになったかを詳しく示すことができる点で、特定要因の評価の利点を述べている。しかし、タスクごとに採点基準を用意するなど、時間や費用などの手間がかかりすぎるため、使われないことが多い。

このように、分析的評価は、ライティングを構成する複数の要因を測っているため、全体的評価や特定要因の評価に比べ妥当性が高いと言える。したがって、分析的評価は、学習者のライティング能力に関する細かい情報を得られる点で、指導内容を確認することができ、大学などの英語教育プログラムにおける英文ライティング科目の評価に適切であると言える。

2.2.3 直接的評価における信頼性の検討

L2ライティングのような行動評価/言語運用評価(performance-based assessment)を行う場合、タスクは実際の場面に近い内容を測ることが求められる(McNamara, 1996)。言語運用を評価する場合、妥当性、真正性の高いタスクが求められると同時に、信頼性の高い結果を導く必要がある。このような言語運用タスクにおいて、採点結果の信頼性を測る方法として一般化可能性理論や多相ラッシュ測定などが挙げられる(Bachman, 2004)。多相ラッシュ測定ではどの評定者が易しく、もしくは厳しく点数をつけたかについて特定化することや、どの項目が特定の受験者にとって難しかったかなど、一般化可能性理論に比べ多くの情報が得られる。しかし本研究では、項目ごとの困難度や個々の評定者に焦点を当てるより、むしろ、評価表の項目数及び評定者数を検討することや、項目全体としてどのような解釈が可能かを調べるのが主な目的であるため、一般化可能性理論を用いる。

2.3 一般化可能性理論

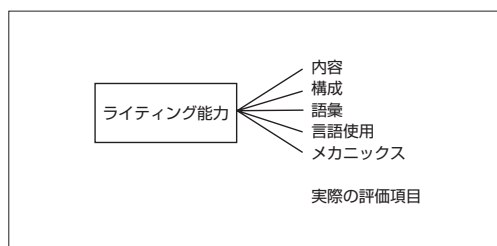
2.3.1 多変量一般化可能性理論

一般化可能性理論とは、観測得点は真値と誤差から成立していることを前提とする古典テスト理論をもとにしている。一般化可能性理論では、測定値に含まれる誤差の原因や誤差の大きさの推定を、分散分析を用いて行う。例えば、テスト項目が誤差の原因になりうるのか、またそれはどの程度の誤差であるのかについて検討できる(山森, 2004)。このような一般化可能性理論を拡張した手法が多変量一般化可能性理論である。池田(1994)は、多変量一般化可能性理論とは、「多数の対象(能力)について同時に測定された複数の測定値プロフィールとそれらの共分散成分の推定を扱う」(p.50)と述べている。多変量一般化可能性理論の利点は、より細かな下位の評価項目を設定できることである。一般化可能性理論では、ライティング能力を、例えば、5つの観点からしか分析できない(図1)。一方、多変量一般化可能性理論では、観点ごとにより詳細な下位項目が設定でき、より多くの視点でライティング能力を分析できる(図2)。

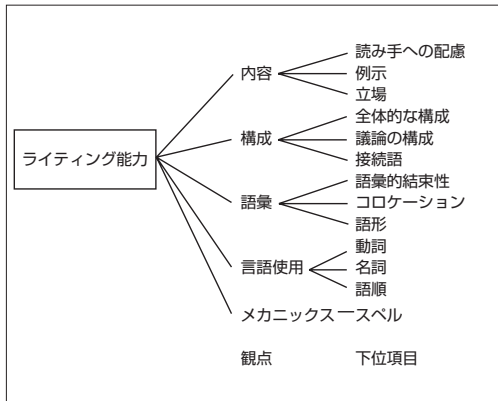
図2のように、多変量一般化可能性理論は、多数の観点(例:内容、構成、語彙、言語使用)の中で、同時に測定された複数の測定値(『内容』の複数の測定値の例:読み手への配慮、例示など)の分散成分と、共分散成分の推定を行うことが可能である。

多変量一般化可能性理論は一般化可能性理論と同様に、一般化可能性研究と決定研究に分かれる。測定の際に現れる誤差のそれぞれがどの程度観測値に影響を与えているかについて、分散分析を用いて推定する。これを、一般化可能性研究と言う。そしてこの誤差の成分のばらつきの大きさ(分散成分・共分散成分、または分散成分・共分散成分の推定値とも言う)をもとに、多変量一般化可能性係数や多変

▼ 図1：一般化可能性理論で測定可能なライティング能力の概念の例



▼ 図2：多変量一般化可能性理論で測定可能なライティング能力の概念の例



量信頼度指数を求め、これらの係数や指数を用いて、適切な評定者数や項目数などを特定化することを決定研究と言う。決定研究の際、相対的な解釈及び集団準拠評価の場合は多変量一般化可能性係数を用い、絶対的な解釈及び目標準拠評価の場合は多変量信頼度指数を用いる (Brennan, 2001a)。

このうち、一般化可能性研究においては、分散と共分散の推定を行う。分散とは、観点の得点のばらつきを示している。一方共分散とは、観点が組み合わされた時のばらつきを示している。例えば、内容と構成という2つの観点が組み合わされたばらつきである。多変量一般化可能性係数を求めるためには、母得点と相対誤差の分散・共分散を求める必要がある (Brennan, 2001a)。母得点の分散・共分散は、

$$\sigma^2 c(\tau) = \sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\tau) \quad (2.1.)$$

相対誤差の分散・共分散は、

$$\sigma^2 c(\delta) = \sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\delta) \quad (2.2.)$$

で表すことができる。 $\omega_{vv'}(\tau)$ はユニバーススコアの共分散であり、 $\omega_{vv'}(\delta)$ は相対誤差の共分散である。相対誤差は測定対象 (例：受験者の得点) と交互作用するすべての変動要因の分散成分、共分散成分の和となる。 ω は分散共分散成分の各要素への重み付けであり、 v は観点を示す。多変量信頼度指数を求める際には、母得点の分散の他に、絶対誤差の分散が必要である。絶対誤差は、測定対象以外のすべての変動要因の分散成分と共分散成分の和

となる。絶対誤差の分散は、

$$\sigma^2 c(\Delta) = \sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\Delta) \quad (2.3.)$$

で求められる。このように、分散共分散行列の推定を行うことを一般化可能性研究と言う。

決定研究では、一般化可能性研究で求められた分散、共分散の値をもとに、多変量一般化可能性係数及び多変量信頼度指数を求める。

多変量一般化可能性係数は、

$$E \rho^2 = \frac{\sigma^2 c(\tau)}{\sigma^2 c(\tau) + \sigma^2 c(\delta)} \quad (2.4.)$$

と表される。

そして、多変量信頼度指数は、

$$\Phi = \frac{\sigma^2 c(\tau)}{\sigma^2 c(\tau) + \sigma^2 c(\Delta)} \quad (2.5.)$$

で求められる。

項目数や評定者数の数を変えながら、多変量一般化可能性係数及び、多変量信頼度指数の変化を予測する。そして、そのシミュレーションに基づき次回の評価で望ましい評価計画を立てることができる点が多変量一般化可能性理論の利点である。

これら一般化可能性係数及び信頼度指数は、古典テスト理論における信頼性係数と見なすことができる。このため一般化可能性係数及び信頼度指数が.80以上になれば、評定結果は信頼できると解釈しうる (山森, 2004)。しかし、ライティング研究においては、2人以上の評定者で、強い妥当性がある場合、教育場面においては.75以上の信頼性でも可とされる (Hamp-Lyons, 2003)。本稿では、多変量一般化可能性係数及び多変量信頼度指数に関しては.75を信頼性がある値として見なし、もし可能であれば.80の数値をめざすことにする。

2.3.2 多変量一般化可能性理論を用いた研究

一般化可能性理論は現在、外国語教育研究において、言語運用を測る際に、評定結果の信頼性の検討のために多くの研究に適用されている (Bulus, Hinofotis & Bailey, 1982; Brown & Bailey, 1984; Weeren & Theunissen, 1987; 山森, 2002; Shoonen, 2005; 山西, 2005a, 2005b)。

一方、多変量一般化可能性理論の外国語教育以外

への応用例としては、Webb and Shavelson (1981), Brennan, Gao and Colton (1995), 入江・鷺坂・舩田・二村 (2005) など多くが挙げられるが、外国語教育に応用された例はまだ数少ない。Webb and Shavelson (1981) は、人材トレーニング及びカウンセリングを行う目的で、雇用者の論理・数学・言語能力を測定するために適切な評定者数、回数を調査した。Brennan, et al. (1995) では、アメリカの大学でリスニングとライティングのテストを実施するにあたり、適切なタスクと評定者の数を調査した。入江他 (2005) では、上司、同僚、部下の評定が、評価の観点によって異なるかについて調査した。このように外国語教育以外への研究事例は多数見られる。

一方、外国語教育への応用例としては、山森 (2003) が挙げられる。この研究では、英語の観点別評価のうち、コミュニケーションへの関心・意欲・態度に関する評価の研究を行っている。中学1年生の2学期に92名の中学生を対象に、3観点9項目 (関心 (3), 意欲 (4), 態度 (2)) に対して2人の評定者で評価を行った。多変量一般化可能性係数が.80以上になったが、下位観点によっては係数が低いため、項目数の改善を行い、3学期の評価では、3観点10項目 (関心 (4), 意欲 (4), 態度 (2)) で実施している。

このように多変量一般化可能性理論の言語教育分野への応用が行われ始めている。また、英作文評価の分析に一般化可能性理論が用いられることも増えてきた (Brown & Bailey, 1984; Shoonen, 2005; 山西, 2005a, 2005b)。しかし、多変量一般化可能性理論を用いた英作文評価の分析はほとんど見られない。多変量一般化可能性理論は英作文の分析的評価にあたり、各観点の下位項目についての検討ができるなどの利点が多い。本研究では、多変量一般化可能性理論を用いた評価表の検討を行う。

3 | 本研究

3.1 目的

本研究では、「論説文」という特定のジャンルに焦点を当てた評価表を作成し、信頼性を調べ、どの程度の項目数と評定者数が適切であるかを検討する。また、評価表に対する評定者の反応を調査し、より

信頼性の高い評価表に必要な事項を検討する。

3.2 リサーチ・クエスチョン

1. 一般化可能性研究においては、対象者、評定者、項目と、それぞれの交互作用の分散成分と共分散成分を算出する。
2. 決定研究においては、一般化可能性研究で求められた分散成分をもとに、多変量一般化可能性係数、及び多変量信頼度指数の値を求め、40枚程度の英作文の評価に必要とされる評定者数、項目数の示唆を得る。また、各変動要因の多変量一般化可能性係数、及び多変量信頼度指数の解釈を行う。
3. 評価表に対する評定者の反応を探る。

3.3 研究方法

3.3.1 実験協力者

参加者は、都内の大学の英文学科 (32名) もしくは国際関係学科 (9名) に所属する大学1年生4人、2年生26人、3年生9人、4年生2人の計41名である。

評定者は、修士以上の学位を有する3人の英語を母語とする英語教師 (高校1名、大学2名) である。2人が20年近い教師経験を有し、1人が5、6年の教師経験を有している。

3.3.2 実施内容

3.3.2.1 参加者への実施内容

・英語テスト

参加者の英語のレベルを測定するために CASEC を用いた。CASEC とは、(株) 教育測定研究所により開発されたコンピュータを使った英語コミュニケーション能力テストである。語彙の知識、表現の知識・用法、リスニング、ディクテーションの4セクションから成立している。1セクション250点であり、1000点満点である (教育測定研究所, 2006)。テスト時間が40分程度と短時間で受験可能であり、英検、TOEIC、TOEFL と比較した目安のスコアが出るため採用した。

・課題

本研究の理論的背景であるジャンル研究においては、ある特定の文脈の中で、何を、だれに対して、どのように伝えるかを重視している。このため、参加者の興味・関心に沿った妥当性の高いタスクを選ぶこ

とが重要である。本研究の参加者は、英文学科、国際関係学科の学生であり、教職課程を取っている学生も多いことから、タスクを以下のように設定した。

An international newspaper plans to feature issues about English education in Japan. This newspaper is collecting opinions from Japanese readers. You are going to write about the following theme:

“In Korea, children are now being taught English in primary school. Japan is considering doing this, too. Do you think that English should be made part of the curriculum for all primary schoolchildren in Japan, or not?”

時間は45分とし、辞書使用可とした。

・アンケート

海外渡航経験及び英作文中の辞書使用についてのアンケートを行った。その結果、海外渡航経験者（観光、語学研修、留学、ボランティア、在住など）は38人、海外渡航経験のない者は3人であった。

辞書使用のべ人数については、英和辞典使用者が44人、和英辞典使用者が32人、英英辞典使用者が6人、英語類語辞典6人、英語活用辞典、連語辞典、日本語辞典がそれぞれ1名ずつであった。

3.3.2.2 評定者への実施内容

・評価尺度

ESL composition profile (Jacobs, et al., 1981) をもとに「論説文」(Exposition) と呼ばれるジャンルに焦点を当てた評価表を作成した。大久保 (2006) では、スペルや句読点を中心とするメカニクスと全体の英作文の評価との相関が低かった。このため、観点からはずし、内容、構成、語彙、言語使用の4観点を評価尺度とした。各観点到それぞれ、3～4の下位項目を設置した(資料1)。

この評価表における、「論説文」ジャンルの特徴としては以下の4点が挙げられる。第1に、Content 中の、Position では、ある立場に対して、賛成か反対かを選択することを求めている。第2点、第3点としては、Organisation の Structure of Exposition と、Structure of Argument の2点が挙げられる。Structure of Exposition は Statement of Position ^

Arguments ^ Reinforcement of Position Statement と段階付けられている (Board of Studies NSW, 1998)。Statement of Position では、書く内容のレビューを行う。Arguments では、議論全体を展開させる。Reinforcement of Position Statement はこれまでの議論をもとにして、再度主張や立場を述べる。Structure of Argument は Argument 内の構造を指す。1つの Argument は Point と Elaboration から成立している。Point はその Argument 内での主要な点を示し、Elaboration は、証拠に基づいた具体的な議論内容を示している。最後に、Vocabulary では The Use of Evaluative Language が挙げられる。論説文では書き手が賛成もしくは反対の選択することが重要であるため、important, significant のような特定の立場を表す語を適切に使用することが求められている。その他の記述については、ESL composition profile をもとに、それぞれの観点到特徴的な項目を下位項目とした。

評定は各項目に対して1, 2, 3の3件法で行った。重み付けに関しては、内容に配点を多くし、語彙に配点を少なくするなど、観点的な重要度に応じて重み付けを行うことも可能である。しかし、Hamp-Lyons (1991) は、積極的な理由がない限り、重み付けを用いた採点は避けるべきであると提言している。このため、本研究においては、重み付けを行わなかった。そして、評価表の点数に関しては、Weigle (2002) は、得点は異なった能力の書き手を区別する必要があるが、評定者が信頼して区別できる数には限界があると述べている。

一方、Hughes (2003) は全体のスコアが高い方が、信頼度が高いとしている。また、分散分析をもとにする一般化可能性理論を使用する場合、得点のばらつきが小さくなる採点方法は望ましくないとする山森 (2003) を踏まえ、山西 (2005a, 2005b) では、Jacobs, et al. (1981) の ESL composition profile の5観点を各10点満点で採点している。しかし、今回は15項目と項目数が多く、5件法、10件法で行った際、それぞれ点数の判断基準を細かく設定することは、評定者への負担になると考え、3件法による得点方法とした。今回のこの評価表は全体で15項目あるため、45点満点となる。

・評定者フィードバック

自由記述方式で、採点中に困難に感じた点、評価表改善のため2点に関してフィードバックを求めた。

3.3.3 データ収集

本データは2005年11月から12月にかけて集められた。参加者はボランティアとして本研究に参加した。参加者は英作文を一定時間内に書いた。その後、アンケートに答え、CASECの説明を受けた。CASEC受験は自宅または大学内の計算センターにて行われた。

集められた英作文はフィードバックシート、評価表、指示文書とともに、2006年1月に評定者に送られ、評定が行われた。評定者への指示は文書にて行われた。

3.3.4 データ分析方法

記述統計の分析には、SPSS Version 14を用いた。多変量一般化可能性理論においては、分散成分及び共分散成分の推定が必要である。また、決定研究の際には、項目数や評定者数を変化させ、シミュレーションする必要がある。中村・豊田（2002）では、多変量一般化可能性係数を確認的因子分析法に基づいて行う方法を提唱しているが、本研究では、mGENOVA（Brennan, 2001b）を用いた（資料2）。

4 結果

4.1 記述統計

本研究の参加者の英語熟達度は表1のように示される。TOEIC及びTOEFLの推定値は、CASECによって提供された。

■ 表1：参加者の英語熟達度（N = 41）

	最小値	最大値	平均値	標準偏差
語彙知識	123	204	168.29	20.18
表現の知識	128	207	169.71	19.82
リスニング	127	224	183.93	25.27
ディクテーション	120	215	169.22	20.79
CASEC 合計	581	807	691.15	59.55
TOEIC (推定値)	570	845	707.32	74.24
TOEFL (推定値)	457	546	500.17	23.48

次に、英作文の記述統計を示す。点数は3名の評定者の平均点とした（表2）。

■ 表2：英作文の結果（N = 41）

	最小値	最大値	平均値	標準偏差
Audience	2.00	3.00	2.74	.27
Example	1.33	3.00	2.29	.47
Task	1.67	3.00	2.30	.40
Position	2.00	3.00	2.81	.31
Exposition	1.33	3.00	2.27	.47
Argumentation	1.67	3.00	2.62	.39
Reference	1.67	3.00	2.57	.38
Conjunction	1.33	3.00	2.31	.54
Lexical	1.67	3.00	2.21	.39
Collocation	1.33	3.00	2.22	.36
Evaluative	1.33	3.00	2.44	.46
Word form	2.00	3.00	2.43	.30
Verb	1.67	3.00	2.68	.40
Noun	1.67	3.00	2.45	.38
Word order	1.67	3.00	2.63	.37
合計点	29.00	43.33	36.94	3.52

4.2 多変量一般化可能性理論を用いた分析

4.2.1 一般化可能性研究の結果

英作文に対する推定された分散成分及び共分散成分を、mGENOVAを用いて算出した（図4）。pは参加者、iは項目、hは評定者、piは参加者と項目の交互作用、phは参加者と評定者の交互作用、ihは項目と評定者の交互作用、pihは参加者、項目、評定者の交互作用を表している。分散成分及び共分散成分行列の見方の例を図3にて示している。なお、この見方は図5の決定研究の見方と共通である。

▼ 図3：分散成分及び共分散成分行列の見方（例）

	内容	構成	語彙	言語使用
内容	1			
構成	5	2		
語彙	6	7	3	
言語使用	8	9	10	4

図3において、1から4の位置にある値はそれぞれ、内容、構成、語彙、言語使用の分散成分を示している。5から10の位置にある値は、それぞれの観点（内容、構成、語彙、言語使用）が交差したところの共分散成分を示している。例えば、8は内容と言語使用の共分散成分を示している。

分散成分と共分散成分を検討した結果、参加者、

評定者、項目、参加者と項目の交互作用、参加者と評定者の交互作用の分散成分及び共分散成分の値はそれほど大きい値を示さなかった。一方、 $\hat{\Sigma}_{ih}$ （項目と評定者の交互作用）において内容と語彙の分散成分の値が比較的大きかった。これは内容と語彙の下位項目において、評価者が互いに異なって評価する傾向があったことを示している。例えば、語彙の下位項目であるコロケーションの判断基準の解釈が評定者によって異なることを示している。参加者、項目、評定者の交互作用の分散成分は、誤差も含め残差と見なすことができ、要因が複雑であるため、通常は解釈しないことになっている（Shavelson & Webb, 1991）。

4.2.2 決定研究の結果

決定研究の結果は図5で表せる。内容4項目、構

成4項目、語彙4項目、言語使用3項目に対して、評定者3人で評価を行った際の分散成分及び共分散成分行列、ユニバーススコアの推定された分散の推定値 $\sigma^2_c(\tau)$ 、相対誤差の分散の推定値 $\sigma^2_c(\delta)$ 、絶対誤差の分散の推定値 $\sigma^2_c(\Delta)$ 、多変量一般化可能性係数 $E\rho^2$ 、多変量信頼度指数 Φ の結果は図5のとおりであった。Brennan (2001a) に従い、決定研究記述の際は、測定対象（p；参加者）以外の項目（l）、評定者（H）に関しては大文字で記述する。

4つの観点に対して等しく重み付けを行った際のユニバーススコアの推定された分散の推定値は $\sigma^2_c(\tau) = .21$ 、相対誤差の分散の推定値は $\sigma^2_c(\delta) = .11$ 、絶対誤差の分散の推定値は $\sigma^2_c(\Delta) = .12$ である。実際に3人の評定者で行った評価に対しては、多変量一般化可能性係数は、 $E\rho^2 = .81$ 、多変量信頼度指数は、 $\Phi = .76$ となり、英作文評定の際の信頼

▼ 図4：英作文に対する分散成分及び共分散成分行列

$$\begin{array}{l} \hat{\Sigma}_p \begin{pmatrix} .051 & & & \\ .069 & .090 & & \\ .039 & .037 & .022 & \\ .036 & .042 & .037 & .032 \end{pmatrix} \quad \hat{\Sigma}_{pi} \begin{pmatrix} -.003 & & & \\ & .046 & & \\ & & .052 & \\ & & & .057 \end{pmatrix} \quad \hat{\Sigma}_{pih} \begin{pmatrix} .270 & & & \\ & .302 & & \\ & & .443 & \\ & & & .182 \end{pmatrix} \\ \hat{\Sigma}_l \begin{pmatrix} .005 & & & \\ & .030 & & \\ & & -.018 & \\ & & & .028 \end{pmatrix} \quad \hat{\Sigma}_{ph} \begin{pmatrix} -.011 & & & \\ -.009 & -.007 & & \\ -.003 & .011 & -.011 & \\ -.009 & .006 & .010 & .035 \end{pmatrix} \\ \hat{\Sigma}_h \begin{pmatrix} -.030 & & & \\ .002 & -.010 & & \\ -.001 & -.005 & -.029 & \\ -.004 & -.006 & .013 & -.020 \end{pmatrix} \quad \hat{\Sigma}_{ih} \begin{pmatrix} .124 & & & \\ & .043 & & \\ & & .168 & \\ & & & .095 \end{pmatrix} \end{array}$$

▼ 図5：決定研究の結果

$$\begin{array}{l} \hat{\Sigma}_p \begin{pmatrix} .051 & & & \\ .069 & .090 & & \\ .039 & .037 & .022 & \\ .036 & .042 & .037 & .032 \end{pmatrix} \quad \hat{\Sigma}_{pl} \begin{pmatrix} .000 & & & \\ & .011 & & \\ & & .013 & \\ & & & .019 \end{pmatrix} \quad \begin{array}{l} \sigma^2_c(\tau) = .21 \\ \sigma^2_c(\delta) = .11 \\ \sigma^2_c(\Delta) = .12 \\ E\rho^2 = .81 \\ \Phi = .76 \end{array} \\ \hat{\Sigma}_l \begin{pmatrix} .001 & & & \\ & .008 & & \\ & & .000 & \\ & & & .009 \end{pmatrix} \quad \hat{\Sigma}_{lH} \begin{pmatrix} .010 & & & \\ & .004 & & \\ & & .014 & \\ & & & .010 \end{pmatrix} \\ \hat{\Sigma}_H \begin{pmatrix} .000 & & & \\ .001 & .000 & & \\ -.000 & -.002 & .000 & \\ -.001 & -.002 & .004 & .000 \end{pmatrix} \quad \hat{\Sigma}_{pIH} \begin{pmatrix} .022 & & & \\ & .025 & & \\ & & .037 & \\ & & & .020 \end{pmatrix} \end{array}$$

度として Hamp-Lyons (2003) が述べている .75 を超えている。このため、「論説文」と呼ばれるジャンルに焦点を当てて作られた15項目からなる本評価表は、3人の評定者で行った場合、十分に信頼性を備えていることが示唆された。

そして、観点ごとに用意されるべき評定者数と項目数をシミュレーションにより検討を行った。まず、評定者1名から3名、及び各観点の下位項目数が1から5項目の時の、多変量一般化可能性係数及び信頼度指数が表3、表4にて示される。

■ 表3：項目数、評定者数と多変量一般化可能性係数の変化

項目数	評定者		
	1	2	3
1	.34	.48	.56
2	.50	.64	.71
3	.59	.72	.78
4	.65	.77	.82
5	.69	.80	.85

■ 表4：項目数、評定者数と多変量信頼度指数の変化

項目数	評定者		
	1	2	3
1	.27	.40	.48
2	.42	.57	.64
3	.52	.66	.73
4	.59	.71	.78
5	.63	.76	.81

表3は相対評価の時に用いられる多変量一般化可能性係数の変化を示している。評定者2人の時は各観点到4項目以上ずつ、評定者が3人の時は各観点到3項目以上の下位項目を設置する必要があることが示唆された。また、絶対評価の際に用いる多変量信頼度指数の場合、表4で示されるように、2人の評定者の時は5項目以上、3人の評定者の時は4項目以上ずつ下位項目を設置する必要があることが示唆された。

次に、項目数15、評価者が3名の時の各観点の一般化可能性係数と信頼度指数を表5に示す。

表5が示すように、評定者3名の各観点に対する一般化可能性係数と信頼度指数は内容と構成においては高い値を示しているが、語彙と言語使用に関し

■ 表5：各観点の多変量一般化可能性係数と多変量信頼度指数

	内容	構成	語彙	言語使用
多変量一般化可能性係数	.69	.71	.31	.39
多変量信頼度指数	.60	.65	.26	.31

ては低い値を示している。

表6は語彙と言語使用において項目数を1～5項目まで、評定者数を1～3人まで変化させた時の一般化可能性係数の変化を示している。

■ 表6：語彙と言語使用における項目数、評定者数と一般化可能性係数の変化

項目数	語彙			言語使用		
	採点者			採点者		
	1	2	3	1	2	3
1	.04	.08	.10	.11	.16	.20
2	.08	.14	.18	.17	.26	.31
3	.12	.20	.25	.22	.33	.39
4	.15	.25	.31	.25	.37	.44
5	.18	.29	.36	.28	.41	.48

3人の評定者で5項目用意したとしても、語彙は.36、言語使用は.48と高い信頼性を得られないことが示唆された。

5 考察

以上の結果を踏まえリサーチ・クエスチョンに沿って考察を行う。

5.1 一般化可能性研究に関して

一般化可能性研究で導き出された分散成分、共分散成分について考察を行う。まず、p(参加者)の分散成分及び共分散成分の値が比較的低かったことは、この評価表を使って評価を行った場合、生徒の間の評価があまり違わなかったことを示唆している。これは記述統計において、英文文の点数のほとんどの項目において、最小値と最大値もしくはどちらか一方の値が、平均からそれぞれ2標準偏差内に収まっている結果と一致している。特に、内容の audience、position に関しては、最低点が2点、最高点

が3点、平均値がそれぞれ、2.73点、2.80点と、ほとんどの生徒が等しく高い水準に達した。このようにpの分散成分が低い要因として、参加者の英作文能力に差がなかった点が挙げられる。

一方、内容と語彙においてih（項目と評定者の交互作用）の分散成分の値が1.24、1.68と大きかった。このことは、この2観点の下位項目において評定者が互いに異なって評価をしたことを示している。つまり、ある特定の項目を、ある評定者が厳しく評定する一方で、別の評定者が易しくつけた可能性がある。内容、語彙の観点とも、h（評定者）、i（項目）単体の分散成分、共分散成分はそれほど高い値を示していない。つまり評価者間の評定の厳しさに差異が見られなかった。また、各観点において、項目ごとに困難度が違うことが起こらなかったことが示されている。つまりそれぞれ単体の要因ではなく、評定者と項目の交互作用が独特に影響し、内容と語彙に関して、評定者によって下位項目の解釈が異なった可能性があることを示唆している。このことは英作文評価において、観点のみならず下位項目を設置する意義を示している。つまり、評価表が内容、語彙などの観点だけだった場合、その観点到に含まれる多数の要素に対して、評価者が互いに異なった評定を行った場合、その異なり具合を十分に反映しきれない可能性がある。よって分析的評価を行う際、観点だけではなく、下位項目を設置することが望ましいと言える。

しかしながら、具体的にどの下位項目に対してどの評定者が厳しくつけたか、易しくつけたかについては多変量一般化可能性理論が提示できる情報には限界があり、多相ラッシュ測定などによる分析が必要である（Bachman, Lynch, & Mason, 1995; McNamara, 1996; Lynch & McNamara, 1998; Kozaki, 2004）。

5.2 決定研究に関して

決定研究の結果、3人の評定者で15項目の評価表を用いた結果、多変量一般化可能性係数が.81、多変量信頼度指数が.76と、共に十分信頼性の高い結果が得られた。この結果をもとにして、評定者の人数や項目数を変化させた際に、十分な信頼性を得られる可能性について検討を行った。その結果、一般化可能性係数及び信頼度指数が.75以上になるためには、評定者1名の場合は各観点の下位項目数が5

項目、全体で20項目あっても十分な信頼性を得ることができない。一方2名の評定者がいる場合は、各観点の下位項目が4項目以上あれば一般化可能性係数を、5項目以上あれば、信頼度指数を満たすことが示唆された。3名以上の評定者の場合は、十分な一般化可能性係数を満たすためには各観点3項目を必要とし、信頼度指数を満たすためには4項目以上必要となることが示唆された。一般化可能性係数を用いるか信頼度指数を用いるかに関しては、相対評価を行うか、絶対評価を行うかによって異なる。相対評価のために一般化可能性係数を用いるとすると、実用性の観点からクラス単位での採点で、3人の評定者を確保することは難しいと考えられる場合には、各観点到4項目ずつ、全体では16項目必要となる。もし項目数を減らす方が望ましい場合には、評定者数を増やす方法として、自己採点や仲間による採点（ピア採点）を取り入れることも一考である。

次に、評定者3名、項目数15の時の各観点の多変量一般化可能性係数と多変量信頼度指数が表5に示されている。内容と構成に関しては比較的高い多変量一般化可能性係数（.69, .71）と多変量信頼度指数（.60, .65）が算出されたため、内容と構成内の項目は十分な信頼性があると示唆された。これに対し、語彙と言語使用は多変量一般化可能性係数が、（.31, .39）、多変量信頼度指数が（.26, .31）とそれぞれ低い値を示している。このため、語彙と言語使用の項目は十分な信頼性を得ていない可能性を示唆している。この要因を、第1に項目と評定者の交互作用、第2に項目数、評定者数、第3に内容・構成と語彙・言語使用の違いの3点から検討する。

第1に項目と評定者の交互作用の影響である。語彙における項目と評定者の交互作用の分散成分は.168と比較的高かった。分散成分が高いことは、語彙の下位項目において評定者が互いに異なった評定を行ったことを示している。このように、評定者ごとに各下位項目の解釈が違ったため、語彙の多変量一般化可能性係数と信頼度指数が低くなった可能性があると言える。ここから、評定者数を増やすことに対して否定的な結論が導き出せる。評定者3名において項目ごとに解釈が異なることは、4人の評定者になった場合、より差が出てくる可能性があることは否定できない。言語使用に関しては、項目と評定者の交互作用の分散成分の値が低いため評定者ごとに項目の解釈が異なったとは考えにくく、他の要

因も強く影響したと考えられる。

第2に、項目数と評定者数の問題である。評定者数を増やすことは実用的ではなく、また、人数を増やすことで評定者間の解釈に差が出るとも限らないことから、評定者3名の際に項目数を变化させた時の語彙と言語使用の一般化可能性係数と信頼度指数の変化を求めた(表6)。各観点に対し、5項目ずつ下位項目を設置したとしても、十分に高い信頼性が得られない。語彙と言語使用の項目数をより増加させることで信頼性を高めることは可能である。しかし、内容・構成は評定者3名、項目数4で比較的高い信頼性を得ているため、語彙と言語使用のみ項目数が多い評価表は不自然である。よって、評定者数、項目数を共に増やすことは望ましいことではないと言える。

第3に、内容・構成と語彙・言語使用の違いである。Schoonen, Vergeer and Eiting (1997) らは、オランダの小学6年生の作文評価を言語教育に携わる評定者と携わらない評定者とで、内容・構成と言語使用(語彙、文法、イディオム、スタイル)の2観点で行った。その結果、両者とも内容に関しては高い信頼性を示す結果となったが、言語使用に関しては、言語教育に携わる評定者の方が、高い信頼性を示す結果となった。よって、内容・構成の方が一般化しやすい観点であると述べている。また、Schoonen (2005) において、言語教育に携わる評定者が作文の評価を内容・構成と言語使用の2観点で行った。一般化可能性理論による分析の結果、内容・構成に比べて言語使用の方が評定者の影響を受けやすかったと述べている。このように、内容・構成は言語使用に携わらない評定者からも、比較的高い信頼性を得ることができ、かつ、評定者の間に差が出ないことが示されている。一方、語彙・言語使用は低い信頼性と評定者間の差異が現れている。本研究も、これらの研究同様に内容・構成に比べて、語彙・言語使用の信頼性が低かったと言える。

この要因として、2つ考えられる。Schoonen, et al. (1997) では、内容・構成に比べて言語使用は適切な言語とは何かについてのメタ言語的知識が必要であるため、言語教育に携わる評定者の方が信頼性の高い結果となり、かつ評定者間に差が現れる傾向が強く出たとしている。一方、Cumming, Kantor, & Powers (2002) は ESL/EFL 教師と TOEFL 評定者のプロトコル分析を行った結果、評価表を使用して

いる際に過去の体験(教師経験、TOEFL 評定者としての経験)が反映される傾向があると述べている。その結果、評価をする際に評価者は、評価表に注意を向けるのと同時に、過去の体験を反映する傾向があるとしている。このような作文評価への評価者の経験の影響は Lumely (2002)、山西 (2005a) でも述べられている。このように、本研究においても、メタ言語的知識の差や、適切な語彙、言語使用についての評定者自身の経験から、評定者間で相違があったのではないかと考えられる。このような評定者間の相違に関しての評定者の考えを分析するため評定者フィードバックの分析を行う。

5.3 評定者フィードバックの分析

本研究の評定者は英語教師歴20年近い2人の英語母語話者と、5年近い英語教師歴を持つ英語母語話者1人の計3人である。内容・構成に対しては、3名とも、評価表の基準に対して従ったものの、評価表に対して、「生徒のより深い思考の結果を反映させること」、「創造性を評価すること」、「内容の展開により強調を置いた評価表を作成すること」、「熟練した書き手は必ずしもルールに従うわけではないこと」を指摘している。このように3人の評定者は英作文評価に対する自分自身の考えを持ちつつも、本研究で使用された評価表に従ったと考えられる。Lumely (2002) では、採点過程のプロトコル分析から、評価表の基準と評定者自身の教師経験で培われた内的基準との間の葛藤がありつつ、評定者が評価表に従った過程を示している。これは本研究で見られたように、内容・構成における評定者自身の内的基準と評価表の間で評定者がすり合わせを行っている様子と似ている状況であると言える。このように、内容・構成に関しては、評定者自身の考えとの間で齟齬(そご)をきたしたものの、評価基準に従い信頼度の高い評価になったと考えられる。

一方語彙と言語使用に関しては、collocation, reference, evaluative language などの定義や評価基準の改定を求めたものの、内容、構成に比べてフィードバックが少なかった。このように、特に評定者自身の考えと評価表の間でさほど大きなずれを見せなかったにもかかわらず、信頼度の低い結果となったのである。Cumming, et al. (2002) のプロトコル分析によると、ESL/EFL 教師は英作文の評定において、言語使用に焦点を当てる傾向が強いと同時に、

タスクの難易度や英作文のレベルに応じて、言語使用への焦点の当て方の度合いを変化させる傾向があったと述べている。このような傾向が本研究の評定者にもあったが、気づきとして起こらなかったため、フィードバックに反映されなかった可能性がある。しかし、本研究では、プロトコル分析は行っていないので、評定者の意識の向け方については今後の課題としたい。

このような内容・構成における評定者の葛藤と、語彙・言語使用に対する評定者間の意識の違いに対して指導と評価を一体化させるためには、評定者トレーニングが必要となる。なぜならば、指導内容をもとに評価表を作成したとしても、評定者間で解釈が異なれば、信頼度の低い評価となるからである。Weigle (1994, 1998) では、評定者トレーニングを通し、評定者が自ら期待する学習者への評価をある統一した基準にある程度すり合わせを行うようになったと述べている。Lumely (2002) も、教師としての内的基準を重視しつつも、ライティングのような複雑な要因から成立している対象を評価する際には、評価者間、評価者内の信頼性を高めることも重要であると述べている。そのため、ある目的のための評価を行う際（例：クラス内における論説文のための評価）、一定の統一性を確保するためには、評定者トレーニングなどを通し、ある程度統一した基準を保つことが重要であると言える。

しかし評価表を絶対視することや、教師が培った経験を無視することは避けなければならない。評定者が感じた葛藤や言語知識に対する意識の差について、フィードバックをとることにより、評価表の改善を試みるのが重要である。より柔軟性のある評価表を作成することにより、指導内容を反映し、かつ、あるレベルの学習者を測るのに適切な評価表を作成するように心がけることが重要であると言える。

6 次回への改善点

本研究における限界と改善点としては以下の2点が主に挙げられる。第1点目としては、タスクの数である。今回、参加者の負担を減らすため、1つのタスクで行った。しかし、英作文評価へのタスクの影響も否定できないため、今回は複数のタスクで行う必要があると考えられる。

第2に、採点者トレーニングである。評定者の居住地域、時間的制約があり、今回は文書による指示と電子メールによる補完的な指示を行った。Weigle (1994) で示されるように、採点基準について互いに話し合う機会を持った方が、評定者同士の基準が統一されやすい。また、評定者トレーニングの重要性が確認できたことから、次回はより充実した評定者トレーニングを行いたいと考えている。

7 まとめ

本研究においては、大学で指導されることの多い、ジャンル別の評価表、特に「論説文」に焦点を置いた評価表を作成し、その採点項目及び評定者に関する信頼性の検討を、多変量一般化可能性理論を用いて検討した。

その結果、評定者3名で15項目の評価表を用い、41名の大学生の英作文の採点を行ったところ、多変量一般化可能性係数、多変量信頼度指数共に十分な信頼性を得る結果が導き出された。しかしながら、下位項目によって、評定者が異なった評定を示す傾向が示唆された。また、各観点において、内容・構成では高い多変量一般化可能性係数及び信頼度指数を示すのに対し、語彙・言語使用は低かった。この要因としては、さまざまあるが、その1つとして、語彙・言語使用に対する評定者間の意識の差が挙げられる。

また、評定者からのフィードバック分析を通して、内容に関しては、評定者が自らの内的基準と評価表の基準をすり合わせる必要があった可能性を示唆している。また、評定者によって言語使用における焦点の当て方が異なった可能性を示している。よって、ある程度の統一性を確保するために、評定者トレーニングなどを通して、信頼性の高い評価を試みるのが重要である。

謝 辞

本研究の実施にあたっては多くの方のご指導とご協力をいただきました。ご多忙にもかかわらず、選考委員の池田央先生、津田塾大学の田近裕子先生から丁寧なご指導を頂戴しました。また、一般化可能性理論に関して、国立教育政策研究所の山森光陽氏、広島大学大学院の山西博之氏には草稿の段階で

有益な示唆をいただきました。

今回この研究に協力して下さった大学生41名と
評定者3名の方がいらっしゃらなければこの研究は

成立しませんでした。感謝の念で一杯です。

最後に、本研究を支えてくださった(財)日本英
語検定協会に心から感謝いたします。

参考文献 (*は引用文献)

- * Bachman, L.F. (2004). *Statistical analysis for language assessment*. Cambridge, UK: Cambridge University Press.
- * Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238-257.
- * Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- * Board of Studies. (1998). *K-6 English syllabus*. Sydney, NSW: Board of Studies.
- * Bolus, R.E., Hinofotis, F.B., & Bailey, K.M. (1982). An introduction to generalizability theory in second language research. *Language Learning*, 32, 245-258.
- * Brennan, R.L. (2001a). *Generalizability Theory*. New York: Springer.
- * Brennan, R.L. (2001b). *Manual for mGENOVA. Version 2.01*. Iowa city, IA: The University of Iowa.
- * Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses work keys listening and writing tests. *Educational and Psychological Measurement*, 55, 157-176.
- * Brown, H.D. (2004). *Language assessment: Principles and classroom practice*. New York: Longman.
- * Brown, J.D., & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34, 21-42.
- * Cumming, A., Kantor, R., & Powers, D.E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive Framework. *The Modern Language Journal*, 86, 67-96.
- * Grabe, W., & Kaplan, R.B. (1996). *Theory and practice of writing*. Essex, UK: Longman.
- * Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 162-189). Cambridge, UK: Cambridge University Press.
- * Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic context* (pp. 241-276). Norwood, NJ: Ablex.
- * Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of Second Language Writing* (pp. 69-87). Cambridge, UK: Cambridge University Press.
- * Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.
- * Hyland, K. (2002). *Teaching and researching writing*. Essex, UK: Pearson Education.
- * 池田央. (1994). 『現代テスト理論』. 朝倉書店.
- * 入江崇介・鷲坂由紀子・舩田博之・二村英幸. (2005). 『多面観察評価における上司・同僚・部下の評定結果の統合について～多変量一般化可能性理論による一般化可能性の検証』. 日本テスト学会第3回大会発表論文抄録集, 60-63.
- * Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F., & Hughey, J.B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- * Johns, A.M. (2003). Genre and ESL/EFL composition instruction. In B. Kroll (Ed.), *Exploring the dynamics of Second Language Writing* (pp. 195-217). Cambridge, UK: Cambridge University Press.
- * 金谷憲 (編). (2003). 『英語教育評価論』. 河源社.
- * Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21, 1-27.
- * 教育測定研究所 (2006). 『CASEC について』.
<http://casec.evidus.com/ex/01/index.html> より引用. (2006年2月5日取得).
- * Lumely, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean the raters? *Language Testing*, 19, 246-276.
- * Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- * McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.
- * 文部科学省. (2003). 『「英語が使える日本人」の育成のための行動計画』.
http://www.mext.go.jp/b_menu/houdou/15/03/03033102.pdf より引用. (2006年3月31日取得).
- * 中村健太郎・豊田秀樹. (2002). 『確認の因子分析による多変量一般化可能性係数の推定』. 第30回日本行動計量学会大会における口頭発表. 於多摩大学. 2002年9月19日.
- * 大久保奈緒. (2006). 『自由英作文評価における信頼性の検討：一般化可能性理論を用いて』. *Tsuda Inquiry*, 27, 145-162.

*Schoonen, R.(2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22, 1-30.

*Schoonen, R. Vergeer, M., & Eiting, M.(1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, 14, 2, 157-184.

*Shavelson, R.J., & Webb, N.M.(1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE.

* 山森光陽. (2002). 『一般化可能性理論を用いた観点別評価の方法論の検討』. *STEP BULLETIN*, vol.14, 62-70.

* 山森光陽. (2003). 『中学英語科の観点別学習状況の評価における関心・意欲・態度の評価の検討—多変量一般化可能性理論を用いて—』. *教育心理学研究*, 5, 195-204.

* 山森光陽. (2004). 「英会話テストの信頼性の検討—一般化可能性理論—」. 前田哲朗・山森光陽 編著. 『英語教師のための教育データ分析入門』. pp.82-89. 大修館書店.

* 山西博之. (2005a). 『一般化可能性理論を用いた高校生の自由英作文の検討』. *JALT Journal*, 27. 169-185.

* 山西博之. (2005b). 『自由英作文評価の改善：評定結果の診断的活用』. *JACET 第44回全国大会要綱*, 219-220.

*Webb, N.M., & Shavelson, R.J.(1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement*, 18, 13-22.

*Weeren, J.V., & Theunissen, T.J.J.M.(1987). Testing pronunciation: An application of generalizability theory. *Language Learning*, 37, 109-122.

*Weigle, S.C.(1994). Effect of training on raters of ESL composition. *Language Testing*, 11, 197-223.

*Wegile, S.C.(1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.

*Weigle, S.C.(2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

資 料

資料１：論説文（Exposition）のための英作文評価基準表

Aspects	Items	Score	
Content	Audience	3	On the whole, the writer makes the audience understand the content clearly.
		2	The writer does not appeal to the audience with appropriate levels of formality.
		1	The writer tends to ignore the audience.
	Examples	3	Adequate and concrete examples explain main points.
		2	Examples are written, but they are abstract.
		1	Few examples are written to support main points.
	Relation to task	3	The writer understands the task and develop his/her idea well.
		2	The writer understands the task, but he/she can not explain his/her ideas fully.
		1	The writer does not understand the purpose of task well.
	Position	3	<i>For or against</i> a particular position is clearly shown.
		2	<i>For or against</i> a particular position is mixedly shown.
		1	<i>For or against</i> a particular position is not chosen.
	Structure of Exposition	3	The structure consists of <i>statement of position</i> ^ <i>arguments</i> ^ <i>reinforcement of position statement</i> .
		2	The structure lacks one of the following stages: <i>statement of position</i> ^ <i>arguments</i> ^ <i>reinforcement of position statement</i> .
		1	There are no clear stages.
	Structure of Argument	3	In argument stages, <i>point</i> is followed by <i>elaboration</i> .
		2	In argument stages, only <i>elaboration</i> sentences are written.
		1	In argument stages, only <i>point</i> is written.

Organisation	Reference	3	Reference chains help readers to understand links between sentences.
		2	Reference chains exist, but these chains are not smoothly linked.
		1	Reference chains are ignored in the text.
	Conjunction	3	Appropriate conjunctions create logical relation.
		2	Conjunctions do not presuppose logical relations clearly.
		1	Conjunctions do not help readers to see the connection between sentences.
Vocabulary	Lexical cohesion	3	Various lexical items are used, and they are well-associated and natural in the text.
		2	Various lexical items are used, but the association between them is not natural.
		1	The same words or expression are repeatedly used.
	Collocation	3	The writer uses collocations naturally in the text.
		2	The writer uses accurate collocations, but some of them are not suitable to the context.
		1	The writer uses few accurate collocations.
	The use of evaluative language	3	The writer shows his/her personal opinion by using accurate and effective evaluative words.
		2	The writer uses evaluative words, but they do not support his/her opinions well.
		1	The writer uses few evaluative languages.
	Word form (prefix & suffix)	3	Word form is accurate.
		2	There are occasional errors of word form.
		1	There are frequent errors of word form.
Language use	Verb (tense, number, agreement)	3	There are few errors of verb agreement.
		2	There are occasional errors of verb agreement
		1	There are frequent errors of verb agreement.
	Noun (article, plural)	3	There are few errors of noun agreement.
		2	There are occasional errors of noun agreement
		1	There are frequent errors of noun agreement.
	Word order	3	Normal word order is followed except for special emphasis.
		2	Word order is effective, but it has minor problems.
		1	Word order is not appropriately arranged.

ID:

Score

資料2：図4，図5，表5の結果を出すために用いられた mGENOVA の入力データ

```

GSTUDY      p x i x h Designs with Covariance Component Design = p x h
OPTIONS     NREC 5 "*"out"
MULT        4 Content  Organisation  Vocabulary  Languageuse
EFFECT      * p41 41 41 41
EFFECT      i  4  4  4  3
EFFECT      #h 3  3  3  3
FORMAT      0 0
PROCESS
  実際のデータ
DSTUDY      p x l x H Designs with Covariance Component Design = P x H
WWTS        .25 .25 .25 .25
DEFFECT     $ p41 41 41 41
DEFFECT     l  4  4  4  3
DEFFECT     #H 3  3  3  3
ENDDSTUDY

```