一般化可能性理論を用いた高校生の自由英作文評価の検討

Using Generalizability Theory in the Evaluation of L2 Writing

山西　博之
広島大学大学院

This paper aims to investigate the characteristics of the evaluation of L2 writing—particularly free English compositions by Japanese high school students—using Generalizability Theory (G theory). Although usually considered to be a difficult topic to examine, the evaluation of free compositions can be thoroughly investigated by using G theory. It enables researchers to provide sufficient information regarding the main effects and the interactions of complicated factors within an evaluation by examining its measurement errors.

I focused on two factors (more specifically, *facets*) in order to obtain the data on the evaluation of free compositions. These facets were: (a) the raters—10 high school teachers (expert raters) teaching English at a national high school and two public high schools, and six university students (novice raters) studying English language education at a national university; and (b) the rating scales, which were Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) *ESL Composition Profile,* and a modified version of *Kantenbetsu Hyoka* of the National Institute for Educational Policy Research (2002). Using these scales, the raters (expert and novice raters) evaluated free compositions written by 20 high school students studying at a national high school in the *Chugoku* region of Japan. The type of G theory design used in this paper is termed a *two-facet crossed design* (all the raters evaluate all the compositions using all the items of the rating scales).

JALT Journal, Vol. 27, No. 2, November, 2005

Studies using G theory are usually comprised of two substudies: a *Generalizability Study* (*G study*) and a *Decision Study* (*D study*). A G study investigates the manner in which the facets and their interactions (termed as *sources of variance*) affected the evaluation results by estimating the magnitude of *variance components*. A D study investigates the degree of reliability of the evaluation by examining *generalizability coefficients*, which correspond to classical test theory's reliability coefficients, using simulations that vary the number of raters or items of the rating scales. The G study in this paper dealt with seven sources of variance—persons (*p*), raters (*r*), rating scale items (*i*), and their interactions (*p* x *r*, *p* x *i*, *r* x *i*, and *p* x *r* x *i*). The D study in this paper particularly focused on varying the number of raters for simulations.

Several observations resulting from both the G study and the D study were as follows: (a) there was a halo effect tendency in the evaluations by the expert raters because the estimated variance components of the interactions of the sources of variance *p* x *r* and *r* x *i* were large; (b) the novice raters' rating experience was insufficient to perform reliable evaluations because the generalizability coefficients of both of the rating scales were low, while the estimated variance component of the interaction of the sources of variance *p* x *r* x *i*, which is regarded as unmeasured error, was large; and (c) the ESL Composition Profile was a more reliable rating scale than the *Kantenbetsu Hyoka* as shown by the D study simulation results.

This paper presents several pedagogical implications based on the results with reference to improvement in the evaluation of free compositions. In particular, I have presented possible methods of diagnostically utilizing the results of G theory to develop and modify the rating scales, and to train the raters.

本研究では，高校生の自由英作文評価に対して一般化可能性理論を用いた検討を行った。一般化可能性理論は，（1）測定に伴う変動要因とその測定誤差の大きさ（分散成分）を推定するための「一般化可能性研究」（G研究）と，（2）分散成分の推定値から求めた一般化可能性係数をもとに評価の改善を行うための「決定研究」（D研究）からなる。本研究では，20名の高校生が書いた自由英作文を高等学校の英語科の教員10名と教員を目指す大学生6名が2種類の分析的評価尺度（ESL Composition Profileと「観点別評価」）によって評価した結果を，一般化可能性理論を用いることで検討した。自由英作文評価に関する生徒，評定者，評価項目といった変動要因とそれらの交互作用がG研究によって，評定者の人数と評価の信頼性の関係がD研究によって，それぞれ詳細に検討された。そして，評価の改善という観点から教育的示唆が示された。

高等学校では2003年度から新しい学習指導要領が施行され，外国語（英語）においては「実践的コミュニケーション能力」を育成するために，従来よりも実践の場で使用できる英語を指導することが目指されることとなった。そのような状況では「ライティング」をはじめとする科目の指導において，実践的な使用場面や目的に応じて英語で自分の意見を主張したり物事を説明したり（文部省, 1999）といった自由英作文（free compositions）を生徒が書き，それを教員が評価する機会は，これまでよりも多くなると考え

られる。しかしながら，そのような実践的な使用場面に応じた自由英作文の評価は，例えば和文英訳よりも評価に主観の入る余地が大きいと考えられ，高い信頼性を得ることが困難であると言える。そのため，自由英作文をどのように評価すれば信頼性の高い評価になるのかという問題は，極めて重要なトピックであるものの未だ十分に議論されていると言えないのも事実である。とりわけ，どの評価尺度を用いた場合にはどの程度の信頼性が得られ，それが十分でなかった場合には，何人の評定者が評価すれば十分に信頼性の高い評価になるのか，といった情報を得ることは，高等学校や大学での入学試験，または学内で行う定期考査や実力テストなどで自由英作文を評価する場合には重要であると考えられる。

そこで本研究では，評定者が複数の評価尺度を用いて高校生の自由英作文を評価した結果を分析することで，上記の問題に対する検討を行っていくものとする。その際，先行研究で行われてきた信頼性係数や相関係数での分析よりも，より詳細な検討が可能になる一般化可能性理論（Generalizability Theory）を用いた分析を行っていく。さらに，日常的に高校生の自由英作文評価を行っている高等学校の教員と，教員を目指すもののまだ評価に不慣れな大学生による評価特徴の差異も，一般化可能性理論を用いて検討していくものとする。そして，それらを総合した考察から，高校生の自由英作文評価に対する教育的示唆を得ることを本研究の目的とする。[1]

自由英作文評価の研究

日本人の高校生が書いた自由英作文の評価における信頼性の問題を取り扱った研究として，工藤・根岸（2002）の研究や山西（2004）の研究がある。

工藤・根岸（2002）の研究では，同一の自由英作文に対して，印象的採点方法（総合的かつ主観的に得点を付ける採点方法），全体的採点方法（総合的かつ採点基準に基づいて得点を付ける採点方法），分析的採点方法（複数の観点に基づいて分析的に得点を付ける採点方法）の3種類の採点方法で評価を行うことで，採点方法の違いから生じる採点結果の信頼性の差異が検討された。工藤・根岸は，14名の評定者が36名の高校生の自由英作文を上記3種類の採点方法（評価尺度）によって評価した結果を用いて，それぞれの採点方法ごとに2名～14名の評定者の全ての組み合わせの採点者間信頼性係数をスピアマン・ブラウンの公式によって算出した。その結果，3種類の採点方法の中では，分析的採点方法で評価を行うESL Composition Profile（Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981）を使用した場合に最も高い採点者間信頼性係数を得ることができることが示された。

また，山西（2004）の研究では，分析的評価尺度としてJacobs et al.（1981）のESL Composition Profileと国立教育政策研究所（2002）を参考にして作成された「観点別評価」用の尺度の2種類が用いられ，10名の教員と6名の大学生によって，20名の高校生が書いた40編の自由英作文が評価された。その評価結果と，作文用の総合的評価尺度（工藤・根岸の用語では全体的採点方法）である「客観的総合評価」と「主観的総合評価」（石田・森, 1985）を用いた評価結果と印象的評価尺度を用いた評価結果が比較された。比較には評価尺度内・評価尺度間の相関係数（ピアソンの積率相関係数）と評価尺度内の信頼性係数（クロンバックのα係数）が用いられ，教員と大学生の評価特徴の差異が検討され

た。その結果，大学生による評価よりも教員による評価の方が評価尺度内・評価尺度間の一貫性が高いことや，大学生の評価においては印象的評価尺度と「観点別評価」尺度の相関係数（印象的評価との併存的妥当性）が他の尺度によるものよりも若干低いことなどが示された。

これらの研究のうち，工藤・根岸（2002）の研究では採点方法（評価尺度）の違いから生じる採点者間信頼性係数の差異が，また山西（2004）の研究では評価尺度内・評価尺度間の相関係数と信頼性係数から明らかになった教員と大学生の評価特徴の差異が，それぞれ主な検討の対象となり教育的な示唆が与えられている。これらの研究のように信頼性係数や相関係数を用いた分析は，その研究の目的とすることに適っていれば有用な方法であると言える。しかしながら，このような分析からは，本研究の目的である，具体的にどのように評価を改善できるのかという情報を得ることは困難である。その理由は，自由英作文の評価結果には，「評定者」という要因のみではなく，「生徒」や「評価項目」といった要因が複雑に絡み合っているため，評価の改善を詳細に検討するためには，それらの要因ごとの影響を見極める必要があるためである。相関係数や信頼性係数を検討する方法は古典的テスト理論と呼ばれるが，古典的テスト理論ではそれらの要因を見極め，同時かつ詳細に検討することは困難である。つまり，本研究で目的とする評価の改善という作業には，それに適った方法を用いることが重要であると言える。

一般化可能性理論を用いた評価研究

池田（1994）は，古典的テスト理論では解決困難な，記述式テストの評価に伴う問題点を，一般化可能性理論を用いることで克服できるとしている。[2]

一般化可能性理論とは，一般化可能性研究（Generalizability Study; 以下，G研究）と決定研究（Decision Study; 以下，D研究）からなる。G研究とは，自由英作文評価などの測定において生じる測定誤差に着目し，その測定誤差の原因である測定に伴う変動要因の成分とそのばらつきの大きさ（分散成分）を推定することで，各変動要因やそれらの交互作用が評価に与える影響を検討するための研究である。一方，D研究とはG研究で得られた分散成分の推定値を用いて，通常の信頼性係数（α係数）に相当する一般化可能性係数（generalizability coefficient）を算出し，どの程度の評価項目数や評定者の人数であれば十分な一般化可能性係数を得られるかのシミュレーションを行い，評価を改善するための研究である（Bachman, 1997; Brennan, 1992; 池田, 1994; Shavelson & Webb, 1991; 山森, 2002, 2004）。このように評価の改善を念頭に置いていることが，一般化可能性理論の持つ古典的テスト理論に対する大きな利点であると言える。

一般化可能性理論を用いた研究で，言語教育の評価に関するものは，海外ではESL環境でのスピーキング技能の測定に関する研究（Lynch & McNamara, 1998）などが見られるが，国内では多くは行われてきていない。そのような中で，英語教育学の分野での先駆的な研究として，山森（2002）の研究がある。山森は，中学校の授業での「観点別学習状況」（コミュニケーションへの関心・意欲・態度，表現の能力，理解の能力，言語や文化についての知識理解，の4観点）の評価が，どのように行われているのかをG研究によって，また，どのように改善されるのかをD研究によって，1年間にわたって調査・検討した。山森の研究は，評価の困難な観点別学習状況の評価を実際に行っていく中で，十

分に信頼性の高い評価を行うための評定者数や評価項目数を示し，それらの改善を行いながら調査を継続させたという点で，極めて有用な情報の提供を行ったものと言える。

作文の評価の研究においても，一般化可能性理論を使用した研究はこれまで国内ではほとんど行われてきておらず，国語科において日本語での児童の作文評価を分散成分の推定値と一般化可能性係数によって検討した梶井（2001）の研究が見られる程度である。梶井の研究では，児童の作文の評価を行っていく中で，作文に対する好意度が高い教員と低い教員の評価特徴の差異と，総合的評価尺度である石田・森（1985）の「客観的総合評価」と「主観的総合評価」による評価結果と学習指導要領から作成された分析的評価尺度による評価結果の差異がそれぞれ検討され，多くの有益な示唆が与えられた。

さらに領域を限定し自由英作文の評価の研究に目を転じると，現在まで分散成分の推定を行うことで自由英作文評価に関連する変動要因を検討するといった研究もほとんど行われてきてはいない。しかしながら，このような困難な領域にこそ一般化可能性理論を用いた研究は有用であると言え，比較的簡便に統計パッケージやソフトウェアで分散成分の推定や一般化可能性係数の算出ができるようになった現状では，山森（2002）の「評価の検討こそ，ITの発展の恩恵を受けるべき分野である」（p. 69）という主張を受けとめ，知見の積み重ねを行っていくことは極めて重要であると言える。

方法

検討内容

そこで本研究では，一般化可能性理論を使用して，高校生が書いた自由英作文に対する評価を検討する。その際，分析のためのデータとして，高校生の自由英作文の評価結果としては比較的新しいものである，山西（2004）の調査で得られたデータを用いる。[3] 具体的には，山西のデータのうち本研究の目的に応じた部分を一般化可能性理論を用いて再分析することで，以下の2点を検討する。

1) 自由英作文評価における生徒，評定者，評価項目という変動要因の主効果とそれらの交互作用の分散成分の推定値を求め，特に評定者の評価経験の違いによる評価特徴の差異に注目した検討を行うこと（G研究）。
2) 分散成分の推定値から一般化可能性係数を算出し，評定者の評価経験の違いと評価尺度の違いに注目したシミュレーションを行うことで，評定者の人数と得られる信頼性の関係に対する検討を行うこと（D研究）。

評価尺度

本研究で検討する評価尺度は2種類あり，1つめはESL Composition Profileである。ESL Composition Profileは，Jacobs et al.（1981）が開発した分析的評価尺度で，content, organization, vocabulary, language use, mechanicsの5項目からなる（Jacobs et al., 1981, p. 30）。この評価尺度は，英作文の分析的評価尺度としては代表的なもので，現在に至るまで多くの研究で使用されてきている。そして，2つめの評価尺度は，国立教育政策研究所（2002）を基に作成された「観点別評価」用の尺度である。この評価尺度は，山西（2004）で用いられた

もので，「言語活動への取組」，「コミュニケーションの継続」，「正確な表現の能力」，「適切な表現の能力」，「言語についての知識」，「文化についての理解」の6項目からなる分析的評価尺度である（山西, 2004, p. 195）。[4] これらの評価尺度はいずれも分析的評価尺度であるが，本研究で分析的評価尺度を検討した理由は以下の通りである。

まず，分析的評価尺度（分析的採点方法）は，工藤・根岸（2002）が指摘するように，その信頼性（採点者間信頼性）が他の評価尺度（印象的評価尺度や総合的評価尺度）に比べて高いことから評定者（採点者）間のぶれの少ない評価が行える一方，評価項目が多いため評価に労力が要求され，実用性（practicality）の面で問題が生じる可能性がある。そのため，特定の分析的評価尺度を用いた場合に，どの程度の評定者数であれば十分に（または，ある程度）信頼性の高い評価がなされるのか，という情報を得ることは，評価における実用性と信頼性のバランスを見極めるために有用であると考えられる。さらに，本研究で使用した2種類の分析的評価尺度においては，世界的に多くの先行研究で用いられているESL Composition Profileを基準として，2003年度以降の高等学校で導入された「観点別評価」に基づいた分析的評価尺度の評価結果を比較・検討することも可能であり，そのような検討は評価尺度の改善のために有用であると考えられる。

評定者と評価対象

自由英作文の評定者は，中国地方の高等学校英語科の教員10名（1校の国立高等学校教員8名と2校の公立高等学校教員各1名；教員歴8〜32年；平均教員歴17.3年；男性7名，女性3名）と中国地方の国立大学で英語教育学を専攻し教員を目指す大学生6名（学部4年生4名と修士課程の大学院生2名；男性1名，女性5名）である。本研究では，便宜的に，教育現場で日常的に高校生の自由英作文評価を行っており教員歴も比較的長い（8年〜）中堅以上の教員を「評価の熟達者」として捉える一方，教育現場での評価経験が教育実習以外にない大学生を「評価の初心者」として捉え，両者の評価経験の違いによる評価特徴の差異を比較・検討するものとする。

また，評価の対象者は，中国地方の国立高等学校の普通科に在籍する生徒20名（高校1年生10名，男女5名ずつ；高校2年生10名，男女5名ずつ）である。彼らは週2コマの「ライティング」の授業中に自由英作文の指導を受けており，評価対象となった作文は，彼らが授業中に書いた自由英作文課題のうちの2種類（課題A，課題B）である。

評定作業のために，2種類の自由英作文（課題の詳細はAppendix A, B参照）はワープロでタイプし直された。タイプされた自由英作文は，A4用紙の横見開き単位（左側に課題A，右側に課題B）で評価項目と併せて印刷され，評価シートとされた。そして，生徒20名分の評価シートは，上述の2種類の分析的評価尺度ごとに1綴りにされた。なお，評価される自由英作文は，生徒の学年・性別などの情報が伏せられた上で，全ての評定者においてランダムオーダーにされ評定者に手渡された。

分析の手順

本研究では，高校生によって書かれた2種類の自由英作文を，評定者が2種類の分析的評価尺度（ESL Composition Profile = 5項目と「観点別評価」= 6項目）を用いて評価した。実際の教育現場では，3段階（A〜C）や5段階（1〜5）のスケールが用いられることが多いと考えられるが，分散分析の手法を応用する一般化可能性理論においては，評価のばらつきが小さくなるスケールを用いることは望ましくないという指摘（山森，2003）があるため，本研究では評価特徴を明確に捉えることを目指し，1〜10点で評価したデータを分析対象とした。なお，本研究では課題種類の違いは検討対象とはせず，2種類の課題の得点（各10点）を合計して，20点満点とした。

本研究でのG研究は，「評定者」と「評価尺度」という2つの相（測定を行うための条件）を設定することで生徒の自由英作文を評価（測定）し，その際に各変動要因（生徒，評定者，評価項目）が完全に組み合わされる（クロスされる）という，「2相完全クロス計画」（全評定者が全評価項目を使用して全生徒の自由英作文を採点する計画）に基づいたものであった。分析ではまず，2種類の課題（課題A，課題B）における生徒の作文に対する評価得点の合計点（20点満点）を従属変数として，評定者（教員，大学生）と評価尺度（ESL Composition Profile，「観点別評価」）の組み合わせごとに分散成分を推定した。特に評価経験の違いによる評価特徴の差異に注目するために，評定者である教員と大学生の比較を行いながら分散成分の推定値の検討を行った。なお，本研究のG研究ではSPSS 11.5J Advanced Modelsに組み込まれているVARCOMPを使用して分散成分の推定を行った。[5]

また，D研究では，分散成分の推定値から一般化可能性係数を算出し，その結果を用いたシミュレーションを行った。その際，2種類の分析的評価尺度ごとに，どれだけの人数の評定者がいればどの程度の信頼性（一般化可能性係数）を得られるのかを検討し，評価尺度の項目数に関しての検討は行わなかった。その理由は，本研究で検討する各分析的評価尺度は，それぞれの項目数（5項目ないし6項目）で1つの技能（作文技能）を測定するためにデザインされたものであるため，項目数を増減させて一般化可能性係数のシミュレーションを行うことは現実的であるとは思われないためである。本研究での一般化可能性係数の算出は，Shavelson and Webb（1991）や山森（2002, 2004）で用いられた2相完全クロス計画用の計算式である式（1）によって行った。なお，式（1）の中のGは一般化可能性係数であり，p，pi，priなどは表1と表2の中の変動要因に対応している。また，Nrは評定者の人数をあらわしており，Niは評価項目数をあらわしている。具体的には，表中のp，pi，priといった変動要因の分散成分の推定値を式（1）の該当箇所に代入していくことで，一般化可能性係数が算出される。

(1)
$$G = \frac{p}{p + \frac{pr}{Nr} + \frac{pi}{Ni} + \frac{pri}{Nr\,Ni}}$$

結果

G研究の結果

G研究の結果として得られた分散成分の推定値は，評価尺度ごとに示し，特に評定者の変動要因の主効果と交互作用を検討することで，評定者である教員と大学生の比較を行った（評価尺度ごとの評価結果の記述統計量はAppendix C, D参照）。ただし，表1と表2に示された分散成分の推定値は，各評定者（教員，大学生）の間で値同士を直接比較することはできない。そこで，比較を可能にするために，全変動要因の分散成分の推定値の合計に対する各変動要因の分散成分の推定値の割合を百分率で求め，各表のカッコの中に示した。なお，各表に示された「生徒×評定者×項目」の交互作用の分散成分の推定値は，残差に相当すると考えられ，各変動要因（生徒，評定者，評価項目）では説明できない要因である（池田, 1994; 梶井, 2001; Shavelson & Webb, 1991）。

まず，表1にはESL Composition Profileの分散成分の推定値とその割合を示した。この表1の変動要因「評定者」の分散成分の推定値の割合（教員61.10%, 大学生60.66%）が全変動要因中の大部分を占めることから，教員・大学生ともに評定者による評価のばらつきは非常に大きかったことが示された。また，「生徒×評定者」の交互作用の分散成分の推定値の割合（教員12.58%，大学生6.27%）から，教員の評価には各生徒に与えた評定値に比較的大きなばらつきがあり，大学生の評価においても若干のばらつきがあったことが示された。そして，「評定者×項目」の交互作用の分散成分の推定値の割合（教員5.76%，大学生0.20%）から，教員の評価には評価項目の捉え方に若干の違いがあった一方，大学生の評価には評価項目の捉え方の違いはほとんどなかったことが示された。

表1. 分散成分の推定値と一般化可能性係数（ESL Composition Profile）

	分散成分推定値	
変動要因	教員(n = 10)	大学生(n = 6)
生徒(p)	1.01 (9.70%)	0.43 (4.34%)
評定者(r)	6.36 (61.10%)	6.00 (60.66%)
項目(i)	0.02 (0.19%)	0.23 (2.33%)
生徒×評定者(pr)	1.31 (12.58%)	0.62 (6.27%)
生徒×項目(pi)	0.10 (0.96%)	0.51 (5.16%)
評定者×項目(ri)	0.60 (5.76%)	0.02 (0.20%)
生徒×評定者×項目(pri)	1.01 (9.70%)	2.08 (21.93%)
一般化可能性係数	0.86	0.61

Note. カッコ外は分散成分の推定値，カッコ内は百分率にした割合.
一般化可能性係数は当該評定者の人数で算出.

次に，表2には「観点別評価」の分散成分の推定値とその割合を示した。なお，大学生の評価において，変動要因の「項目」の分散成分の推定値は負であったものの値が小さかった（σ^2 = -0.05）ため，Brennan（1992）の方法を適用して値を0に修正した。ただし，分散成分の推定値の割合の計算には負の値をそのまま使用した。[6] この表2の変動要因「評定者」の分散成分の推定値の割合（教員39.69%，大学生35.37%）が全変動要因中で最も大きかったことから，教員・大学生ともに評定者による評価のばらつきは大きかったことが示された。また，「生徒×評定者」の交互作用の分散成分の推定値の割合（教員16.02%，大学生13.58%）から，教員・大学生ともに各生徒に対して与えた評定値に比較的大きなばらつきがあったことが示された。そして，「評定者×項目」の交互作用の分散成分の推定値の割合（教員16.02%，大学生0.57%）から，教員の評価には評価項目の捉え方に比較的大きな違いがあった一方，大学生の評価には評価項目の捉え方の違いはほとんどなかったことが示された。

表2. 分散成分の推定値と一般化可能性係数（「観点別評価」）

	分散成分推定値	
変動要因	教員(n = 10)	大学生(n = 6)
生徒(p)	0.77 (9.01%)	0.45 (8.60%)
評定者(r)	3.37 (39.69%)	1.85 (35.37%)
項目(i)	0.77 (9.01%)	0.00 (-0.96%)
生徒×評定者(pr)	1.36 (16.02%)	0.71 (13.58%)
生徒×項目(pi)	0.05 (0.59%)	0.82 (15.68%)
評定者×項目(ri)	1.36 (16.02%)	0.03 (0.57%)
生徒×評定者×項目(pri)	0.81 (9.54%)	1.42 (26.96%)
一般化可能性係数	0.83	0.61

Note. カッコ外は分散成分の推定値，カッコ内は百分率にした割合.
一般化可能性係数は当該評定者の人数で算出.
下線は負の値（-0.05）をBrennan（1992）の方法で0に修正.

D研究の結果

ここでは，G研究で得られた分散成分の推定値を使用して一般化可能性係数を算出し，評定者数と信頼性の関係のシミュレーションを行った。

本研究のD研究ではまず，G研究を行った際の評定者数（教員= 10名，大学生= 6名）で得られた一般化可能性係数を，表1と表2の下段に示した。この結果からESL Composition Profile，「観点別評価」ともに教員の方が大学生よりも信頼性の高い，つまり一貫性の高い評価を行っていたことが示された。一般化可能性係数の解釈は，古典的テスト理論の信頼性係数（α係数）と同様に行うことが可能であるため（山森，2002），信頼性が高いと解釈する1つの基準は一般化可能性係数が0.80以上である。そのため，本研究で評価を行った10名の教員の平均点（または合計点）を使用した場合，本研究での生徒の自由英作文は十

分な信頼性をもって評価されたと解釈できる。逆に本研究での大学生の評定者6名の平均点や合計点では，若干信頼性が低い評価であったと解釈できる。ただし，このような解釈は，通常の信頼性係数による分析でも行うことができるため，より有用な知見を得るために，表1と表2に示された分散成分の推定値を式（1）に当てはめることで，一般化可能性係数のシミュレーションを行った。自由英作文の評価は主観の入る余地の大きいものであると考えられるため，一概にカッティングポイントを0.80以上の一般化可能性係数とするよりは，実用性を考慮に入れて，例えば工藤・根岸（2002）が指摘するように0.60であってもある程度信頼性が高いと判断することもあり得る。そのような場合にもシミュレーションは有効な手段であると言える。

そこで次に，表1と表2に示された分散成分の推定値を式（1）に当てはめていくことで，一般化可能性係数の変化のシミュレーションを行った。本研究の目的の1つは，上述したように，自由英作文の評価を行う評定者の評価経験や人数が異なれば，信頼性（一般化可能性係数）がどのように異なるかという情報を得ることである。そのため，評定者の評価経験の違い（評価に熟達した教員であるか，評価に慣れていない大学生であるか）を基軸にして，評定者の人数を変化させていくことでシミュレーションを試みた。なお，評定者の人数を変化させてシミュレーションを行う際には，式（1）の*Nr*の部分の値を希望の人数の値に変えて計算をすればよく，また，本研究では行わないが評価尺度の項目数を変えるのであれば*Ni*の部分の値を希望の値に変えればよい。[7]

図1には，評価尺度としてESL Composition Profileを用いた場合の一般化可能性係数のシミュレーション結果を示した。今回のシミュレーションでは，評定者の人数を1～20名とした。この結果から本研究に参加した評定者のうち，教員の評価では7名であれば一般化可能性係数が0.80を超えるのに対し，大学生の評価では20名でも0.80を超えないことが示された。また，基準を緩めて0.60のラインを見ると，教員の評価では3名であれば0.60を超えるのに対し，大学生の評価では6名必要なことが示された。そして，1名での評価における一般化可能性係数は，教員の評価では0.40を超える一方で，大学生の評価では0.30に満たないことが示された。

図2には，評価尺度として「観点別評価」を用いた場合の一般化可能性係数のシミュレーション結果を示した。この結果からESL Composition Profileと同様に教員の方が大学生よりも信頼性の高い評価を行っていたことが示された。しかしながら，子細に比較していくと，「観点別評価」の方がESL Composition Profileよりも若干信頼性（一般化可能性係数）が低いことが，つまり0.80，0.60といった一般化可能性係数を満たすには，教員・大学生ともに「観点別評価」の方が概ね1名多くの評定者を必要とすることが示された。

考察

G研究の考察

ここでは，検討内容1）のG研究の結果から，特に評定者の変動要因に関する3点の考察を行う。

1点目として，2種類の評価尺度のいずれにおいても，教員・大学生ともに変動要因「評定者」の主効果の分散成分の推定値の割合が，全変動要因中で最も大きいことが示された。このことは，評価尺度や評定者によって程度の差はあ

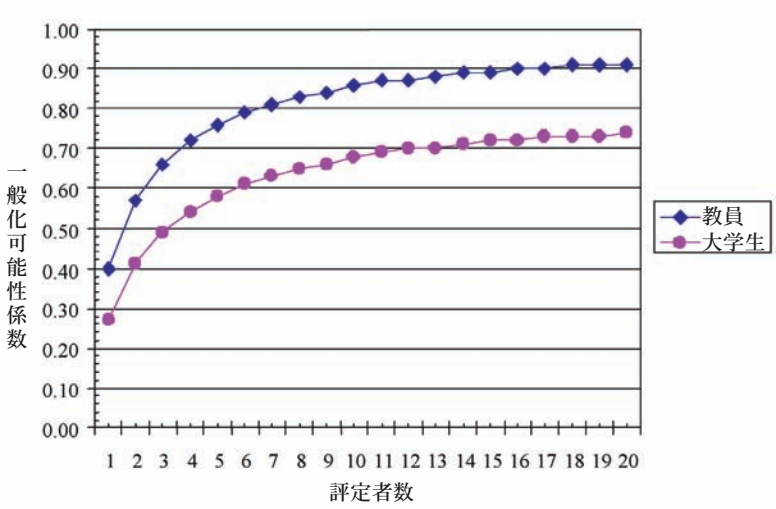

図1. 一般化可能性係数のシミュレーション結果（ESL Composition Profile）

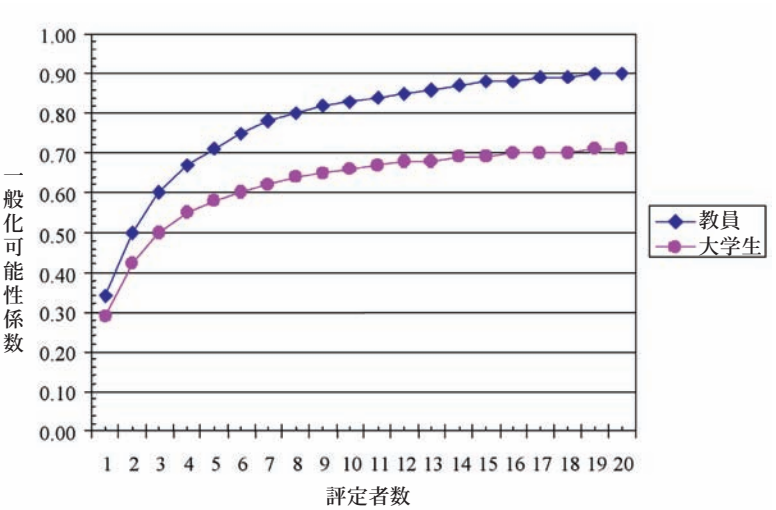

図2. 一般化可能性係数のシミュレーション結果（「観点別評価」）

れども，自由英作文の評価には評定者の主観の入る余地が大きいということが示されたものと解釈できる。ただし，このことは本研究で1〜10点という幅の広い評価尺度を用いたことにも起因すると考えられ，実際の教育現場で多く使用

されると考えられる3段階や5段階の尺度を用いた場合には，「評定者」の分散成分の割合は小さくなることが予想される。

2点目として，教員・大学生ともに「生徒×評定者」の交互作用の分散成分の推定値の割合が比較的大きく，特に教員の評価で顕著であった。この割合が大きいということは，例えば，ある生徒に対してどの評価項目でも高い評価を与えた評定者がいた一方で，その生徒に対してどの評価項目でも低い評価を与えた評定者がいたという，与えた評定値にばらつきがあった傾向を示している。また，教員の評価で「評定者×項目」の交互作用の分散成分の推定値の割合が比較的大きかった。この割合が大きいということは，例えば，ある評価項目でどの生徒に対しても高い評価を与えた評定者がいた一方で，その評価項目でどの生徒に対しても低い評価を与えた評定者がいたという，評価項目の捉え方にばらつきがあった傾向を示している。これらの傾向が教員に顕著であったことは，同じデータを用いた山西（2004）の研究で，相関係数による分析結果から解釈されたように，「教員は日常的な作文に対する指導と評価の中で独自の評価観を確立」（p. 202）していることに起因すると言える。しかし，G研究の結果からは，教員のその独自の評価観は，一度ある生徒やある評価項目に対する良し悪しの基準ができあがると以後はその基準に基づいて評価を行うという評価の偏り，いわゆるハロー効果を生じさせていた可能性があると解釈できる。

3点目として，大学生の評価において，残差に相当する「生徒×評定者×項目」の交互作用の分散成分の推定値の割合が大きかった（ESL Composition Profileで21.93%，「観点別評価」で26.96%）。この残差が大きいということは，大学生の評価は，本研究のG研究での各変動要因（生徒，評定者，評価項目）では説明困難な部分が大きかったことを意味する。つまり，本研究に評定者として参加した大学生は，本研究で用いられた評価項目を生徒の自由英作文評価に十分に結びつけられなかった部分が大きかったことを意味し，その結果として一般化可能性係数が教員の評価よりも低いものであったと解釈できる。

D研究の考察

次に，検討内容2）のD研究の結果から，評定者数と信頼性に関する考察を行う。

本研究では，使用した2種類の評価尺度のいずれにおいても，教員の評定者の方が大学生の評定者よりも信頼性（一般化可能性係数）の高い評価を行っていたことが示された。このことは山西（2004）の研究において，信頼性係数と相関係数による分析によって示された結果と同様であるが，本研究ではD研究のシミュレーションによって，より具体的に何人の評定者であれば十分に（または，ある程度）信頼性の高い評価を行うことができるのかを検討することが可能になった。例えば，十分に信頼性の高い評価を一般化可能性係数が0.80以上であるとするならば，本研究で用いた「観点別評価」尺度を本研究に参加した教員が用いた場合には8名の評定者が必要であること，自由英作文に対するある程度信頼性の高い評価を工藤・根岸（2002）に倣って0.60以上とするならば，3名の評定者が必要であること，といった検討を行うことが可能である。このような検討は，以下の教育的示唆に示されるように，評価の実用性と信頼性に関しての診断的な改善に結びつくと考えられる。

教育的示唆

ここでは以上の考察に基づいて，本研究での評定者と評価尺度に対する教育的示唆を3点，一般化可能性理論を利用した評価の改善という観点から述べていくものとする。

1点目として，教員に関しては，D研究の結果から大学生との比較においては信頼性（一般化可能性係数）の高い評価を行っていることが示され，特にある程度信頼性の高い評価（0.60以上）を行うことは，比較的少ない人数（3名）で可能であることが示された。ただし，教員であっても1〜2名では，自由英作文に対する信頼性の高い評価を行うことは困難であったと指摘できる。さらに，G研究の結果から生徒や評価項目に対する捉え方に対する偏り（ハロー効果）が示されたため，評価への「熟達」が必ずしも適確な評価に結びつかなかった可能性も指摘できる。そのため，評価の改善という観点からは，評価前に他の教員と評価項目の読み合わせや吟味を行うことや，評価後に他の教員と評価結果の違いを検討しあうようなトレーニングが重要であると言える。特に，本研究で用いた「観点別評価」のような新しい評価尺度や各学校での独自の評価尺度を使用する場合には，そのようなトレーニングを十分に行っておくことが重要であると言え，そのために一般化可能性理論を用いたシミュレーション結果などの情報を参照することは有用であろう。

2点目として，大学生に関しては，D研究の結果から十分に信頼性の高い評価（0.80以上）を行うことは困難であったことが示された。また，ある程度信頼性の高い評価（0.60以上）を行うためにも，教員の倍程度（6名）の人数が必要であることが示された。このこととG研究の結果から，大学生は自由英作文の評価そのものに慣れることが重要であると考えられる。そのため，評価の改善という観点からは，例えば大学の教員養成課程で大学生に対して自由英作文の評定作業の指導を行うならば，その際に大学生の評定者に対して一般化可能性理論を用いたシミュレーション結果などの情報を示すことで，効果的に指導を行っていくことができると考えられる。そして，ある程度評価に慣れてからは，大学生も上述の教員と同様に，評定者同士による評価結果の検討を行っていくことが有効であろう。

3点目として，2種類の分析的評価尺度に関しては，D研究の結果から本研究の2種類の評価尺度は同じ分析的評価尺度であっても，教員・大学生ともに「観点別評価」の方がESL Composition Profileよりも若干信頼性が低かったことが示された。このことから，先行研究で多く用いられているESL Composition Profileを基準とするならば，本研究の「観点別評価」で評価することには，評定者の評価経験に関係なく相対的に若干の困難が伴ったと考えられる。そのため，評価尺度の改善という観点からは，一般化可能性理論を用いたシミュレーション結果などの情報を参照することで，評価尺度の実用性と信頼性の兼ね合いを考慮に入れながら評価項目の内容や評価基準などの吟味を行っていくことが有効であろう。

結語

最後に，本研究全体に関する結語を述べる。本研究では，高校生の自由英作文評価の検討に一般化可能性理論を用いたことで，本研究での評価結果に対する具体的かつ詳細な検討を行うことができただけでなく，今後の評価の改善の

ための問題点を提示することもできた。得られた結果は，あくまで本研究での評定者と評価尺度に関するものであるため，それをそのまま他の状況に適用することには留意する必要があるものの，1つの事例とそれに関連した改善案が示された点において有意義なものであったと言える。今後，本研究のような事例が，他の生徒，評定者，評価尺度，作文課題などにおいても多く報告されていくことで，困難な領域であると思われてきた自由英作文評価に対する知見が積み重ねられていくことが期待される。

\

山西博之は，広島大学教育学研究科の博士課程に在籍する大学院生である。また，広島大学附属中・高等学校で英語科の非常勤講師を務めている。研究テーマは，高校生の自由英作文の指導と評価に関する諸問題である。

注

1. 本研究には補足的な目的が2点ある。それらは，1）一般化可能性理論を用いた研究の方法を示すこと，2）古典的テスト理論を用いた研究結果との比較を可能にすること，である。1）に関しては，一般化可能性理論についての解説を行うこと，結果を導くための計算式や結果の解釈の方法を示すこと，そして基本文献を示すことによって配慮した。2）に関しては，古典的テスト理論を用いた山西（2004）と同じデータを用いることで，両者の比較を行うことと相補的な知見を得ることが可能になるようにした。
2. 池田（1994）の主張は以下の通りである。
「記述式テストではその教育的価値の重要さと裏腹に，対象とする受験者個人の能力に加えて，評定者の主観的判断の差異が混入し，さらには課せられる課題差などが複雑に関連し合ってそれらを明確に区別することが困難である。そのため公平な測定評価ができないものとして，避けられる傾向が強かった。そうした状況で得られた数値からどこまでの一般化が可能であるのか，その理論的基礎を与えるものとして登場してきたのが一般化可能性理論である。それは個人間の能力差だけではなく，見方を変えれば，評定者間の差異分析や課題間の差異分析，あるいは教授法の比較などの実験的研究にも同様に扱える一般性のある理論である。」（p. iii）
3. 評価尺度や評価の手順に関する詳細な記述は，山西（2004）を参照。
4. 本研究の「観点別評価」用の尺度は，中学校用の評価規準（国立教育政策研究所， 2002）に基づいて作成され，評価は2003年8月に行われた。その後，高等学校用の評価規準は2004年3月に完成し，同年6月にウェブ上に公開された（国立教育政策研究所, 2004）。
5. 分散成分の推定には，分散分析に基づいたSPSSなどの統計パッケージを用いた方法のほかに，構造方程式モデリング（SEM）に基づいたAMOSなどのソフトウェアを用いた方法（例えば，中村， 2003）やBrennanの開発したソフトウェアであるGENOVA（http://www.education.uiowa.edu/casma/computer_programs.htm）を用いた方法（例えば，山森, 2003）などがある。
6. Brennan（1992）によると，分散成分の推定値に負の値がある場合，その値を0に修正して一般化可能性係数を算出することができる。ただし，そ

の値が他の分散成分の推定などに使われる場合は，推定結果の偏りを避けるために，負の値をそのまま使用する。本研究では，分散成分の推定値の割合を算出するために，負の値をそのまま使用した。

7. シミュレーションのためにSPSSやMicrosoft Excelなどのソフトウェアで式（1）と同様の計算式を作成しておくことで，計算を一度に自動的に行うことが可能になり，手計算の手間を省くことができる。

参考文献

Bachman, L. F. (1997). Generalizability theory. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol.7. Language testing and assessment* (pp. 255-262). Dordrecht: Kluwer Academic Publishers.

Brennan, R. L. (1992). *Elements of generalizability theory* (Rev. ed.). Iowa City: ACT Publications.

池田央 (1994).『現代テスト理論』朝倉書店.

石田潤・森敏昭 (1985).「小学生の文章表現の発達的変化」『広島大学教育学部紀要』33, 125-131.

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.

梶井芳明 (2001).「児童の作文はどのように評価されるのか？－評価項目の妥当性・信頼性の検討と教員の評価観の解明－」『教育心理学研究』49, 480-490.

国立教育政策研究所 (2002).「評価規準の作成，評価方法の工夫改善のための参考資料－評価規準，評価方法等の研究開発（報告）－ . Retrieved December 18, 2004, from http://www.nier.go.jp/kaihatsu/houkoku/saisyu.htm

国立教育政策研究所 (2004).「評価規準の作成，評価方法の工夫改善のための参考資料－評価規準，評価方法等の研究開発（報告）－」. Retrieved December 18, 2004, from http://www.nier.go.jp/kaihatsu/kou-sankousiryou/html/tobira.htm

工藤洋路・根岸雅史 (2002).「自由作文の採点方法による採点者間信頼性について」*Annual Review of English Language Education in Japan (ARELE), 13*, 91-100.

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.

文部省 (1999).『高等学校学習指導要領解説 外国語編 英語編』開隆堂出版.

中村健太郎 (2003).「一般化可能性係数の算出」第67回日本心理学会ワークショップ『構造方程式モデリングはこう使う！』発表資料. Retrieved December 20, 2004, from http://www.littera.waseda.ac.jp/faculty/tyosem/jpa67/jpa67.html

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

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, 14*, 62-70.
山森光陽 (2003).「中学校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』51, 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－」三浦省五（監修）前田啓朗・山森光陽（編著）『英語教師のための教育データ分析入門－授業が変わるテスト・評価・研究』（pp. 82-89）大修館書店.
山西博之 (2004).「高校生の自由英作文はどのように評価されているのか－分析的評価尺度と総合的評価尺度の比較を通しての検討－」 *JALT Journal, 26*, 189-205.

Appendices

Appendix A: 課題A

指示文：下の絵の内容を英語で説明してください。この絵がどのような状況を表しているかをよく考えて，その内容が他の人に分かるように説明してください。

（出典：実用英語技能検定準2級問題集）

Appendix B: 課題B

指示文：Describe something strange or frightening you have witnessed or experienced in your life.

Appendix C: ESL Composition Profileの記述統計量（20点満点）

	Min		Max		*M*		*SD*	
評価項目	教員	大学生	教員	大学生	教員	大学生	教員	大学生
Content	7.50	11.83	12.60	16.50	9.72	13.58	1.21	1.39
Organization	7.30	11.00	13.60	16.33	9.71	13.20	1.32	1.40
Vocabulary	8.10	11.33	13.00	16.00	9.58	13.41	1.13	1.15
Language use	7.70	10.33	12.80	14.83	9.30	12.36	1.05	1.29
Mechanics	8.60	10.83	13.50	15.17	10.12	13.62	1.04	1.09

Note. 教員（$n = 10$），大学生（$n = 6$）が20名の生徒に与えた評定値の平均値.

Appendix D:「観点別評価」の記述統計量（20点満点）

	Min		Max		*M*		*SD*	
評価項目	教員	大学生	教員	大学生	教員	大学生	教員	大学生
言語活動への取組	10.30	10.17	15.00	17.33	12.48	13.58	1.14	1.85
コミュニケーションの継続	9.60	9.50	14.30	17.00	11.82	13.32	1.15	1.92
正確な表現の能力	9.20	11.67	13.20	15.00	10.59	13.13	0.97	1.15
適切な表現の能力	8.80	11.00	13.20	16.00	10.50	12.81	1.00	1.36
言語についての知識	9.10	11.17	13.00	16.33	10.44	12.84	0.92	1.42
文化についての理解	8.22	11.50	12.82	15.50	10.02	12.92	1.02	0.99

Note. 教員（$n = 10$），大学生（$n = 6$）が20名の生徒に与えた評定値の平均値.