# 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

兵庫県／神戸市立大池中学校　教諭　占部 昌蔵

**概要**　自由英作文を評価するときに，評価の信頼性を議論されることがある。では，どのようにしてその信頼性を高くすることができるのか。本研究では，トレーニングを受けた評価者が行う評価が，前回の評価に比べてどの程度変化するのかを調べることによって，項目応答理論を応用した評価者トレーニングの有効性を検討した。同時に，評価者の背景によって信頼性に違いが見られるのかも検討した。評価基準は，ESL Composition Profile を使用した。この評価基準は，内容，構成，語彙，言語使用，（句読点，文法などの）メカニクスの５観点から構成されている。12名の英語科教員と８名の大学院生が，100名の高校生が書いた英作文を，この評価基準を用いて評価した。

結果，今回の評価者トレーニングは効果があることが確認された。また，評価者の背景によって信頼性に大きな違いは見られなかった。

## 1 はじめに

近年，インターネット時代と呼ばれることが多くなり，それに伴って，英語による優れた文章を書くことは，今まで以上に重要になってきている。そして，ライティングに関しては，学校外では，特に，社会的には文章技術の需要はますます増加する傾向にある。しかし，学校内では「実践的コミュニケーション能力」の育成を「スピーキング・リスニング能力の育成が中心」と誤解されてきたためか，ライティングの重要性は，それほど強くは認識されてこなかった。そして，高校段階において，「複数の文は書くことはできても，内容的に一貫した文章を書くことができない」（国立教育政策研究所, 2004）という報告もあるほどである。しかし，大学入試問題に占める自由英作文問題は増加（旺文社, 2005）する傾向にあったり，英検の１級に自由英作文の問題が導入されたり，G-tec for student でも自由英作文の問題が必ず出題されたりするなど，テスト団体によるテストにおいても，まとまりのある英文を書く能力が今まで以上に求められるようになってきている。しかし，自由英作文の指導は，評価に信頼性が得られにくいと思われているためか，高校の授業では敬遠されがちのようである（宮田（編）, 2002）。確かに，自由英作文の評価は，その性格上，完全には客観的に評価しにくく，その評価に評定者の主観が入ることが避けられない面がある。では，どのようにしたら評価の信頼性を高めることができるのだろうか。そのような疑問への１つの解決策として，本研究では，FACETS（ラッシュ・モデル分析ソフトウェア）を用いて評価を分析し，その分析から得られる情報を評価者トレーニングに生かし，その結果，信頼性向上にどの程度貢献できるのかを試みる。

## 2 研究の背景

### 2.1 評価

#### 2.1.1 評価の信頼性

評価では，いくつかの要素が重要であるのだが，信頼性（reliability）に関して，Weir（1990）では，テストにおいて信頼性は，妥当性（validity），実用性（practicality / feasibility）とともに，重要な要素

の 1 つで，あると主張している。これに加えて，真正性（authenticity）も重要な要素の 1 つであると考えられている（Bachman & Palmer, 1996）。

### 2.1.2　評価方法

英作文を評価するときの評価方法としては，全体的評価（holistic scoring），分析的評価（analytic scoring），特定要因の評価（primary trait scoring）などがある。

全体的評価とは，評価者個人の判断をもとにして全体評価として 1 つのスコアをつける方法である。代表的なものとして，TOEFL の writing section での評価があり，6 段階の全体的評価になっている。しかし，Weigle（2002）は，全体的評価は 1 つのスコアしか得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。ただ，短時間で採点が可能で，実用性が高いという利点があるために，テスト団体からこの評価法が採用されることも多い。

分析的評価とは，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。それぞれの下位能力ごとに評価されるため，テスト受験者に診断的情報を提供できるという点でプラスの波及効果（washback effect）が高いことが挙げられる。現在まで頻繁に利用されているのは，Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey（1981）による ESL Composition Profile（資料参照）である。この評価法では，内容（content），構成（organization），語彙（vocabulary），言語使用（language use），（句読点，文法など）のメカニクス（mechanics）の 5 つの項目をそれぞれ30, 20, 20, 25, 5 の配点で採点し，その合計を得点と見なすものである。しかし，時間がかかるという欠点や部分の総和は全体と同じではないという批判的な指摘もある。

特定要因の評価とは，タスクごとに，そのタスクで最も測りたいものを想定し，その採点規準を示した評価である。

それぞれの評価法に，長所，短所があるのだが，高等学校レベルの学習者の立場で考えると，診断的情報を提供できるという点で分析的評価が優れているように思われる。

そして，このような言語運用能力を測るテストにおいては，まず，妥当性の高いタスクが求められる。また，それと同時に，信頼性の高い評価結果も求められる。そして，採点結果の信頼性を測る方法として，伝統的に行われてきた古典的テスト理論，項目応答理論，古典的テスト理論の発展形である一般化可能性理論などがある。

## 2.2　項目応答理論

### 2.2.1　項目応答理論

項目応答理論とは，ある困難度を持ったテストの項目に，ある能力を持った受験者がどのように応答するか，ということに関して確率的なモデルを設定し，それに基づいて受験者の応答データを分析したり，テストを開発したりするための理論（静・竹内・吉澤, 2002）である。項目応答理論では，素点や正答数に基づく得点などを利用するのではなく，自然対数を用いて logit という単位で数値を求め，その数値をもとに項目の特性や受験者の能力を推定する。そして，項目応答理論には，古典的テスト理論では難しかった等間隔の目盛りを持つ尺度が比較的簡単に得られるという利点やテストに依存しない受験者能力の推定（test-free person measurement）が行えるという利点（中村・大友, 2002）などがある。近年，欧米だけでなく日本でも，この理論を応用した英語熟達度テストが見られるようになってきている。

### 2.2.2　ラッシュ・モデル（Rasch model）

デンマーク人数学者 Rasch が創出した数学的モデル。項目応答理論の中の最も基本となるものとして，1 パラメーター・ロジスティック･モデル（1-parameter logistic model）と呼ばれる場合もある（静他, 2002）。もとは，ラッシュ・モデルは二分するデータの分析に制限されていた。その後，このラッシュ・モデルは，項目困難度，受験者の能力という 2 つの相（facets）に加えて他の相も分析できる多相ラッシュ測定（many-faceted Rasch measurement）に発展してきた。そして，Linacre（1996）は，その代表的なソフトウェアとして FACETS を開発し，今日まで発展させてきている。

## 2.3　評価者

### 2.3.1　評価者内信頼性

評価者内信頼性とは，ある評価者がある基準の使用においてどの程度一貫しているかの指標である。言い換えると，評価者による評価が，時間経過によって判断の一貫性が保たれているかどうかの程度

を表しているとも言える。古典的テスト理論では，評価者内信頼性の計算は，同一の評価者による，時間経過を伴う同一試験の 2 回の評価結果を比較し，相関係数を求めることによって算出する。しかし，多相ラッシュ分析を使用すれば，同一試験の 2 回の評価結果を必要とせず，評価者内信頼性は，1 回の評価結果から算出することができる。

### 2.3.2 評価者トレーニング

評価者トレーニングとは，評価に採点する側の主観が入る場合に，複数の採点者が一貫した基準で採点，評定できるようになるために実施されるトレーニングのことである（静他 , 2002）。また，McNamara（2000）によると，このトレーニングでは，各評価者は事前にいくつかの異なるレベルのパフォーマンスを評定するよう求められる。そして，グループ内で自分の下した評価と他人が下した評価を比較し，その差異を検討するのであるとある。そして，このトレーニングの重要性を認めながらも，Weigle（1994）は，トレーニングによってエラーの数を減じることはできても，評価者間における厳しさの差異は完全には取り除くことができないと述べている。また，Fulcher（2003）によると，トレーニングは評価者間における点数の差異は減じることができるが，評価者はそれぞれ違った採点方法で採点している事実が研究で明らかになっていると述べている。

## 2.4 先行研究

英作文評価の信頼性については，評価方法の違いによって（全体的評価方法，分析的評価方法など），どの方法を使った方がその信頼性が高くなるのか，評価者を何人用意すればある程度高い信頼性が得られるか，などについては国内・国外を問わず研究されてきている。工藤・根岸（2002）の研究では，日本人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い，3 種類（印象的，全体的，分析的）の採点方法で評価を行い，採点方法の違いから生じる信頼性の差異が検討された。被験者（評定者：14名の大学院生）が高校生の自由英作文（36名分）を上記 3 種類の採点方法によって評価した結果を用いて，採点者間信頼性を，相関係数をもとにした公式によって算出した。その結果，分析的採点方法（ESL Composition Profile 使用）が最も少ない人数で，ある程度の信頼性係数を得ることができたと報告している。

スピーキングテストの評価についてではあるが，項目応答理論を使用した研究に秋山（2002）がある。この研究では，日本人中学生対象のスピーキングテストの評価に関して，項目応答理論を応用し，評価者の信頼性を取り扱い，評価者間信頼性と評価者内信頼性という異なった信頼性の差異の比較，検討が行われた。7 名の評価者で中学生（109名）のスピーキングテストを採点した。評価者間信頼性はピアソンの相関係数を用い，その結果，信頼性係数は全体で約0.9と高い数値が得られた。しかし，評価者内信頼性は項目応答理論応用ソフトウェアを用いて計算され，その結果，2 名の misfit rater（不適合評価者）が見つかったとしている。このことは，評価者間信頼性が高い場合でも，評価者の中には一貫性が欠ける採点をしている場合があるということを示している。

評価者信頼性に影響を及ぼすものとして，評価経験の有無や評価者トレーニングなどが考えられる。しかし，評価経験の有無や評価者トレーニングによって，どの程度その評価の信頼性が高いのか，または，向上するかについての研究はあまり多くは見られない。山西（2005）の研究では，評価者の評価経験の違いによって信頼性がどの程度異なるかを調べている。もとになった作文は高校生が書いた英作文20部で，被験者（高校教師10名，大学生及び大学院生 6 名の計16名）が分析的採点方法（ESL Composition Profile 使用）に従って採点した。その採点結果を，一般化可能性理論を用いて検討している。その結果から，まとめとして，高校教師の方が大学生・大学院生より少ない人数である程度の高い信頼性が得られるとしている。また，Weigle（1998）の研究は，アメリカの大学で，ESL 作文評価者へのトレーニングの効果を調べたもので，16名の評定者が30名分の作文を評価した。作文評価経験の有無とトレーニングの事前事後で評価がどう異なるかを報告している。結果は，トレーニング後の方が信頼性を表す数値が全体的に見ると良くなっており一貫性が向上したと言える。しかし，個人で見るとその数値が高くなっているが，一貫性が低下したと言える者もいた。また，トレーニングによる効果は，評価の一致よりも一貫性の向上にあったとしている。このように，評価者信頼性に影響を及ぼす評価経験の

有無や評価者トレーニングを扱った研究は少なく，結論を一般化するには，さらなる研究が必要になる。

# 3 本研究

## 3.1 目的

本研究では，FACETS の結果を用いたトレーニングを受けた評価者が行う評価は，前回の評価に比べてその信頼性をどの程度向上できるかを検討する。

## 3.2 リサーチ・クエスチョン

本研究の目的及び先行研究から，以下のようにリサーチ・クエスチョンを立てた。

1：一貫性が低い評価者の評価は，評価者トレーニングによって，どのような変化が見られるか
2：英語科教員と大学院生の間では，評価における信頼性において違いは見られるのであろうか

## 3.3 研究方法

### 3.3.1 実験協力者

評価者として，英語科教員12名（中学校所属 4 名・高等学校所属 8 名：教員歴 7 年～23年），大学院生（英語教育及び国際コミュニケーション分野専攻）8 名の計20名が参加した。

### 3.3.2 手順

まず，予備実験として，公立高等学校 2 年生の書いた英作文 3 クラス（108名）分のコピーをもとに，3 人の評価者（筆者を含む）で ESL Composition Profile を採点基準として評価し，本実験で使用するためにレベルを代表するベンチマーク作文 4 部を抽出した。そして，この 3 人の評価者の平均をもとに，等質になるよう英作文50部を 2 セット（セット A とセット B）を用意した。そして，採点基準に従って，20人の評価者 1 人につき英作文50部を採点してもらった（セッション 1 ）。その結果に基づき，複数の実験協力者にインタビューを実施した。セッション 2 を行う前に，1 回目の分析の結果，misfit rater と判断できる評価者を対象に，評価者トレーニングを行った。後日，セッション 1 と同様に 2 回目の採点を実施（セッション 2 ）した。

### 3.3.3 分析方法

まず，評価者内信頼性を分析するために，FACETS を使用する。McNamara（1996）より，得られるデータの Infit 値が（MnSq の場合）1.3以上（fit 値はデータがモデルにどの程度適合しているかを示す。値が1.0ならば最も適合していると解釈する。MnSq 以外に Zstd も FACETS では表示される。Zstd はその値が 0 ならば最も適合していると解釈する）の数値を示す評価者を評価の一貫性が低い misfit rater として判断する。次に，FACETS から得られるデータの Unexpected Responses の表（これは，ある作文に対して与えた評価が，全評価者の与えた平均値からどれくらいかけ離れていたかを表しているのではなく，その評価者の厳しさに応じて，ある作文に対してこのくらいのスコアを与えるべきであろうという予想値から大きく離れた場合にだけ，挙がってくるものである）を利用する。今回は，その基準を residuals を±2以上のものとした。ここで，misfit rater に該当する評価者とそうでない者の予期せぬ反応数を比較する。そして，その表から，該当する評価者の採点傾向，予想する以上のかけ離れた反応だった作文を抽出し，評価者トレーニングに生かすものとする。また，評価者間信頼性（相関係数を算出）についても比較検討する。

そして，トレーニングの成果が上がったかどうかの指標として，Infit 値，Unexpected Responses の数，評価者間信頼性の数値について変化の度合いを検討する。

### 3.3.4 評価者トレーニング実施内容

セッション 2 を行う前に，1 回目の分析の結果，misfit rater と判断できる 4 名の評価者に，一種の評価者トレーニング（セッション 1 での結果報告と採点方法変更の指示）を行った（実験協力者に時間の制約があり，予定していた通常のトレーニングを行う十分な時間は確保できなかったため，このような内容のトレーニングとした）。

セッション 1 での結果報告では，面接をしながら該当する評価者の採点傾向の報告と各項目における点数が最高点と最低点が多かったために分布がいびつになっていることなどを伝えた。

採点方法の変更は，FACETS の分析結果から一貫性の高かった評価者が行っていた方法を参考に，①20部の作文ごとに区切りを入れる（20人目の作文の

採点が終わった後休憩を入れ，もう一度１～20人目までを見直す。すぐには21人目の作文を採点しない），② 評価項目ごとに採点をする（一度に１人の作文について５項目を採点しない），③ 採点の見直しをする，の３点とした。

# 4 結果

## 4.1 FACETS による分析概観

図１は，１回目の採点を FACETS により分析した結果の全体像を表した図である。左端の logit は，項目応答理論で使われる単位であり，この図では０が平均を示している。例えば，０よりこの数値が上の生徒は平均より能力が高いことを表しており，０から離れる距離が遠いほどより能力が高いことを表している。２番目の列は，生徒の能力を表しており，＊の数と位置で度数分布を表している。次の列は，項目の困難度を表しており，それぞれの項目がどれくらい難しいかを表している（ここでは，数字の１は内容，２は構成，３は語彙，４は言語使用，５は句読点，文法などのメカニクスを表している）。右端の列は，評価者の厳しさを表している。上位に位置する評価者ほど厳しく採点し，下位に位置するほど甘く採点することを表している。

▼図１：分析結果（全体像）

```
------------------------------------------------------------
|   logit   | 受験者 | 項目    | 評価者                    |
------------------------------------------------------------
|           |        |         |                           |
|           |        |         |                           |
|           |        |         |                           |
|           | .      |         |                           |
|           |        |         |                           |
+      1    +        +         +                           +
|           | .      |         |                           |
|           | .      |         |                           |
|           | **     |         | 12                        |
|           | *.     |         |                           |
|           | *.     |         | 7                         |
|           | ***    |         |                           |
|           | *****  |         | 18                        |
|           | ***    |  1      | 11  5                     |
|           | *****. | 4  5    |                           |
*      0    *  *****  *         *                           *
|           | .      | 2  3    | 4  6  9  14  13  15       |
|           | **.    |         |                           |
|           | ****   |         | 3  8  10  17              |
|           | ***    |         | 2  20                     |
|           | ***    |         | 16                        |
|           | .      |         |                           |
|           | *      |         | 19                        |
|           | *.     |         | 1                         |
|           |        |         |                           |
+     -1    +        +         +                           +
|           | .      |         |                           |
|           |        |         |                           |
|           | .      |         |                           |
|           |        |         |                           |
|           |        |         |                           |
------------------------------------------------------------
```

## 4.2 評価者のデータ

■表１：評価者データ１（セッション１）

| 評価者 | 厳しさ | 誤差 | Infit 値（Mn） |
|---|---|---|---|
| 1 | -0.85 | 0.05 | 0.69 |
| 2 | -0.40 | 0.04 | **2.69** |
| 3 | -0.33 | 0.04 | 0.72 |
| 4 | -0.10 | 0.04 | **1.39** |
| 5 | 0.22 | 0.04 | **2.17** |
| 6 | -0.07 | 0.04 | **1.38** |
| 7 | 0.56 | 0.04 | 1.08 |
| 8 | -0.33 | 0.04 | 0.85 |
| 9 | -0.15 | 0.04 | 0.41 |
| 10 | -0.36 | 0.04 | 1.17 |
| 11 | 0.17 | 0.04 | 0.69 |
| 12 | 0.68 | 0.04 | 0.62 |
| 13 | -0.12 | 0.04 | 0.99 |
| 14 | -0.06 | 0.04 | 1.27 |
| 15 | -0.12 | 0.04 | 0.94 |
| 16 | -0.14 | 0.04 | 0.39 |
| 17 | -0.26 | 0.04 | 0.52 |
| 18 | 0.32 | 0.04 | 1.10 |
| 19 | -0.74 | 0.04 | 0.56 |
| 20 | -0.40 | 0.04 | 0.94 |

（評価者 2, 3, 4, 7, 15, 16, 17, 18 は大学院生，他の評価者は英語科教員）

表１は，セッション１における評価者の厳しさなどをまとめたもので，２列目は，評価者の厳しさを表す。例えば，12の評価者が最も厳しく，１の評価

者が最も甘いということが読み取れる。3列目は，誤差を表す。どの評価者についてもほぼ同じで，非常に小さい値であることがわかる。4列目は，Infit値を表しており，この数値が1.3以上の場合は，その評価者を misfit rater と判断でき，その評価者の評価は不安定であるということを意味している。この結果から今回は，評価者2，4，5，6，の4名を misfit rater とし，この4名をトレーニング対象者とした。

表2は，同様に，セッション2の結果を表している。

■表2：評価者データ2（セッション2）

| 評価者 | 厳しさ | 誤差 | Infit 値（Mn） |
|---|---|---|---|
| 1 | -1.03 | 0.05 | 0.54 |
| 2 | -0.75 | 0.05 | **2.42** |
| 3 | -0.70 | 0.05 | 0.75 |
| 4 | -0.77 | 0.05 | 1.29 |
| 5 | -0.61 | 0.05 | **1.49** |
| 6 | 0.79 | 0.05 | 1.22 |
| 7 | 0.56 | 0.05 | 0.72 |
| 8 | -0.22 | 0.05 | 1.12 |
| 9 | -0.51 | 0.05 | 1.13 |
| 10 | -0.83 | 0.05 | 0.86 |
| 11 | 0.22 | 0.05 | 0.41 |
| 12 | -0.06 | 0.05 | 0.45 |
| 13 | -0.15 | 0.05 | 1.08 |
| 14 | -0.93 | 0.05 | 1.29 |
| 15 | 0.21 | 0.05 | 0.82 |
| 16 | -0.74 | 0.05 | 0.51 |
| 17 | -0.67 | 0.05 | 0.88 |
| 18 | 0.33 | 0.05 | 1.17 |
| 19 | -0.56 | 0.05 | 0.97 |
| 20 | -0.68 | 0.05 | 1.03 |

## 4.3　予想以上の反応

表3は，Unexpected Response をまとめたもので，上記の各評価者によるモデルで予想する以上のかけ離れた反応数と全員の評価者による反応数を表している。

## 4.4　評価者間信頼性

表4は，セッション1及び2における評価者間信頼性（それぞれの評価者と他の評価者との相関係数（ピアソン）の平均値を表している。

■表3：予期せぬ反応数

| 評価者 | セッション1 | セッション2 |
|---|---|---|
| 2 | 53 | 36 |
| 4 | 19 | 5 |
| 5 | 35 | 15 |
| 6 | 11 | 16 |
| 計 | 118 | 72 |
| その他計 | 129 | 92 |
| 総計 | 247 | 164 |

■表4：評価者間信頼性

| 評価者 | セッション1 | セッション2 |
|---|---|---|
| 1 | 0.54 | 0.59 |
| 2 | 0.45 | 0.57 |
| 3 | 0.54 | 0.48 |
| 4 | 0.48 | 0.50 |
| 5 | 0.44 | 0.49 |
| 6 | 0.46 | 0.53 |
| 7 | 0.34 | 0.46 |
| 8 | 0.54 | 0.53 |
| 9 | 0.43 | 0.54 |
| 10 | 0.52 | 0.60 |
| 11 | 0.39 | 0.58 |
| 12 | 0.54 | 0.57 |
| 13 | 0.47 | 0.57 |
| 14 | 0.45 | 0.50 |
| 15 | 0.38 | 0.53 |
| 16 | 0.52 | 0.52 |
| 17 | 0.33 | 0.50 |
| 18 | 0.56 | 0.49 |
| 19 | 0.56 | 0.38 |
| 20 | 0.45 | 0.54 |

■表5：評価者背景ごとの信頼性の平均値

| 評価者 | セッション1 | セッション2 |
|---|---|---|
| 大学院生 | 0.45 | 0.50 |
| 英語科教員 | 0.48 | 0.53 |

# 5　考察

## 5.1　評価者トレーニング

リサーチ・クエスチョン1についてであるが，セッション1と比べてトレーニング実施後のセッ

ション 2 では該当する 4 名の評価者の一貫性を示す数値が下がっている（表 1，表 2）。このことは，このトレーニングが一貫性を向上させることに効果がある可能性を示している。また，セッション 1 でこの数値が低かった評価者がセッション 2 で基準値を上回るほど高くなることはなかった。このことより，該当する 4 名以外にはトレーニングの必要者はいなかったと判断できる。ただし，表 2 の数値が示すとおり，4 名中 2 名は，トレーニング後も misfit rater を示す基準値（1.3）を上回っている。このことは，このトレーニングは効果はあるが，すべての評価者の Infit 値を基準値以下にできるほどの効果はないと考えるのが妥当ではないかと思われる。また，表 3 から，トレーニング実施後のセッション 2 では該当する 4 名の評価者のうち 3 名は，予想以上の反応数を減らしている。このことも，このトレーニングは効果はあるが，すべての評価者に，同様には効いていないと考えられる。Weigle（1998）では，評価者はトレーニングを通して，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られている。今回も同様に，トレーニングによってある程度のすり合わせは行われるようになったのかもしれない。

### 5.2 評価者の背景

リサーチ・クエスチョン 2 についてであるが，英語科教員（N = 12）と大学院生（N = 8）の間では，評価者間信頼性（表 5 ）において U 検定の結果，両グループ間において有意な差は見られなかった（セッション 1 ：U = 46, $p < 0.5$，セッション 2 ：U = 38, $p < 0.5$）。これは，いくつかの先行研究の結果とは異なる。この結果の理由を以下のように考えた。

Weigle（1994, 1998），山西（2004, 2005）では，評価経験の有無が信頼性や一貫性に大きく影響すると述べられている。しかし，評価行動はもっと複雑な要素（評価者の信条，性格，偏見など）がからんでいるので，トレーニング後にもかかわらず信頼性が落ちた評価者が出てきたり（Weigle, 1998），評価経験のない大学生でも評価経験のある教員と同等の信頼性数値を出す者がいたり（山西, 2004）するのではないだろうか。そして，そのような信頼性が落ちた者を例外としてとらえるのではなく，通常の評価者トレーニングだけでは不十分な場合があり，評価経験の有無という背景だけが信頼性に大きな影響を与えているのではないと考えるのが妥当ではないだろうか。また，Fulcher（2003）によると，最近では，評価規準作成者は評価者の受け止め方を反映させるべきではないかという議論が出てきていることや，一貫した判断のためには，トレーニングよりも評価者の評価基準に対する受け止め方と評価規準の内容との間にある溝をどうするかの方が重要になってきているのではないかと述べている。

よって，評価者の考えをある基準へ押し込むようなトレーニングだけではなく，さまざまな方法で評価行動を導き出し，改善していくようなトレーニングが必要であろう。そして，評価の一貫性がある程度維持できるのなら，Bonk, Ockey（2003）の言うように，評価のずれを統計的に調整することも 1 つの手段として考えられる。

また，トレーニングに前後して，評価者からのフィードバックを通して基準表の改善やサンプル作文の改良を実施することや，評価者へのフィードバックを通して評価者自身の内的基準の変容，基準表へのすり合わせが促進されることも考えられる。上記のような方法は，大規模な試験団体や high stakes な試験（その試験の結果が受験者の将来を左右する可能性のある試験）では難しいかもしれないが，中学校・高等学校現場や市町村の地区単位での研修では十分可能であろう。

## 6 研究の限界と教育的示唆

### 6.1 本研究の限界

過去の先行研究に比べれば，実験協力者数26名は，依然として一般化に厳しい規模ではあるが，過去の研究と比べても多い。ただ，最終的な分析の対象となったのは20名による評価という点は，筆者の反省すべき点である。より多くの実験参加者の協力が得ることができれば，違った結果が得られたかもしれないし，より多様な採点者傾向を発見できたかもしれない。

### 6.2 教育的示唆

本研究の結果により，通常考えられる評価者トレーニングだけではなく，評価者との話し合いやインタビュー，採点方法を変えることによっても，評価者内信頼性が向上することがわかった。したがっ

て，テスト実施者は，単に採点を繰り返すだけの評価者トレーニングとその結果によって評価者の信頼性を測るだけでは不十分な場合も出てくるということ留意しておくべきであろう。例えば，そのようなときには，なぜそのような採点傾向になっているのかを採点者からよく聞いた上で，評価者トレーニングを実施することや採点者に採点方法の変更を要求することが望ましいと思われる。また，テストの信頼性を高めるために，Hughes（1989）は，さまざまな提案をしているが，その提案の中にある採点者トレーニングのところで主張している，「採点規準から大きくかつランダムにはずれる採点をする採点者は二度と使わない」というような措置は，指導と評価を一体で進めていく学校現場，つまり，指導した教員または指導した教員と同じ学校の教員が評価をしていく学校現場ではなじまないだろう。

次に，項目応答理論自体は，難解な理論であり，難解な数式の理解も必要であるが，コンピュータとこの理論に基づくソフトウェアがあれば，そのソフトウェアを目的に応じて使用することによって，中学校・高等学校現場においても，それほど煩雑な手続きなしで利用可能である。例えば，校内研修，または市町村の地区単位での，主観的評価を伴うスピーキングやライティングの評価者トレーニングにおいても十分利用可能であると思われる。

## 謝　辞

まず，本研究を行うすばらしい機会を与えてくださった（財）日本英語検定協会と選考委員の先生方に感謝いたします。特に，担当してくださった池田央先生に厚く御礼申し上げます。お忙しいにもかかわらず，中間報告から最終原稿に至る過程で貴重なご助言をいただきました。また，実験計画から実験実施に至る過程で助言をくださった兵庫教育大学の今井裕之先生に深く感謝いたします。また，本研究に協力してくださった兵庫県内，及び，県外の公立中学校・高等学校の先生方にも心から感謝します。それと，兵庫教育大学大学院，神戸大学大学院の院生の皆さんにもご協力いただきました。他にも多くの方からの助言や励ましなどをいただきました。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」. *STEP BULLETIN*, vol. 12, 67-78.
＊Bachman, L.F. and Palmer, A.S.(1996). *Language Testing in Practice*. Oxford: Oxford University Press.
＊Bonk, W.J. and Ockey, G.J.(2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89-110.
Brown, J.D.(1996). *Testing in Language Programs*. NJ: Prentice Hall Regents.
＊Fulcher, G.(2003). *Testing Second Language Speaking*. Essex, UK: Pearson Education.
＊Hughes, A.(1989). *Testing for Language Teachers*. Cambridge, UK: Cambridge University Press.
池田央.(1994).『現代テスト理論』. 朝倉書店.
＊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.
金谷憲(編).(2003).『英語教育評価論』. 東京: 河源社.
＊国立教育政策研究所.(2004).『平成14年度高等学校教育課程実施状況調査』. 東京: 国立教育政策研究所.
小室俊明(編).(2001).『英語ライティング論』. 東京: 河源社.
＊工藤洋路・根岸雅史.(2002).「自由英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.
＊Linacre, J.M.(1996). *A user's guide to Facets*. Chicago: MESA Press.
Linacre, J.M.(2005). Facets for Windows version (Version3.57.0). Chicago: MESA.
＊McNamara, T.F.(1996). *Measuring second language performance*. London and NewYork: Longman.
＊McNamara, T.F.(2000). *Language Testing*. Oxford, UK: Oxford University Press.
＊宮田学.(編).(2002).『ここまで通じる日本人英語』. 東京: 大修館書店.
＊中村洋一・大友賢二.(2002).『テストで言語能力は測れるか―言語テストデータ分析入門』. 東京: 桐原書店.
＊旺文社(編).(2005).『全国大学入試問題正解　英語(国公立大編)』. 東京: 旺文社.
＊静哲人・竹内理・吉澤清美.(2002).『外国語教育リサーチとテスティングの基礎概念』. 関西大学出版部.
＊Weigle, S.C.(1994). Effects of training on raters of ESL composition. *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, UK: Cambridge University Press.
＊Weir, C.J.(1990). *Communicative Language Testing*. Englewood Cliffs, NJ: Prentice Hall.

＊山西博之.(2004).「高校生の自由英作文評価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.

＊山西博之.(2005).「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」. *JALT Journal*, 27, 169-185.

**資料：ESL Composition Profile**

## ESL COMPOSITION PROFILE

STUDENT　　DATE　　TOPIC

| | SCORE | LEVEL | CRITERIA | COMMENTS |
|---|---|---|---|---|
| CONTENT | | 30-27 | EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic | |
| | | 26-22 | GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail | |
| | | 21-17 | FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic | |
| | | 16-13 | VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate | |
| ORGANIZATION | | 20-18 | EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive | |
| | | 17-14 | GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing | |
| | | 13-10 | FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development | |
| | | 9-7 | VERY POOR: does not communicate • no organization • OR not enough to evaluate | |
| VOCABULARY | | 20-18 | EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register | |
| | | 17-14 | GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured* | |
| | | 13-10 | FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured* | |
| | | 9-7 | VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate | |
| LANGUAGE USE | | 25-22 | EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions | |
| | | 21-18 | GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured* | |
| | | 17-11 | FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured* | |
| | | 10-5 | VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate | |
| MECHANICS | | 5 | EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing | |
| | | 4 | GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured* | |
| | | 3 | FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured* | |
| | | 2 | VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate | |

TOTAL SCORE　　READER　　COMMENTS

Copyright © 1981 by Newbury House Publishers, Inc. All rights reserved.