自由英作文における評定者評価の種類と信頼性

その他のタイトル	Types of Evaluation by Raters and Reliability in an English Essay
著者	水本 篤
雑誌名	統計数理研究所共同研究リポート215　「学習者コーパスの解析に基づく客観的作文評価指標の検討」
ページ	43-49
発行年	2008
URL	http://hdl.handle.net/10112/12986

統計数理研究所共同研究リポート 215
「学習者コーパスの解析に基づく客観的作文評価指標の検討」(2008), 43-49.

自由英作文における評定者評価の種類と信頼性

水本　篤
流通科学大学
E-mail:　atsushi@mizumot.com

あらまし　本研究では，自由英作文（エッセイ）の評価方法である，総合的評価（holistic rating）と分析的評価（analytic rating）の 2 種類の内部構造の関係を明らかにし，どちらか一方が他方より優れているのかという点について調査した。また，一般化可能性理論（Generalizability Theory）を利用することにより，測定における誤差の要因，一般化可能性係数，および評価尺度の項目数と評価者人数の検討を行った。結果として，総合的評価・分析的評価のどちらの尺度でもかなりの信頼性が確保できることがわかった。また，一般化可能性理論を用いることにより，誤差の要因の特定や，項目数や採点者数の検討を行い，評価の改善が可能になることを実証した。
キーワード　自由英作文，総合的評価，分析的評価，一般化可能性理論

Types of Evaluation by Raters and Reliability in an English Essay

Atsushi, MIZUMOTO
University of Marketing and Distribution Sciences (Assistant Professor)

Abstract　In this study, a comparison of two common types of rating methods of the essay test, holistic rating and analytic rating, was made in order to clarify which type of rating methods works better. First, the internal constitution of these two rating methods was explored by examining the intercorrelation of the scores of the holistic rating scale and those of each item in the analytic rating scale. Next, variance components and generalizability coefficient were estimated utilizing Generalizability theory, and thus simulating the number of items and rates necessary for reliable measurement. The results indicate that both rating methods function equally well in terms of reliability. Furthermore, it was confirmed that improvement of measurement would be possible with Generalizability theory.
Keyword　English essay, Holistic scoring, Analytic scoring, Generalizability Theory

1. はじめに

自由英作文（エッセイ）の評価（採点）方法には，総合的評価（holistic rating）と分析的評価（analytic rating）の 2 種類があり（Bacha, 2001；山西, 2004），評価の目的によって使い分けられる。総合的評価は自由英作文の全体的な印象で 1 つのスコアをつけるもの

であるが，分析的評価は，いくつかの項目（たとえば，語彙，文法，構成）に対して，それぞれにスコアをつけることになるので，学習者の発達段階に応じて，総合的評価ではできない細かなフィードバックが可能になる。Weigle (2002) による 2 つの評価の比較では，分析的評価の方が総合的評価よりも信頼性が高い反面，分析的評価は時間とコストがかかるので，総合的評価の方が実施可能性（practicability）の面で優れているとされている。このような側面は，それぞれの評価尺度の特徴を考えてみればわかることだが，実際に，学習者の自由英作文を評価するときに，どちらの評価を用いればよいかは，「目的次第」となってしまい，いずれを選択しても，片方を用いたほうが，もう一方よりも信頼性・妥当性の高い測定ができたかもしれないという疑問が残ってしまう。そこで，本研究では，山西（2005a）と同じ研究デザインを利用し，それぞれの評価尺度の比較，信頼性の検討を行うことを目的とした。

2. 方法

2.1. 本研究の目的と分析方法

まず，総合的評価尺度と分析的評価尺度の内部構造の関係を明らかにし，どちらか一方が他方より優れているのかという点を調査した。そのために，同一の自由英作文を 2 人の評定者が総合的評価尺度と分析的評価尺度の両方を用いて，2 回評価し，その評価の平均値を使って，相関分析を行った。

そして，山西（2005b）や，山森（2002）に見られるように，一般化可能性理論（Generalizability Theory）を利用することにより，測定における誤差の要因，および評価尺度の項目数と評価者人数の検討を行った。一般化可能性理論は，測定において存在する測定誤差の成分と大きさを検討する方法で，誤差が評定者の違いによるものなのか，項目の違いによるものなのかなどという点を明らかにし，必要な評定者の数や，項目の数をシミュレーションにより検討することができる（山森, 2004）。一般化可能性理論には 2 つの段階がある。はじめに，評定者，項目などの変動要因が評価に与える影響の大きさ（分散成分）の推定を行う「一般化可能性研究」（Generalizability study; G-study）があり，次に，その一般化可能性研究で得られた分散成分の推定値をもとに，古典的テスト理論での信頼性係数に相当する一般化可能性係数（generalizability coefficient）を算出し，具体的に必要な評定者の数や，項目の数をシミュレーションし検討する「決定研究」（Decision study; D-study）がある（田中 他, 2007; 山森, 2002; 2004; 山西, 2005a; 2005b）。

2.2. 参加者と道具

分析に使用したデータは，水本（2008）で収集されたもので，実験参加者は，関西の私立大学の 2 回生（商学部）を中心とした 40 名の EFL 学習者であり，英語習熟度は TOEIC 300 点台前半の false-beginners（疑似初心者）が大半であった。実験参加者のライティング能力測定には，授業時間内に自由英作文を以下のトピックで書くように指示し，提出さ

れたものを用いた。

自由英作文のトピック

Do you agree or disagree with the following statement?
"It is important for college students to have a part-time job."
Use reasons and specific details to support your answer.

作成の際には，パソコンのワープロソフトで 150 語以上，制限時間を 60 分，辞書の使用は禁止と設定し，本論文の著者の監督の下，授業時間内に提出することを義務づけた。語数は「150 語以上」を自由英作文作成時の条件としていたが，規定語数を満たせない者もいたため，そのような場合には制限時間内に書けた語数を対象とした。

2.3. 自由英作文の評価

自由英作文の評価は，本論文の著者（評定者 A）と，中学校で 3 年間の英語指導経験を有する外国語教育学専攻の大学院生（評定者 B）の 2 名で行った。総合的評価の尺度としては，TOEFL（Test of English as a Foreign Language）の TWE（Test of Written English），および TOEFL CBT（Computer-Based Testing）のライティング・セクションの評価基準を使用した。TOEFL CBT の評価は，2000 年から 2006 年の TOEFL iBT（Internet-Based Testing）の実施まで利用されていたもので，0 点から 6 点からなるが，これを総合的評価尺度（holistic scoring）10 段階に変更して使用した。

分析的評価の尺度としては，Jacobs et al.（1981）の ESL Composition Profile を山西（2004, 2005b）に倣い，10 段階に変更したものを使用した。ESL Composition Profile は Content（内容），Organization（論理・構成），Vocabulary（語彙），Language use（文法・構文），Mechanics（句読点などの形式）の観点から分析的に評価するものである。[1]

表 1　2 名の評定者の採点の平均値と標準偏差

		平均（*M*）	標準偏差（*SD*）
分析的評価	Content	4.55	1.98
	Organization	4.50	1.91
	Vocabulary	4.93	1.60
	Language use	3.68	1.55
	Mechanics	4.56	1.66
	分析的評価の平均	4.44	1.63
総合的評価		3.75	1.95

[1] 今回の研究で使用した ESL Composition Profile (Jacobs et al.,1981）を日本語に訳したものは http://www.mizumot.com/ESL-CompositionProfile.xls を参照のこと。

評価者 2 名の評価基準を合わせるために，今回の実験参加者と同レベルの学習者 5 名が書いた自由英作文を，評定者 A が総合的評価と分析的評価の両方を用いて先に採点し，採点者 B がそれに合わせるという形の評定者トレーニングを行った上で，参加者 40 名分の自由英作文を採点した。2 名の評定者の採点順序（総合的評価と分析的評価の採点の順番）と評価順序（40 名の採点）は順番の影響を相殺するためにランダムにして評価を行った。

表 1 は 2 名の評定者の採点の平均値，標準偏差をまとめたものである。総合的評価の評定者間信頼性を示すピアソン相関係数は $r = .81$ で，分析的評価の信頼性係数は $\alpha = .96$ であったため，2 つの尺度はともに十分な信頼性を持った評価尺度であると判断した。

3. 結果と考察

3.1. 評価尺度間の関係

表 2 は分析的評価と総合的尺度の相関係数をまとめたものである。分析的評価尺度では，Mechanics（句読点などの形式）以外の Content（内容），Organization（論理・構成），Vocabulary（語彙），Language use（文法・構文）は相関係数がかなり高い値（$r = .87$～$.96$）であり，山西（2005）の報告よりもかなり高い値になっている。これは，今回対象とした学習者のレベルが低く，ライティング能力の高い学習者と低い学習者の違いがはっきりと表れた結果ではないかと推測できる。また，総合的評価と分析的評価の各項目の相関係数に関しても，Mechanics（$r = .67$）を除いて，Content（$r = .85$），Organization（$r = .88$），Vocabulary（$r = .91$），Language use（$r = .90$）となっており，分析的評価の平均と総合的評価の平均も $r = 90$ と高い相関となっていることから，分析的評価でも自由英作文の総合的な能力が測定されていると考えられる（山西, 2005）。

表 2　分析的評価尺度と総合的評価尺度の相関係数

	1	2	3	4	5	6	7
1. Content	－						
2. Organization	.96	－					
3. Vocabulary	.91	.94	－				
4. Language use	.87	.89	.93	－			
5. Mechanics	.69	.70	.74	.74	－		
6. 分析的評価の平均	.96	.97	.97	.95	.83	－	
7. 総合的評価	.85	.88	.91	.90	.67	.90	－

また，分析的評価でも，総合的評価でも，結果に差がほとんどでないと考えられるので，片方を用いたほうが，もう一方よりも信頼性の高い評価ができたかもしれないという可能性は少ないであろうと考えられる。ゆえに，2 つの評価尺度の使用は，分析的評価による細分化されたライティング能力の測定を行い，フィードバックを学習者に与えるのか，もし

くは，単純に総合的評価により全体的な学習者のライティング能力を測定したい（もしくは，能力の差を明らかにしたい）のかという，どのような目的がテストにあるかによって使い分けられるべきであるといえよう。

3.2. 評価尺度間の関係

3.2.1. 一般化可能性研究（Generalizability study; G-study）

表 3 は，評定者，項目などの（変動）要因が評価に与える影響の大きさ（分散成分）の推定を行う「一般化可能性研究」(Generalizability study; G-study) を行った結果である。分散成分の推定には分散分析を用いた方法 (池田, 1994; 山西, 2005a; 2005b) を用い, SPSS Advanced Models に含まれる VARCOMP を使って分析した。[2]

今回の研究では，表 3 の 7 つの要因が測定における誤差の要因となりえる。割合に注目して見てみると，受験者の要因（60.1%）は，測定における目的でもあるため（受験者によって評価が違う), 当然のことである (山森, 2004)。次に大きな分散成分を占めているのは，評定者の要因（13.4%）であり，評定者によって，評価の厳しさに差があったことを表している。また，受験者×評定者の要因（7.2%）は，受験者によって，評価者 2 人の評価に違いがあったことを示しており，受験者×項目の要因（5.7%）は，受験者によって，評価の項目ごとに評価に違いがあったことがわかる。このように，一般化可能性研究（G-study）では，変動要因ごとに分散成分の検討が可能である。継続的に実施するようなテストの場合には，評価において改善すべき点を明らかにすることができる。

表 3　分散成分の推定値と割合

	分散成分	
変動要因	推定値	割合（百分率）
受験者（*p*）	2.43	60.1%
評定者（*r*）	0.54	13.4%
項目（*i*）	0.15	3.6%
受験者×評定者（*pr*）	0.29	7.2%
受験者×項目（*pi*）	0.23	5.7%
評定者×項目（*ri*）	0.03	0.7%
受験者×評定者×項目（*pri*）	0.38	9.4%

3.2.2. 決定研究（Decision study; D-study）

一般化可能性研究で得られた分散成分の推定値をもとに，古典的テスト理論での信頼性係数に相当する，一般化可能性係数（generalizability coefficient）を算出し，具体的に必

[2] 詳細な説明は，http://home.att.ne.jp/banana/yamanishi/JACET05.html を参照。

要な評定者の数や，項目の数をシミュレーションし検討する「決定研究」（Decision study; D-study）を行った。図 1 の計算式（山西, 2005a による）によって得られた一般化可能性係数は，$G = .91$ である。「.80 以上であればその採点の結果は信頼できると解釈する」（山森, 2002, p. 64）とされているため，今回の研究で用いた評価方法は，一般化可能性係数の結果からも信頼性の高いものであったと考えられる。

$$G = \frac{\text{受験者の分散成分}}{\text{受験者の分散成分} + \frac{\text{受験者×項目の分散成分}}{\text{項目の数}} + \frac{\text{受験者×評定者の分散成分}}{\text{評定者の数}} + \frac{\text{受験者×項目×評定者の分散成分}}{\text{項目×評定者の数}}}$$

図 1　一般化可能性係数算出のための計算式

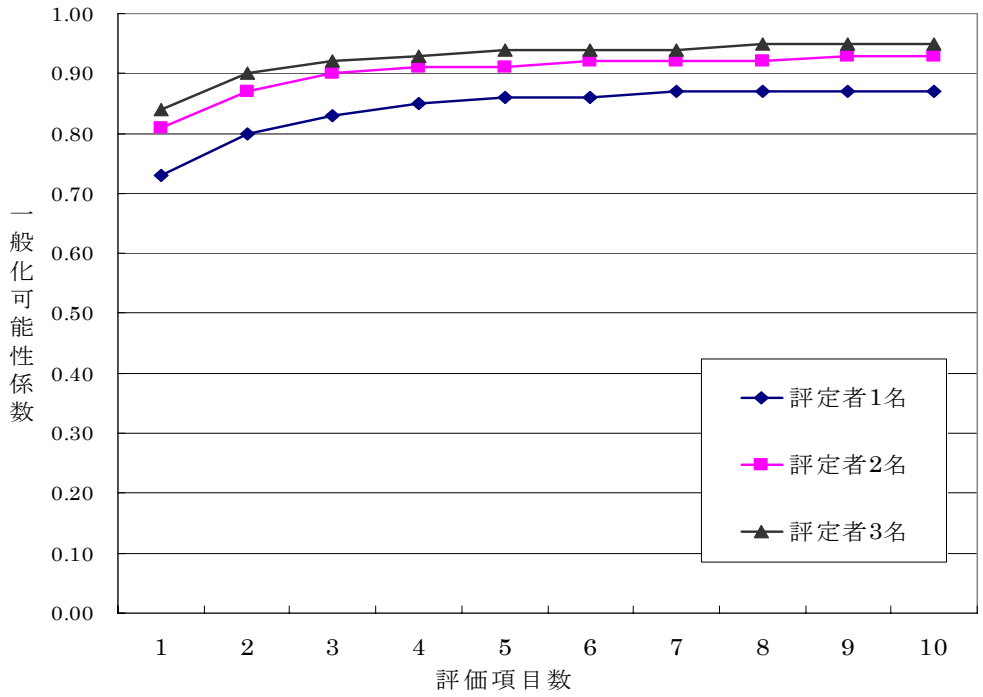

図 2　項目数と評定者数が一般化可能性係数に及ぼす影響（シミュレーション）

図 2 は，項目数と評定者数が一般化可能性係数に及ぼす影響をシミュレーションしたものである。評定者が 1 名の場合にはいくら項目数を多くしても，評定者 2 名の場合の一般化可能性係数には届かないことがわかるが，2 項目以上であれば，一般化可能性係数で .80 以上となるので，1 名でも高い信頼性が確保できる。また，評定者を 3 名にしてもそれほど一般化可能性係数は変わらないので，評定者 2 名で採点を行えば，3 名の場合よりも評定者トレーニングなどの手間が省けると判断できる。評定者 2 名の場合には，評価項目数を 5 つにすれば，項目数が 1 つの場合よりも，一般化可能性係数で .10 ほど上昇する。このよ

うに，決定研究（D-study）では，評定者数や項目数をテストがどれだけ high-stakes なものであるのかなどを考慮に入れながら検討することが可能になる。

4. おわりに

今回の研究では，かなり英語能力の限定された学習を対象としたが，このような尺度に関する信頼性・妥当性の検証の場合には，より大きな幅広い能力を持ったサンプルで実施していくべきである。今回の研究では対象の質的な制限があったものの，得られた結果をまとめると，今回対象としたレベルの学習者の自由英作文を評価する際には，総合的評価・分析的評価のどちらの尺度でもかなりの信頼性が得られることがわかった。テスト実施者の立場から総合的，分析的のどちらかを選ぶとなれば，実用性の観点から総合的評価を選んでも問題なく，また，分析的評価でも同程度の信頼性を持った測定を期待できる。その場合には，本研究で示したように，一般化可能性理論を用いることにより，誤差の要因の特定や，項目数や採点者数の検討を行い，評価の改善が可能になるだろう。

謝　辞

本研究を行うにあたり，愛媛大学の山西博之先生から，たいへん有益なアドバイスをいただきました。末筆ながらここに記して謝意を表します。

文　献

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*, 371-383.

水本篤 (2008).「自由英作文における語彙の統計指標と評定者の総合的評価の関係」『統計数理研究所共同研究レポート』

田中博晃・廣森友人・山西博之・広瀬恵子 (2007).「教育現場に根ざした英語ライティング研究を目指して : 英作文の指導と評価」『大学英語教育学会中国・四国支部研究紀要』4, 55-72.

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

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 (2004).「英会話テストの信頼性の検討－一般化可能性理論―」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (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. 82-91). 東京：大修館.

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.

山西博之 (2005a).「自由英作文評価の改善：評定結果の診断的活用」第 44 回大学英語教育学会全国大会シンポジウム発表資料.

山西博之 (2005b).『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.