分析的評価と全体的評価

― 高校生英語弁論のデータより ―

キーワード：音声言語、評価、相関

内藤　徹

1．はじめに

スピーチの評価は音声言語を対象とするので、文字言語と比べて後戻りできず、また時間的にも制約があり困難な点が多い。そして、その評価法についても、ANALYTIC(AL) EVALUATION（以下 A.E.と略）と HOLISTIC（IMPRESSIONISTIC) EVALUATION（以下 H.E.と略）が考えられ、どちらがどういう場合に効果的なのかは明らかにされていない。

今回、高校生を対象とする英語弁論において、この評価に関するデータを得ることができた。この研究が SPEECH のみならず ORAL COMMUNICATION の評価法にも１つのヒントになれば幸いである。

2．先行文献研究

（１）SPEAKING および SPEECH の評価

測定法として 直接測定法(DIRECT MEASUREMENT)、半直接測定法(SEMI-DIRECT MEASUREMENT)、間接測定法(INDIRECT MEASUREMENT)が考えられる[1]が、本論の中では直接測定法のみを取り扱う。

また、評価者（審査員など）は評価の信頼性を高めるために、目標言語の高い運用能力を持ち、専門的に熟練した者が、明確に定義づけられた採点基準に従って評価を行う必要がある[2]。そして、客観的な評価をすることも大切である。ハロー効果(HALO EFFECT)によってスピーキングやスピーチの信頼性が約３％低下したとの報告もあり[3]、注意を要する点である。

実用性からみた評価の問題点は①採点に要する時間と採点のしやすさ、②実施に要する時間と場所や設備などの実施のしやすさ、の２点が考えられる[4]。①の場合、A.E.は評価が複雑でかつ煩雑となり、時間の面でも問題になるのだが、評価の客観性を高めるためには評価基準を複数設定することが望ましい。しかし、これは短い評価時間、評価のしやすさなどの実用性と反することになり評価の適切さを損なうことにもなる。この点に関しては、H.E.の方が勝っていると言えるが、信頼性の面では問題点が残る。

（２）WRITING の評価からの示唆

Stiggins and Bridgeford(1985)の定義では HOLISTIC SCORING と ANALYTIC SCORING は次の様になる。

> Holistic scoring calls for the reader to rate overall writing proficiency on a single rating scale. Analytical scoring breaks performance down into component parts (e.g organization, wording, ideas) for rating on multiple scales. [5]

Cooper(1977)、Kaczmarek(1980)、Harris(1988)などが述べているように、H.E.は主観的なので信頼性が時々問題になる。従って、Homburg(1984)が言うように INTER-RATER RELIABILITY と INTRA-RATER RELIABILITY に充分注意する必要がある[6]。前者は今回の研究で取り扱う部分であり、後者は訓練などによってカバーできる部分でもある。さらに、彼は Cooper(1977)と同様 H.E. の信頼性を増すために、評価者の訓練が必要であるとも述べている[7]。

そして、Kaczmarek(1980)の次の言及は注目に値する。

> Subjective methods of evaluationg essays work about as well as objective scoring techniques and the former methods are strongly correlated with other measures of ESL proficiency which have independent claims to validity. ..
> The teacher's subjective score may actually be as consistent as a similar evaluation by a small number of trained raters. [8]

その他、Perkins(1983)は、次のように述べている。

> Holistic scoring may have the higest construct validity when overall attained

writing proficiency is the construct to be assessed. [9]

また、Homburg(1984)は、次のように述べている。

Many ESL professionals consider the holistic grading of ESL compositions to be valid. [10]

そして、小谷(1989)や Heaton(1989)も H.E. を肯定的にとらえている。さらに、Heaton(1989)は次のようにも述べている。

Impression marks must be based on impression only. Impression marking is generally found to be much faster than analytic or mechanical marking. [11]

H.E. が実用性において優れているということは、多くの者の意見の一致するところであろう。

最後に Heaton(1989)は初級レベルの TESTER は FLUENCY よりも文法や語彙などの細部にとらわれるが、上級レベルになるに従って FLUENCY を重んずるようになると述べている[12]。これは、初級レベルの評価者は ANALYTIC になりがちであるが、上級レベルになると全体的にとらえられるようになると言うことであろう。すなわち、上級レベルでは HOLISTIC な評価も可能になると考えられる。文字言語と音声言語の違いはあるが、SPEECH においてもこのことが言えるのではないだろうか。

３．実験研究

（１）被験者と評価者および評価方法

１）被験者：高校英語弁論大会の出場者（５番－１５番）

２）評価者：日本人教師４名とＡＬＴ６名

３）評価方法：A.E. (CONTENT, DELIVERY, PRONUNCIATION の各項目)と H.E.の２種類。

（２）仮説

１）英語の運用能力がある者が評価者(EVALUATOR)になる場合、A.E. と H.E. の間には強い相関がある。

２）評価は、特に CONTENT と相関が強い。

（３）分析方法

相関(PEARSON PRODUCT MOMENT COEFFICIENT)

有意差(t-TEST、ONE-WAY ANOVA→RYAN'S METHOD)

分布(STANDARD DEVIATION) 等

（４）結果および分析

☆Table 1（１０名の評価者の評価：N=11(以下同様)）

	A	B	C	D	E	F	G	H	I	J
TOTAL	302	325	314	326	346	333	375	264	354	351
MEAN	27.5	29.5	28.5	29.6	31.5	30.3	34.1	24.0	32.2	31.9
SD	2.40	3.34	2.64	3.08	3.09	3.47	3.56	3.93	2.40	1.68
MAX	32	34	33	34	36	36	39	33	37	35
MIN	23	24	25	24	27	23	30	19	27	30
RANGE	9	10	8	10	9	13	9	14	10	5

NOTES:
A - E = A.E.
F - J = H.E.

A,B,F,G=日本人教員
C,D,E,H.I,J=ALTs

☆Table 2 (Table 1 の分散分析およびライアンの法)

***** ANOVA *****

FACTOR	SS	DF	MS	F
SSB	796.719	9	88.524	8.821
SSW	1003.620	100	10.036	
TOTAL	1800.340	109		

SIGNIFICANCE OF F
*** P<0.001

***** RYAN'S METHOD *****
LOCATION OF SIGNIFICANCE

	A	B	C	D	E	F	G	H	I
B									
C									
D									
E									
F									
G	*	*	*	*					
H	=	*	*	*	*	*	*	*	
I	*			=	=	=	*		
J	*						=	*	

* = P<0.05
= = P<0.001

☆Table 3（１０名の評価点の相関行列）

	A	B	C	D	E	F	G	H	I	J
A		.82**	.44	.83**	.69*	.72*	.64*	.44	.62*	.81*
B			.11	.88***	.78**	.91***	.69*	.19	.38	.61*
C				.17	.37	.14	.24	.58+	.37	.59+
D					.68*	.81**	.75**	.15	60+	.68*
E						.64*	.78**	.55+	.41	.66*
F							.61*	.11	.32	.57+
G								18	.37	.70*
H									.65*	55+
I										.59+

***	P<0.001
**	P<0.01
*	P<0.05
---------	--
+	P<0.1

有意相関の数

A	B	C	D	E	F	G	H	I	J
8	7	0	7	7	6	7	1	2	6

A.E. = 29　　H.E. = 24

☆Table 4 (A.E.の平均点とH.E.の平均点、t-TEST　相関)

	A.E.	H.E.	t-TEST	相関係数
TOTAL	322.6	335.4	t=1.166	
MEAN	29.3	30.5	df=20	0.89*** (P<0.001)
SD	2.38	2.21	P< 0.3	
MAX	33.0	35.8	NO SIGNIFICANCE	
MIN	26.2	28.0		
RANGE	6.8	7.8		

☆Table 5 (A.E.とH.E.の日本人とＡＬＴの比較)

	A.E.		H.E.	
	日本人(1)	ＡＬＴ(2)	日本人(3)	ＡＬＴ(4)
TOTAL	313.5	328.7	354.0	323.0
MEAN	28.5	29.9	32.2	29.4
SD	2.74	2.30	3.13	2.32
MAX	33.0	33.3	37.0	35.0
MIN	24.5	27.0	27.0	26.3
RANGE	8.5	6.3	10.0	8.7

☆Table 6（Table 5 の分散分析およびライアンの法）

***** ANOVA *****

FACTOR	SS	DF	MS	F
SSB	82.055	3	27.352	3.619
SSW	302.280	40	7.557	
TOTAL	384.335	43		

SIGNIFICANCE OF F

* P< 0.05

***** RYAN'S METHOD *****

LOCATION OF SIGNIFICANCE

	1	2	3
2			
3	*		
4			

* = P<0.05

☆Table 7 (A.E.の日本人とＡＬＴ、H.E.の日本人とＡＬＴの評価点の相関行列)

	1	2	3	4
1		.85***	.87***	.53+
2			.77**	.71*
3				.40

***	P<0.001
**	P<0.01
*	P<0.05

+	P<0.1

☆Table 8 (A.E.平均(a)、H.E.平均(b)、CONTENT合計(c)、DELIVERY合計(d)、
PRONUNCIATION合計(e)：満点を全て４０点に換算) [c X 0.4, d X 0.8, e X 0.8]

	a	b	c	d	e
TOTAL	322.6	335.4	347.2	312 0	291.2
MEAN	29.3	30.5	31.6	28.4	26.5
SD	2.38	2.21	1.92	4.03	2.96
MAX	33.0	35.8	34.4	34.4	31.2
MIN	26.2	28.0	29.2	22.4	23.2
RANGE	6.8	7.8	5.2	12.0	8.0

☆Table 9 (Table 8 の分散分析およびライアンの法)

***** ANOVA *****

FACTOR	SS	DF	MS	F
SSB	169.094	4	42.273	4.891
SSW	432.136	50	8.643	
TOTAL	601.230	54		

SIGNIFICANCE OF F

P< 0.003 **

***** RYAN'S METHOD *****

LOCATION OF SIGNIFICANCE

	a	b	c	d
b				
c				
d				
e	*	*		

* = P<0.05

* = P<0.01

☆Table 10 (A.E.平均(a)、H.E.平均(b)、CONTENT合計(c)、DELIVERY合計(d)、
PRONUNCIATION合計(e)の評価点の相関行列)

	a	b	c	d	e
a		.89***	.93***	.87***	.89***
b			.83**	.75**	.76**
c				.78**	.71*
d					.66*

*** P<0.001

** P<0.01

* P<0.05

（５）考察

標準偏差は 1.68-3.93 と分布に幅が見られる。最高点と最低点の幅は 5-14 とやはり幅がある。(Table 1)分散分析によれば１０人の評価点の平均には 0.1% 水準で有意差があり、ライアンの法では G と H が他の評価者と 5% 水準での有意な点差が多い。そして、標準偏差は H が 3.93 で最も大きく、次に G の 3.50　F の 3.47 となり、また逆に J の 1.68 が最も小さく、次いで I の 2.40 となっている。また、H は最高点と最低点の幅も 14 と最大であるが、J は 5 で最小である。従って、H.E.の方に評価者間の幅の大きさが見られる。(Table 2) また、H は評価点の平均が 24.0 で最も厳しく、続いて A は 27.5 である。英語学の専門家は評点が厳しくなる傾向があると言われているが、この場合 A も H も英語学の専門である。相関に関しては、A.E. の方が H.E. よりも有意な相関の数がやや多い。[A.E. 29 個、H.E. 24 個] しかし、C , H , I を除けば、全体的にお互いの相関はかなり強い。[有意相関の数 6 以上] (Table 3)

A.E. の平均点と H.E. の平均点は、有意差はない(P<0.3)。そして、両方の相関は .89*** で 0.1% 水準の相関があり、かなり強い相関と言える。(Table 4)

そして、A.E. と H.E. の日本人とＡＬＴを比較すると(Table 5)、全体的に 5% 水準で有意差があり、ライアンの法では A.E. の日本人と H.E. の日本人に 5% 水準で有意差があるが、その他は有意差は見られない。(Table 6)　相関については、H.E. のＡＬＴは A.E. や H.E. の日本人とは相関はあまり強いとは言えないが、その他は有意な相関がある。従って、日本人の方が A.E. と H.E. の相関がより強いと言える。(Table 7)

最後に、A.E. の３つの項目と A.E. と H.E. の評価の比較である。CONTENT, DELIVERY, PRONUNCIATION の項目の中で DELIVERY が一番標準偏差が大きく(4.03)、評価点にバラつきがあることを示している。(Table 8)　分散分析では、全体的に 0.3% 水準で有意差があり、ライアンの法では PRONUNCIATION は H.E. の平均と 5% 水準で有意差があり、特に CONTENT とは 1% 水準で有意差がある。発音は厳しく評価されているようである。(Table 9)　そして、相関に関しては全てにおいて有意な相関があり、各々かなり強い相関と言える。H.E. の平均とは CONTENT が 0.83 **で最も強く、続いて PRONUNCIATION の 0.76**、DELIVERY の 0.75**となっている。ただし、A.E. の３つの

項目は A.E. の平均とは内部相関であるので高くはなるが、中でも CONTENT とは .93***と最も強い。(Table 10)

それでは、仮説の検証である。A.E. と H.E. は有意な評価点の差はなく、相関は 0.1% 水準で強い相関(0.89***)が見られる。従って、仮説１）は証明されたことになる。英語の専門家が評価をすればどちらを用いてもあまり変わらないと言えよう。［前年度の調査でも双方に 0.82*** と言う強い相関関係が見られた。N=35］ そして、A.E., H.E. とも A.E. の１項目である CONTENT とは 0.1% と 0.5% 水準でかなり強い相関(0.93***, 0.83**)があり、項目の中では最も高い係数である。従って、仮説２）も証明されたことになる。

その他、英語学の専門家は評価点が厳しくなるという松居(1970)の研究がある[13] が、今回のデータでもそれがある程度裏づけられている。

4．おわりに

今回の研究の中で２つの仮説が証明されたが、評価者が英語の専門家であれば H.E. でも充分にその効果を上げることができると言えよう。H.E. は PRACTICALITY の面で優れているが、VALIDITY や RELIABILITY が下がらないよう充分に注意をする必要がある。従って、今まで見てきたように H.E. はバラつきが出やすい傾向もあるので、評価にあまり慣れていない評価者には A.E. の方が向いているのかも知れない。

なお、今回の SPEECH CONTEST においてデータを提供して下さった英語の先生方に、お礼を申し上げる次第である。

（福井県立　鯖江高等学校）

NOTES

1 小川芳男 『英語教授法辞典』, 三省堂, 1982. 595-598.
2 Valette, R.M. Modern Language Testing. 2nd Ed. New York: Harcourt Brace Jovanovich, Inc., 1977. 157-161.
3 Yorozuya, Ryuichi and J.W. Oller, Jr. "Oral Proficiency Scales: Construct Validity and the Halo Effect," Language Learning, 30,1, 1980. 135-153.
4 福井　保 「英語テストとその形式(5)音声テスト－話すことのテスト」『現代英語教育８月号』, 1969. 14.
5 Stiggins, R.J. and N.J. Bridgeford "An analysis of published tests of writing proficiency" Educational Measurement: Issues and Practices, 1983. 26.
6 Homburg, T.J. "Holistic Evaluation of ESL Compositions: Can It Be Validated Objectively?" TESOL Quarterly, 18(1), 1984. 87.
7 ------ 1984. 102.
8 Kaczmarek, C.M. "Scoring and Rating Essay Tasks" in J.W. Oller, Jr. and K. Perkins(eds), Research in language testing. Rowley, Mass.: Newbury House Publishers, Inc., 1980. 151-157.
9 Perkins, K. "On the Use of Composition Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability. TESOL Quarterly, 17(4): 1983. 652.
10 Homburg, T.J. 1984. 87.
11 Heaton, J.B. Writing English Language Tests (New Edition) Harlow, Essex: Longman Group Ltd., 1989. 147.
12 ----- 1989. 148.
13 松居司 「Oral Production のテスト問題作成と評価という観点から」『現代英語教育２月号』 研究社出版, 1970. 17.

その他の参考文献

青木昭六　『英語の評価論』大修館書店　1985.
岡　秀夫　『英語のスピーキング』大修館書店　1984.
沖原勝昭　『英語のライティング』大修館書店　1985.
吉田一衛　『英語のリスニング』大修館書店　1984.
上野景理　A Comparative Study of Holistic and Analytic Evaluations of EFL Compositions Written by Japanese High School Students, M.A. Thesis Joetsu University of Education, 1992.
Hatch, Evelyn and Hossein Farhady Research Design and Statistics for Applied Linguistics, Newbury House Publishers, Inc., 1982.
Larsen-Freeman, Diane and Michael H. Long An Introduction to Second Language Acquisition Research, Longman Group Ltd., 1992.
Underhill, Nic Testing Spoken Language, Cambridge University Press, 1989.