

## クローズテストにおけるテキストと削除変数について

広島大学 西田 正

### はじめに

過去十数年におよぶクローズテスト(以下 CT)の評価は大きく二分されている。好意的な意見として、CTは言語運用能力を総合的に測定できる(Oller 1973, 1976), TOEFL などの英語標準テストや大学での配置テストに有効である(Oller & Conrad 1971, Stubbs & Tucker 1974, Hinofotis 1980), 読解テストになりうる(Anderson 1972, Ibe 1975, Wainman 1979), 英語以外の外国語テストにも適用できる(Caulfield & Smith 1981) などがある。しかし、その一方で、CTに採用するテキスト、語の削除間隔、採点法を変えると、テストの信頼性と妥当性が変わるという批判が出ている(Alderson 1979a, b, 1980, 1983, Klein-Braley 1983)。

本研究の目的は、この指摘に沿って、CTの材料であるテキストの内容と語の削除間隔がCTの信頼性と妥当性にどのような影響を与えるかを考察することにある。具体的にはテキストの親密度(familiar / unfamiliar passage)と削除間隔(6th / 10th)が異なるCTを大学生に実施し、同テストの結果と英語標準テスト(CELT)及び書取りテストの結果との相関を求めて、CTの信頼性と妥当性を検討する。

### 1. テキストの親密度

CTの作成に当たってテキストの難易度が受験者のレベルに合っているかを考えることが多い。一般的には、テキスト自体が難解すぎるCTでは、得点が低い平均点の周りに、易しすぎる場合は逆に高い平均点の周りに、それぞれ集まり結果的にテストの判別能力が低下すると予想できるからである。Mullen (1979)は、易しいテキストのCTを文脈に合う複元語を正答と見なす方法で採点すると、TOEFLなどの外部基準テストの成績を最も良く予測でき、CTの妥当性が高くなると述べている。一方、Oller (1972)は、難しいテキストがproficiency testとの相関が高いと報告している。このような報告を見る限りはテキストの難易度がテスト結果に与える影響は一定していないが、ここで注目したい点は、両報告を含めCTの実験ではテキストの難易度はreadability公式に基づく統語的複雑さの程度を意味し、テキストの内容や話題にはほとんど実質的な検討がされていないことである。

CTを作成する際 passageの内容は無視できない。その内容が受験者の知的レベルに合わないとか、受験者の文化的背景から離れている時は、テストが受験者に不利に働く(Johnson 1982)。Oller (1979)が一部の受験者のみが理解できるような知識を要求するテキストは避けるべきだと指摘するのもCTに関わる内容上のbiasを認めるからであろう。テキストの内容は受身者にとって身近なものから全く疎遠なものまで幅がある。この親密度がCTの得点、信頼性、妥当性に影響すると仮定する理由は次の2点である。

(1)  
第1の理由はschema理論から裏付けである。テキストの親密度は内容に対応するschemaの量と質の問題である。後述する「野球」と「クリケット」のCTのように、読み手が野球のテキストに親近感を覚えるのは野球についてのschema、あるいは先行知識を豊富に持っているからである。

逆に、クリケットについてのschemaが不足していれば、テキストの英語の統語的理解は可能であっても、内容の理解は進まず、この話題は疎遠なままである。第2の理由は読解モデルからの示唆である。読解は上昇型処理 (bottom-up) と下降型処理 (top-down) の相互作用である (Adams 1982)。そして、この相互作用の成否を決定する程の影響を持つ要因は、読み手が獲得している内容についての schema である。身近な内容の読みが進むのは、読みの推理解作業において、体系的な知識である schema が下降型処理を押し進め、かつ、上昇型処理をも円滑にする結果、読解全体がはかどる為である。このような CT に含まれる読みの行為の内実から判断すれば、テキストの親密度は CT の結果に影響を与えると思われる。つまり、親密度の高いテキストの CT は得点が高くなり、妥当性と信頼性も高くなる。

## 2. 語の削除と採点法

削除方法には、テキスト内の n 番目の語を削除する random deletion と削除語を意図的に前以て決めておく rational deletion がある。間隔が問題となるのは前者の場合である。Alderson (1983) は、削除間隔が CT の妥当性に大きく影響すると指摘している。同報告によれば、易しいテキストの間隔を 6 語から 8 語に変えると、標準テストとの相関が 0.59 から 0.70 に、難易度が中程度のテキストを 10 語から 6 語に間隔を狭くすると相関が 0.57 から 0.86 に上昇した。

採点法には、原典通りの復元語のみを正答と認める Exact Scoring Method (以下 EXACT)、文脈に適合する復元語を正答とする Semantically Acceptable Scoring Method (以下 SEMAC)、母国語話者の反応を基に受験者の素点を換算する Clozentropy Scoring Method、多肢選択の解答を準備する M-C 法などがある。Brown (1980) は妥当性、信頼性、項目判別力などの基準からこの4つの採点法を比較し、SEMAC が最も優れ EXACT がこれに続くと判断している。SEMAC の EXACT に対する優位は Mullen (1979)、Alderson (1979b)、Oller (1972) などでも報告されている。

### 3. 1. テキストと削除変数に関する実験

#### 実験 1

#### 目的

CT のテキストの親密度が同テストの信頼性と妥当性に影響を与えるか否かを検証する。

#### 方法

- 1) 材料 *Everyman's Encyclopedia* (1967) の野球とクリケットの説明文を一部抜粋し (約 480 語)、第 2 番目の文頭から 6 語おきに語を削除し 77 項目から成る CT-B (野球) と CT-C (クリケット) を作成した。各テスト用紙に野球とクリケットのタイトルと選手の配置図を載せた。テキストの内容が受験者にとって身近である程度を調べるため 5 point scale の親密度指数 (Familiarity Index 以下 FI) を添えた<sup>(2)</sup>。外部テストとして CELT と書取りテスト<sup>(4)</sup> を利用した<sup>(5)</sup>。
- 2) 受験者 広島大学 1・2 年生 44 名
- 3) 手順 事前に CT の予備練習を約 10 分間全員に与えた後、CT-B と CT-C の用紙を配布し、FI に記入してから 2 つのテストに取り掛かるように指示した。テストの順序の得点への効果を相殺する為に、奇数番号の受験者には CT-B から偶数番号の受験者には CT-C からそれぞれ解答させた。各テストの所要時間は 40 分であった。書取りテストではテープの内容を 3 度繰り返し聞く間に問題を完成させた。実施期日は昭和 59 年 6 月下旬であった。CT の採点は EXACT と SEMAC を併用し、同年 4 月下旬に実施済みの CELT 及び今回の書取りテストの得点と CT の得点との相関を見た。

結果 CELT の各下位テスト, CELT の総合点, CT, 書取りテストの成績と CT への反応数<sup>(6)</sup> は表 1 に, CT の信頼性 (KR 21), 親密度, テスト間の相関は表 2 に示す。

Table 1.

Results of CELT and Cloze Tests (N=44)

Test	$\bar{X}$	SD
L-A	40.45	8.70
V-A	37.34	7.82
S-A	59.50	9.60
Total	137.30	19.04
B-E	16.59	7.13
B-S	20.00	8.29
B-R	45.36	16.12
C-E	5.89	4.06
C-S	6.39	4.22
C-R	25.41	16.89
Dic.	21.20	5.29

B=Baseball C=Cricket E=Exact Word Scoring  
S=Semantically Acceptable Scoring R=Response  
Dic.=Dictation

Table 2.

Correlation of Cloze Tests Scored by Exact Word and Semantically Acceptable Scorings with CELT (N=44)

	Test	L-A	V-A	S-A	Total	Dic.	R	EXACT
Baseball	EXACT (KR21=0.74)	0.53***	NS	0.34*	0.52***	0.57***	0.76***	---
	SEMAC (KR21=0.79)	0.51***	NS	0.36*	0.53***	0.53***	0.76***	0.98***
Cricket	EXACT (KR21=0.68)	NS	NS	0.38**	0.29*	0.39**	0.73***	---
	SEMAC (KR21=0.68)	NS	NS	0.40**	0.31*	0.40**	0.75***	0.98***

FI=Familiarity Index NS=Not significant at the .05 level \*\*\*=P < .001 \*\*=P < .01 \*=P < .05

上記の 2 表から次のことが指摘できる。① CT-B は CT-C よりも身近なテキストである ( FI の平均 2.7 vs 1.0 )。②いずれの採点法によっても CT-B が CT-C よりも得点が高い ( B-E > C-E  $t_0 = 11.70$  P < .001, B-S > C-S  $t_0 = 12.66$  P < .001 )。③同一の CT を比較すると SEMAC の得点が高い ( B-S > B-E  $t_0 = 11.84$  P < .001, C-S > C-E  $t_0 = 3.91$  P < .001 )。④ CT-B は CT-C よりも反応数が多い ( B-R > C-R  $t_0 = 10.74$  P < .001 )。⑤ CT-B は CT-C よりも文法と語彙を除くすべての外部テストとの相関が高い。⑥ CT-B は書取り, 聴取りテスト, CELT の総合点との相関が高く, CT-C は書取りと文法テストとの相関が比較的高い。⑦信頼性係数では野球と SEMAC の組み合わせが最も高く ( 0.79 ), 野球と EXACT が続き ( 0.74 ), クリケットは EXACT, SEMAC とともに 0.68 であった。

以上の結果から, テキストの親密度の高い CT が CELT と書取りテストとの相関が高く, 併存的妥当性と信頼性も高いことがわかる。また採点法では SEMAC と EXACT との相関は 0.98 であったが, 外部テストとの相関と信頼性係数からして, SEMAC の方が優れていた。

## 実験 2

目的 テキストの親密度と新たに加えた削除間隔が CT の信頼性と妥当性に及ぼす効果を検証

する。

方法

- 1) 材料 実験1の反省から統語的に単純な*Encyclopedia Americana* (1958)より, 野球とクリケットの説明文を一部抜粋し(約420~630語), 各テキストから6語おきと10語おきに原文中の語を削除し, 60項目からなるCT B6(野球・6語), B10(野球・10語), C6(クリケット・6語), C10(クリケット・10語)の4種類を作成した。各テスト用紙にタイトル, 図, FIを添えた点は実験1と同様である<sup>(7)</sup>。また, 書取りテストとCELTも前回と同じ内容である。
- 2) 受験者 広島大学 1・2年生 160名
- 3) 手順 無作為に抽出した受験者を4つのグループに分け, 各グループ(40名)に1種類のCTを実施した。所要時間は40分。書取りテストの実施は実験1に準じた。実施期日は昭和59年9月上旬であった。

結果 各グループのCELT, 書取りテスト, CTの得点, 親密度及び反応数は表3に, CTの得点を親密度と削除間隔の変数から分散分析した結果は表4と表5に, 各テスト間の相関は表6に, それぞれ示した。

Table 3.

Basic Descriptive Statistics for Cloze Tests, CELT Subtests, Dictation, and Responses (N=40 for each group)				
	B 06 (FI=2.9)	B 10 (FI=2.5)	C 06 (FI=1.0)	C 10 (FI=1.0)
L-A	$\bar{X}=42.45$ SD= 9.42	$\bar{X}=38.10$ SD= 8.57	$\bar{X}=42.35$ SD=11.48	$\bar{X}=40.45$ SD= 8.75
V-A	$\bar{X}=39.58$ SD=10.80	$\bar{X}=40.25$ SD=11.65	$\bar{X}=40.83$ SD= 8.80	$\bar{X}=42.10$ SD= 9.72
S-A	$\bar{X}=61.05$ SD=12.03	$\bar{X}=60.23$ SD=10.58	$\bar{X}=59.83$ SD=11.16	$\bar{X}=60.75$ SD=13.32
T-L	$\bar{X}=143.43$ SD= 25.89	$\bar{X}=138.25$ SD= 23.44	$\bar{X}=143.00$ SD= 24.41	$\bar{X}=143.23$ SD= 24.93
Dic.	$\bar{X}=21.83$ SD= 7.31	$\bar{X}=20.13$ SD= 7.42	$\bar{X}=19.83$ SD= 6.00	$\bar{X}=19.43$ SD= 7.14
Res.	$\bar{X}=45.70$ SD= 9.55	$\bar{X}=41.20$ SD=10.52	$\bar{X}=31.73$ SD=14.28	$\bar{X}=37.30$ SD=15.75
Ex.	$\bar{X}=17.60$ SD= 6.18	$\bar{X}=12.56$ SD= 6.74	$\bar{X}= 5.35$ SD= 3.60	$\bar{X}=11.10$ SD= 5.20
Sm.	$\bar{X}=20.88$ SD= 6.95	$\bar{X}=15.35$ SD= 7.42	$\bar{X}= 7.63$ SD= 5.07	$\bar{X}=12.53$ SD= 6.17

B=Baseball text    C=Cricket text    Ex.=Exact word scoring  
Sm.=Semantically acceptable scoring    06=6th deletion    10=10th deletion  
Res.=Response

Table 4.

ANOVA for gains in EXACT related to Text Familiarity and Deletion				
Source	S.S.	D.F.	M.S.	FO
A	1876.9000	1	1876.9000	60.7941**
B	4.9000	1	4.9000	0.1587
AxB	1166.4000	1	1166.4000	37.7805**
E	4816.2000	156	30.8731	
T	7864.4000	159		

Source    A=Text Familiarity    B= Deletion

Table 5.

ANOVA for gains in SEMAC related to Text Familiarity and Deletion				
Source	S.S	D.F.	M.S.	FO
A	2584.0563	1	2584.0563	61.8572**
B	3.9063	1	3.9063	0.0935
AxB	1086.8063	1	1086.8063	26.0160**
E	6516.8250	156		
T	10191.5938	159		

Table 6.

Correlation of Cloze Tests Scored by Exact word and Sementically Acceptable Procedures, with CELT, Dictation, and Response

	Test	L-A	V-A	S-A	Total	Dic.	R.
B 06	EXACT (KR21=0.69)	NS	NS	0.37*	0.41**	NS	0.59***
	SEMAC (KR21=0.73)	0.35*	NS	0.38*	0.42**	NS	0.63***
B 10	EXACT (KR21=0.79)	NS	0.35*	0.47***	0.45**	0.51***	0.78***
	SEMAC (KR21=0.81)	NS	0.42**	0.48***	0.49***	0.54***	0.78***
C 06	EXACT (KR21=0.63)	0.50***	NS	0.43**	0.50***	0.51***	0.76***
	SEMAC (KR21=0.75)	0.53***	0.32*	0.47***	0.58***	0.58***	0.72***
C 10	EXACT (KR21=0.68)	0.35*	0.34*	0.59***	0.58***	0.47***	0.78***
	SEMAC (KR21=0.75)	0.34*	0.34*	0.62***	0.58***	0.44***	0.81***

\*\*\*=P<.001      \*\*=P<.01      \*=P<.05      NS=not significant at the .05 level

以上の表から次のことが判明する。①テキストの親密度は「野球」が高い (FI B 06 = 2.9, B 10 = 2.5, C 06 = C 10 = 1.0)。②各グループ間には英語力の有意差がなく均一な英語力を持つグループである〔聴取りテスト  $F_0(3.156) = 1.80 P > .05$ , 語纂テスト  $F_0(3.156) = 0.25$ , 文法テスト  $F_0(3.156) = 0.99$ , CELT 総合点  $F_0(3.156) = 0.41$ , 書取りテスト  $F_0(3.156) = 0.91$ 〕。③反応数は CT-B と C 間に有意差があり〔 $F_0(3.156) = 8.61 P < .01$ 〕, Post Hoc. の結果 B 06 が最も反応数が多く B 10, C 10, C 06 の順であった。④ EXACT と SEMAC のいずれの方法で採点しても CT 間に有意差があり〔 $F_0(3.156) = 32.91 P < .01$ ,  $F_0(3.156) = 29.33 P < .01$ 〕, B 06 が最も得点が高く, 次に B 10 が高く, C 06 が最低であったが, B 10 と C 10 の間には有意差がない。⑤ CT の得点にはテキストの親密度の主効果があり〔 $F_0(1.156) = 60.71 P < .01$ ,  $F_0(1.156) = 61.85 P < .01$ 〕, 削除変数と親密度と交互作用が現われた〔 $F_0(1.156) = 37.78 P < .01$ ,  $F_0(1.156) = 26.01 P < .01$ 〕。従って, 親密度の高い CT は 6 語おきの削除間隔の場合に限って, 親密度の低い CT よりも容易である<sup>(a)</sup>。⑥各外部テストのとの相関では得点最も高い B 06 が最も低く, 最も得点低い C 06 が最も高く, C 10, B 10 の順序である。C 06 は聴取りテスト, 書取りテスト, CELT の総合点と 0.50 ~ 0.58 の相関があり, B 06 は CEL の総合点と 0.41 ~ 0.42, 文法テストとは 0.37 ~ 0.38 の相関があるが書取りテストと語纂テストとの相関は有意でなく, 同テストを EXACT で採点すると聴取りテストとも有意な相関がない。⑦ KR 21 の信頼性係数は B 10 が最も高く, C 10, B 10, C 06 の順序である。⑧ SEMAC は EXACT よりも外部テストとの相関が強く信頼性も高い。

### 3.2 実験1と2に関する考察

実験1と2を通じて、親密度の高いテキストのCTでは得点と反応が多くなったが、テストの信頼性と妥当性についてはかならずしも一致した結果が出ていない。実験1では、親密度の高いCTが外部基準テストと高い相関関係にありテストの妥当性が高くKR21から判断できる信頼性も高かった。しかし、実験2では、逆に、親密度の低いCTが妥当性と信頼性が高くなった。削除変数に関しては、10語間隔のCTが併存的妥当性が高い。採点法は2つの実験結果ともSEMACの優位さを立証しておりOller(1972)やAlderson(1979b)の報告と同様である。

テキストの親密度がCT得点に影響しテストの信頼性と妥当性を変えることは明らかになったが、何故一致した影響力が見られなかったのであろうか。この問題は受験者の英語学力の特性と関連があるように思われる。実験2で検討したように、CT得点は受験者の英語力の差ではなく、テキストの親密度が関与しているが、根本的な問題はCTに含まれる読解過程におけるtop-down処理にある。前述したように読解はbottom-up処理とtop-down処理のsynchronyである(Haberlandt 1982)。しかし、日本人大学生が英語を読む場合bottom-up処理が自動化する程発達しておらず、努力の大半は語と語の関係から意味を取る作業に注がれ、内容全体から意味を取るというtop-down処理に手が届かない。従って、自分にとって身近でschemaが豊富なテキストに対しても、また疎遠な内容に対しても同じようなstrategyを用いており、テキストの親密度が読みの活動に直接効果を及ぼさないのである。Carrell(1983)も、ESLの大学生はこのtop-down処理能力が欠けていると指摘しているので、本実験の受験者のようなEFLの大学生にはさらにこの傾向が強いと思われる。

CTと各テストとの相関については、実験1と2において語果テストとの相関が最も低く、続いて聴取りテストが低い。比較的高い相関があったテストは文法、書取りとCELTの総合であった。しかし、CTからは文法力の38.4%、書取り力及び総合点で示された英語力の33.6%が最大限予測できるにすぎず、相関は先行研究に比べて低い<sup>(9)</sup>。また、従来、CTと書取りテストとの相関が他の言語テストよりも高いことからCTは書取りテスト同様integrativeなテストと思われるが、本実験では他の外部テストとの相関に比べて両テスト間に著しく高い相関は検出されなかった。

#### 結 論

CTが教育実践の場で注目されて来た理由のひとつにテストの経済性を挙げることができる。CTは作成が簡単で得点の処理も短時間で完成する。しかし、この経済性の中にこそ未解決な課題が残っている。今後はそのひとつひとつを検証し、結果を積み上げCTの有効性と限界を見きわめる必要がある。

#### 〔注〕

- (1) schema理論と読解については西田(1983)を参照。
- (2) 親密度の尺度は5=非常によく知っている、4=よく知っている、3=普通、2=あまりよく知らない、1=全く知らない、である。実験においてはこの段階づけを口答で説明した。
- (3) クローズテストの実例は紙面の都合上割愛した。
- (4) CELTの各下位テストの得点は100点満点に換算した。
- (5) 書取りテストはVOA News in Cassette(International Communication Incorporated 1984)を利用した。内容は4種類のpassage中の削除した語と語句を書取るテストで50点満点。
- (6) 反応数は正答と誤答を加えた数。
- (7) 実験2の使用したクローズテストも割愛した。

- (8) 文脈が多い場合はテキストの親密度はテストの成績に強く関与しない。このことは文脈の利用の面で興味深い。
- (9) 例えば, Oller (1972)では, クローズテストと文法テストは0.66~0.80, 書取りテストとは0.68~0.85の相関があり, Oller & Conrad (1971)では, 文法テストとは0.58, 書取りテストとは0.82の相関が報告されている。

#### References

- Adams, Marilyn Jager (1982) "Models of Reading," Le Ny, Jean-François and Walter Kintsch (eds.) *Language and Comprehension* (North-Holland Publishing Company) (1982) 193-206.
- Alderson, J. Charles (1979a) "The Cloze Procedure and Proficiency in English as a Foreign Language," *TESOL Q*, 13, 2, 219-227.
- \_\_\_\_\_ (1979b) "Scoring Procedures for Use on Cloze Tests," Yorio, Carlos A., Kyle Perkins, and Jacquelyn Schachter (eds.) *On TESOL '79: The Learner in Focus* (TESOL) (1979) 193-205.
- \_\_\_\_\_ (1980) "Native and Nonnative Speaker Performance on Cloze Tests," *LL*, 30, 1, 59-76.
- \_\_\_\_\_ (1983) "The Cloze Procedure and Proficiency in English as a Foreign Language," Oller, John W. Jr. (ed.) *Issues in Language Testing Research* (Newbury House Publishers, Inc.) (1983) 205-217.
- Anderson, J. (1972) "The Application of Cloze Procedure to English Learned as a Foreign Language in Papua and New Guinea," *ELT*, 27, 1, 66-72.
- Brown, James Dean (1980) "Relative Merits of Four Methods for Scoring Cloze Tests," *MLJ*, 64, 3, 311-317.
- Carrell, Patricia L. (1983) "Three Components of Background Knowledge in Reading Comprehension," *LL*, 33, 2, 183-207.
- Caulfield, Joan and William C. Smith (1981) "The Reduced Redundancy Test and the Cloze Procedure as Measures of Global Language Proficiency," *MLJ*, 65, 1, 54-58.
- Haberlandt, Karl (1982) "Reader Expectations in Text Comprehension," Le Ny, J. F. and W. Kintsch (eds.) *Language and Comprehension* (1982) 239-249.
- Hinofotis, Frances Butler (1980) "Cloze as an Alternative Method of ESL Placement and Proficiency Testing," Oller, John W. Jr. and Kyle Perkins (eds.) *Research in Language Testing* (Newbury House Publishers) (1980) 121-128.
- Ibe, Milagros D. (1975) "A Comparison of Cloze and Multiple Choice Test for Measuring the English Reading Comprehension of Southeast Asian Teachers of English," *RELCJ*, 6, 2, 24-32.
- Johnson, Robert K. (1982) "Questioning Some Assumptions About Cloze Testing," Heaton, J. B. (ed.) *Language Testing* (Modern English Publications Limited) (1982) 59-72.
- Klein-Braley, Christine (1983) "A Cloze Is a Cloze Is a Question," Oller, John W. Jr. (ed.) *Issues in Language Testing Research* (1983) 218-228.
- Mullen, Karen A. (1979) "More on Cloze Tests of Proficiency English as a Second Language," Brière, Eugène J. and Frances Butler Hinofotis (eds.) *Concepts in Language Testing: Some Recent Studies* (TESOL) (1979) 20-32.

西田 正 (1983) 「英語の聴解と読解に及ぼす絵文脈の効果——スキーマの視点」『言語文化研究』  
( 広島大学総合科学部紀要V ) Vo1. 9, 135-150.

Oller, John W. Jr. (1972) "Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language," *MLJ*, 56, 3, 151-158.

\_\_\_\_\_ (1973) "Cloze Tests of Second Language Proficiency and What They Measure," *LL*, 23, 1, 105-118.

\_\_\_\_\_ (1976) "Language Testing," Wardhaugh, Ronald and H. Douglas Brown (eds.) *A Survey of Applied Linguistics* (The University of Michigan Press) (1976) 275-300.

\_\_\_\_\_ (1979) *Language Tests at School* (Longman)

\_\_\_\_\_ and Christine A. Conrad (1971) "The Cloze Technique and ESL Proficiency," *LL*, 21, 2, 183-195.

Stubbs, Joseph Bartow and G. Richard Tucker (1974) "The Cloze Test as a Measure of English Proficiency," *MLJ*, 58, 5-6, 239-241.

Wainman, H. (1979) "Cloze Testing of Second Language Learners," *ELTJ*, 33, 2, 126-132.