教室におけるシャドーイング技術の評価

玉井 健

本研究は，外国語の聴解力向上を目的とした訓練法としてのシャドーイングについて，その評価法を提案し，妥当性，信頼性，実用性を検討するものである。シャドーイングは聞こえてくる入力音声を，ほぼ同時にしかもできるだけ正確に再生する行為をいうが，確立された評価法はまだない。評価対象が音声出力であり，評価基準が入力音声の正確な再生であることを勘案して，まず実質的に評価可能な最小単位である音節レベルでの評価（音節評価法）を試みた。この結果をもとに，より簡便で実用的な評価法としてのチェックポイント法の結果を算出比較し，その妥当性と信頼性及び実用性の検討を行った。音節評価法は信頼性，妥当性に問題はないものの，実用性に問題があり，チェックポイント法はある一定の項目数を確保すれば，一つのの評価法として十分使用可能であることが示唆された。

はじめに

シャドーイングは，通訳のためのリテンション技術向上を目的として一般の通訳養成機関で広く用いられてきたが，近年は LL 教育の普及と共に時事英語教材を教室に導入する方法として，大学その他の教育機関でも積極的に応用されている（西村:1996）。また聴解訓練法としての有用性については，玉井 (1992) のディク

テーションとの比較実験によって明らかにされている。しかし，シャドーイング技術の実体とその評価法についての研究はまだほとんどなされていない。これは次のような理由によると思われる。

第一は，音声という不安定な言語形式を評価対象にする際の扱いにくさ。第二は，通訳養成機関におけるシャドーイングはあくまで訓練法の一つであり，それ自体が最終目標ではなく，そのためにシャドーイング技術の進歩が評価の対象となっていないこと。第三は，シャドーイング技術の実体に対する把握が十分になされていないこと，が考えられる。本稿では，まずシャドーイングの聴解行為における位置づけをし，さらにその評価法について試案を提示し，考察を加えてみたい。

認知行為としてのシャドーイング

Baddeley (1986) の作動記憶モデルをベースに考えると，シャドーイングは高度に認知的な行為であり決して機械的な行為ではないことがわかる。以下に二谷 (1994) が図式化したモデルをさらに簡略化したものを用いて説明する。

注意によって選別的に入った音声情報はパターン認識にかけられ，その一部が作動記憶右側の音声ループにまわされる。Schweickert (1986) によると，この入力音声は何もしなければ約 1.5 秒で消失するが，内語発声 (subvocalization) にかけられる限りは保持される。つまり，音声ループ上では入力情報のうちうまく内語化された部分のみが残り，内語化されなかったものは消失することになる。このことは，作動記憶上で保持できる情報量は 1.5 秒のうちに正確に内語発声できる情報量に従属することを意味し，例えば，有声無声を問わず，早口に音声化できる人はループ上に保持できる情報量が増えるということになる。入力音声を早口に内語化できれば，中枢装置が意味理解のために利用できる情報量が増え，その結果理解が助長されることが考えられる。

この内語発声を意識的に有音声化して行うのがシャドーイングである。シャドーイングでは，入力音声の後追いをしながら正確に反復してゆくが，このリアルタイムでの瞬間的な反復行為が学習者の内語発声技術を高め，結果的により多くの情報を意味理解処理にまわすことが可能になるのだろう。

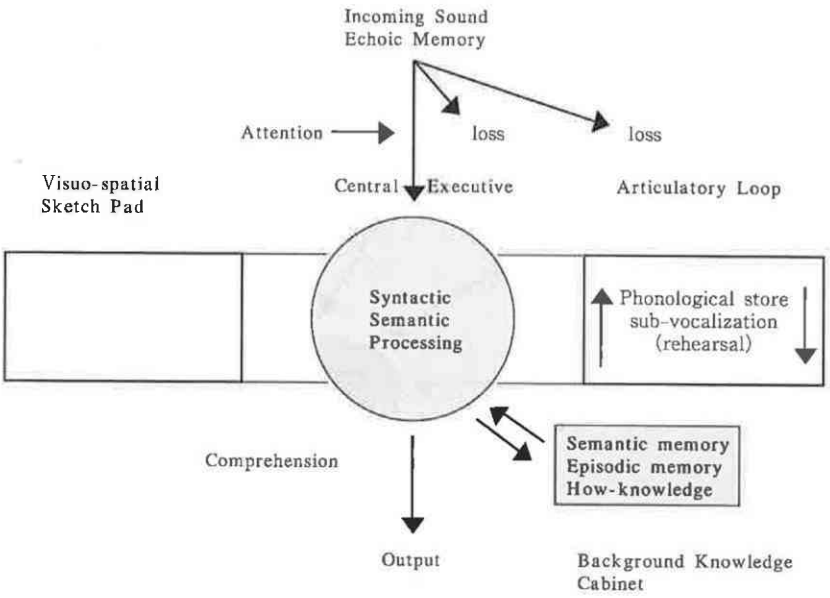

図 1: Baddeley の作動記憶モデル（二谷:1994 を改変）

シャドーイング技術を評価することの意味

シャドーイングの評価はどういった意味合いを持つのだろう。その評価は聴解行為の最終的結果，つまり意味理解の程度を測定するものなのか，あるいはその過程のどこか一部を測定するものなのだろうか。Atkinson & Shiffrin (1968) のモデルでも Baddeley のモデルでも瞬間的な聴解行為はいくつかの段階に分けられている。シャドーイングが入力音声の正確な再現を目的としていて，意味理解の程度を問題にしない点を考えると，少なくとも最終段階ではなくそれ以前のいずれかの段階ということになる。上図において，入力音声が意味理解以前に保持されているのは右側の音声ループ上であることを考えると，仮説的にではあるが，シャドーイングの評価は，上図の音声ループ上に保持されている情報形態の正確さとその量を測るものであると考えてよいかもしれない。最終的な意味理解でないことは，意味の把握そのものをシャドーイングが目的としていないことからも

わかる。図 2 のモデルを参考にしながらこの点を考えてみたい。

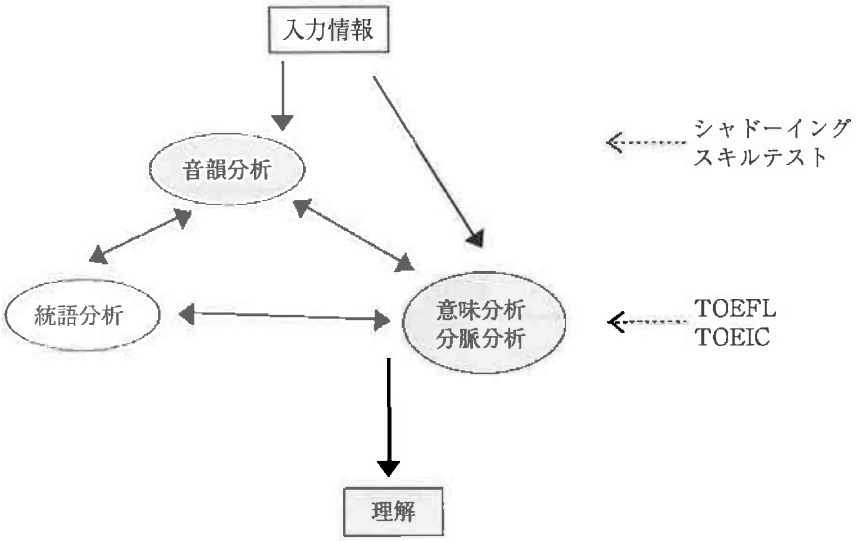

図 2: 理解への流れと測定されるレベル

TOEFL や TOEIC などの総合的な聴解力テストを実施した際に測定器具としてのそれらのテストの測るものは，意味理解の程度である。聞いた会話の全てを正確に聞き取ったり記憶していなくても，全体の大筋をきちんと捉えていれば，内容に関する質問にはかなり答えられる。つまり入力された音声情報の持つ意味を最終的にどれほど正確に再構築できたかが評価のターゲットである。ただし意味理解に至るには，音韻パターンの分析，語彙認識，統語論的分析，語用論的分析あるいは文脈分析等を前段階として経ている（そうでない場合もある）から，そういった様々な能力も含めた総合的な聴解力を測定のターゲットとしている。上図でいえば，分析の最終結果である理解の程度ということになる。

これに対してシャドーイング技術を測るということは，どれだけ正確に入力情報を音声レベルで再生できるかということであるから，必ずしも意味の把握を前提としない。もちろん，統語論的な知識や語彙知識，語用論的知識や文脈の類推，

意味分析における結果が，相互補完的に音韻分析を助けることは考えられるが，基本的にはどれだけ正確に音韻分析ができるかという能力の測定と考えられる。

ただしこれは，学習者がシャドーイングをする際に意味をとらない，ということでは決してない。シャドーイングしながら注意を向ければ意味は追えるし，意味を追わず機械的なリピート作業に徹することも可能である。あくまで，測定器具としてのシャドーイング技術テストのターゲットが，意味理解の程度ではないということである。

筆者の過去の実験 (玉井:1992) では，高校生を被験者にしたシャドーイングスキル・テストと総合的な聴解力テストである SLEP との相関は（$r = 0.285$），大学生を被験者にした場合では，TOEFL との相関が ($r = 0.396$) と低いレベルの値しか確認されなかった。これは一つの言語処理過程においてもいくつもの側面があり，測定するターゲットが違えば結果も異なってくることを補足的に説明すると考えられる。結論として，TOEFL や TOEIC，SLEP などの言語テストのリスニング部門を総合的聴解力テストとして位置づけるならば，シャドーイング技術テストは同じリスニングという領域にありながら，入力音声情報の再構築という異なる側面にスポットを当てていることになる。

シャドーイング技術の測定：全音節評価法とチェックポイント法

上記の議論をふまえ，まずシャドーイング技術について次のように定義した。

> Shadowing is the act or a task of listening in which the learner tracks the heard speech and repeats it as exactly as possible while listening attentively to the incoming information.
>
> （シャドーイングとは聞こえてくるスピーチに対してほぼ同時にあるいは一定の間をおいてそのスピーチと同じ発話を口頭で再生する行為，または聴解訓練法）

またシャドーイング技術を測るテストをシャドーイングスキル・テストと呼ぶことにする。シャドーイングスキル・テストでは評価対象がリアルタイムでの音声であるために，口頭による出力部分を録音し，採点者がテープを聴いて現スピー

チに対する再生率を評価することにした。採点に際しては，どういった単位で分析を行うかを決定しなければならないが，聴いて採点できる最小単位として，音節ごとの正誤を判断することにした。

ここで問題になるのは，採点者が異なった場合に評価に一貫性が保てるのか，つまり採点者間でどのような差，あるいは揺れが起こるのだろうかということである。この点については採点者間信頼度係数を算出することにした。

シャドーイングスキルテストは，入力も音声，出力も音声であるために，文の内容・文法的複雑さ・文の長さ・語彙・読み方（natural / teacher talk など）・スピード等の変化によって難易度が変わる。全ての変数を統制することは不可能であるが，文の難易度については Flesch-Kincaid の Readability Index（読みやすさ指数 RI）[1]を用いて統制を図った。スピードは 1 分間の語数で統制した。概要は次の通り。

材料： 難易度，スピード，読み手，読み方の異なる 3 種類の材料をもとにテープを作成した。英語の難易度は，米国の学校での読みやすさ指数で 5 年生から 10 年生レベルの範囲のものを 3 点採用した。スピードは 1 分間あたり 109 語から 151 語の範囲にとどめた。読み方については統制はしていない。

被験者： 高校生　47 人

表 1: 使用した英文の難易度と 1 分間あたりの語数

タイトル	読みやすさ指数	語数/分	音節数
Eye on The World	RI=4.75(5^{th} grade)	109 words	134 syllables
Cultural Difference	RI=9.57(10^{th} grade)	151 words	237 syllables
Guide Dog	RI=8.95(9^{th} grade)	133 words	170 syllables

[1]従来の readability formula は基本的に，特定の言語要素の特徴（単語の長さ，単語の使用頻度の高低，文の長さなど）が，文章のなかでどのくらいの割合を占めているか，その平均値を調べることによって，文の読解難易度を決定しようとするものである。—単語の使用頻度の高低とは，いくつかの種類の読み物に使用されている単語を集計し，それらを頻度の高い順にリストアップし，作成された語彙リストの中に，任意のある文章に使用されている単語がどの程度の割合で含まれているかを計算することによって，読解難易度算定の基礎の一つとして利用することを意味する）(佐藤：1988)

テスト要領： LL教室でスピーカーから流れるスピーチを3つ順にシャドーイングさせ，出力した音声を録音した。

採点方法： 採点は外国人教師1名と日本人教師2名がスクリプトと照らし合わせながらテープを聴き，自然に再生されていると判断された音節数(541箇所)を計算し，100%に換算するという方法で行われた。スクリプトの形態は次のように音節単位で切られている。

（例）The o-zone lay-er pro-tects us from the sun-'s ra-di-a-tion, but it has a hole in it. A very big hole that is get-ting big-ger eve-ry day. The US passed a law that all o-zone de-story-ing chem-i-cals must stop being used by two thou-sand. Eu-rope has laws too, but eve-ry-one must stop if we want to save the o-zone lay-er. baselineskip16pt

上記の文章の場合は80音節あることになり，そのうちの自然に再生されている音節数が評価点となる。これを音節評価法(The syllable method)と呼ぶことにする。3人の採点結果の平均と標準偏差(表2)，及び採点者間の相関(表3)は次の通り。

音節評価法について妥当性，信頼性，実用性の3点から考えることにする。シャドーイングスキル・テストの目的は，原スピーチがどれほどの正確さをもって再生されているかであるから，評価はテープを聴いてスクリプトと対照させた上で行われることになる。この際，評価単位が小さいほど正確性は保たれることになる。そういう意味では，音節単位で評価することは，入力音声に対する再生力を計測する上で，評価単位を単語よりもさらに小さくできるという点で適切な処置であると考えられる。ただし，それは評価過程を煩雑にする欠点をも併せ持つことにもなる。

また，テープを聴いていると，各文章の出だしの箇所は出力されていないことが多かった。今回は3つの文章を材料にしたが，いずれの文章においても最初の1文は採点の対象にしないほうが良いと思われる。学習者にとって，最初の文に対しては，聴解における内容の予測がほとんど行えないためであろう。表面的妥当性という面からは，採点の煩雑さを別にすれば大きな問題点はないと思われる。

表 2: シャドーイングスキルテスト採点結果（音節評価法）

採点者	$\bar{X}$	SD
R(1)	69.2	10.04
R(2)	63.8	11.06
R(3)	64.3	10.18

表 3: シャドーイングスキルテスト採点者間の相関

	R(1)	R(2)	R(3)
R(1)	1.00		
R(2)	0.870	1.00	
R(3)	0.903	0.845	1.00

注（R(1) ー外国人 R(2), R(3) ー日本人）

信頼性という点ではどうか。採点者相互の平均点の差については t-検定の結果, R(1) > R(2), R(1) > R(3) ($p < 0.01$) というふうに，外国人評価者と日本人評価者との間には平均点に有意な差が見られたが，これは，曖昧な出力（複数の-s の脱落や無声化された-ed の脱落など）に対する採点者の許容度の差であると考えられる。しかし，表 3 の示す通り，3 人の採点者間の相関は高く，外国人教師と日本人教師との間にも差は見られなかった。採点者間信頼度の推定値は 0.872 で高いレベルにあり，当初考えられた採点者間の評価差の問題はないと考えられる。ただし，次のような問題点が指摘された。

1. 出力された音声を採点者が聴いて判断するため，採点者自身が，正答とすべきかそうでないか迷う箇所が多いことが報告されている。また，a game → again, support her → supporter というような同音異義的な誤りについては，本テストは誤りと見なしえない。ただし本テストの性質としての，意味をターゲットとしないという点から見ればこれは問題とはならないとも言える。

2. 採点者の集中力が低下した時に，採点が不安定になることが採点者から報告された。録音された音声を資料として扱うゆえの問題と考えられる。

実用性という面ではどうか。この点から見ると音節評価法は大きな問題を抱えていると言わざるを得ない。全音節を集中して聴くことは採点者に多大な時間的及び体力的な負担を強いることになるし，これは評価そのものを不安定なものにしてしまう。それは同時に日々の教室での評価法としては非実用的ということになる。

チェックポイント法

ここで，音節評価法の信頼性を保ちながら，もっと簡便な方法として，音節ではなく単語レベルの聞き取りにすること，さらにチェックポイントを決めて採点することにより採点の労力を減ずる方法を考えた。具体的には，スクリプトの全単語を5語ごとにチェックするというやり方である。チェックポイント法 (The check point method) と呼ぶことにする。5語間隔とした理由は，スクリプト全文中の各節の平均的な長さが7.8語であり，5語ごとのチェックを行えば，意味単位としての各節のうち必ず1カ所は採点の対象になるだろうと考えたからである。採点用紙のチェックポイントは次のようになる。

（例）The ozone layer protects us from the sun's radiation, but it has a hole in it. A very big hole that is getting bigger every day. The US passed a law that all ozone destroying chemicals must stop being used by two thousand. Europe has laws too, but everyone must stop if we want to save the ozone layer.

日本人採点者 R(2) と R(3) の行った採点項目についてチェックポイント法を用いて再度採点し直してみたところ，平均と標準偏差は表4のようになった。また，音節評価法との相関は表5に示す。

まず信頼性について考えてみる。チェックポイント法の同じ採点者の二つの異なる方法による採点結果の相関は，R(2) において ($r = 0.89$)，R(3) において ($r = 0.87$) と高い相関が見られ，二つの方法間の結果による実質的な差はない

表 4: シャドーイングスキルテスト採点結果（チェックポイント法）

採点者	$\bar{X}$	SD
R (2)	43.1	8.45
R (3)	42.3	8.96

表 5: 音節評価法とチェックポイント法の採点者間の相関

	R(2) 音節評価法	R(3) 音節評価法	R(2) チェックポイント法	R(3) チェックポイント法
R(2) 音節評価法	1.00			
R(3) 音節評価法	0.85	1.00		
R(2) チェックポイント法	0.89	0.80	1.00	
R(3) チェックポイント法	0.76	0.87	0.85	1.00

と判断できる。76 箇所のチェックポイントの均質性については，折半法[2] による検定を行ったところ（$r = 0.89$）という値が得られた。チェックポイントの設定については機能語，内容語にかかわらずランダムに設定した。この点については，内容語，機能語とも外国人学習者には同じくらいむずかしく，その間には有意差はないという Wainman (1979) の報告を参考にした。テスト項目トータルとしての均質性は確保されたようである。

また，音節評価法における採点者相互間の相関と，チェックポイント法における採点者相互間の相関が同一であることは，二つの採点結果の等質性をサポートすると考えられるだろう。音節評価法の信頼性はチェックポイント法においても維持されていると考えられる。

[2]一つのテストについて，奇数番号の問題項目と偶数番号の問題項目の 2 つに折半し，その両者の得点間の相関係数を求めて，それを信頼度係数とする方法。テスト項目の均一性を検定しているので，テストそのものの信頼性の検定ではない。

適正なチェックポイントの数について

音節評価法では 541 音節，チェックポイント法では 76 箇所が採点対象になったが，その数はどれぐらいが適切なのだろうか。教室で日々の評価に用いるならば，評価項目が少なければ少ないほど簡便であるし，かといって少なすぎると測定道具としての信頼性に問題が生じてくる。そこで一人の採点者の採点結果に絞って，全体を構成する 3 つの部門それぞれの結果と全体との比較をしてみた。(表 6)

表 6: 音節評価法の各部門と全体との相関 (採点者 R(2))

	全体 (541 音節)	Eye on the World	Cultural Difference	Guide Dog
全体 (541 音節)	1.00			
Eye on the World (134 音節)	0.84	1.00		
Cultural Difference (237 音節)	0.93	0.69	1.00	
Guide Dog (170 音節)	0.90	0.65	0.76	1.00

各部門とも全体とは高い相関がみられるものの，各部門相互の相関は中程度の相関に下がっている。次にチェックポイント法においても一人の採点者に絞って同じ処理をしてみた。(表 7)

全体との相関は若干下がり，各部門相互の相関にもはっきりとした減少が見られる。各部門のチェックポイント数がそれぞれ 21 カ所，30 カ所，25 カ所であることを考えると，このレベルの数で十分な信頼性が確保できるとは言い難い。今回の結果に限って言うならば，70 カ所程度の項目数を確保すれば音節法の結果に近い評価が得られていた。以上の議論を結論としてまとめてみる。

表 7: チェックポイント法の各部門と音節評価法による結果との相関

	全体 (541 音節)	Eye on the World	Cultural Difference	Guide Dog
全体 (541 音節)	1.00			
Eye on the World (21 箇所)	0.71	1.00		
Cultural Difference (30 箇所)	0.78	0.52	1.00	
Guide Dog (25 箇所)	0.76	0.49	0.64	1.00

結論

本稿ではシャドーイング技術を評価する方法としての音節評価法とチェックポイント法について検討した。どちらも，録音された音声出力を採点者が聴きながら採点するという点では基本的に同じであり，この点から生じる不安定さを解消してはいない。二方法の相違点は，前者が音節単位で全てのテキストに対して採点を行うのに対して，後者は 5 語間隔に定められたポイントについて，単語レベルの採点を行うという点である。この差は，前者が 541 カ所の音節単位を採点対象にしたのに対して，後者は 75 カ所の限定された箇所という形で反映された。

音節評価法は，テキスト全体に対して音節単位での評価を行うという点で，シャドーイング技術を測るテストとしては適正と思われる。採点者相互の信頼度係数 (inter-rater reliability) は非常に高いレベルに維持されており，3 人の採点者がそれぞれ一貫性を持って採点できていたことが伺える。出力された音声を聴いて評価するという不安定要因は完全には拭えないものの，実質的なテストとしての信頼性は十分確保されたと言える。音節評価法の短所は，一定の信頼性を保ちなが

らも，実用面から見ると採点の労が大変であり，必ずしも日々の授業で使用可能な実用的な評価法とは言い難い面である。

チェックポイント法は，音節評価法と比べると，実用性という面ではかなり簡便になっている。信頼性という面については，項目数が30程度では音節評価法の結果との相関も十分でないが，三種類の文章を組み合わせてその合計（76箇所）をとった場合は，音節評価法の結果との間に高い相関が見られた。チェックポイント法の場合はある一定の項目数を確保する必要がある。今回の結果から言えば70項目程度あれば一定の信頼性が確保されると言えるだろう。

ただし，シャドーイングスキル・テストのデザインを考える場合，採点者が聴いて判断するという点は，いずれの方法においても解決されておらず，安定性の確保という面での問題は依然残されている。この点については，採点者が目標言語において十分な聴解力を持ち，かつ，曖昧な出力に対して一定の方針を定めて臨むことが前提条件となるだろう。

シャドーイング技術の客観的評価には様々な制約があるが，本稿で検討した二つの方法は，それぞれの潜在的な問題点に配慮しながら用いれば，十分評価法として成り立つ可能性が示唆されたように思う。シャドーイングそのものが，他のリスニング指導法よりも比較的短期間で聴解力の伸長効果が現れることが確認されているだけに，トレーニング法としての教室での積極的な使用はこれからもどんどん拡大することが考えられる。そのためにも，さらに高い信頼度を持ち，かつ効率的な評価法の開発は今後も継続されねばならない。

References

阿部純一 et al.『人間の言語情報処理』，サイエンス社， 1994.

Aitchison, Jean. *Words in the Mind — An Introduction to Mental Lexicon*, Cambridge: Blackwell,1987.

Baddeley, Alan. *Human Memory — Theory and Practice*, Cambridge: Lawrence Erlbaum Associates, 1990.

Baddeley, Alan. *Working Memory,* New York: Oxford University Press, 1986, 71.

バッドリー・アラン 川幡政道訳 『記憶力—そのしくみとはたらき』誠信書房, 1982.

二谷広二 「注意の認知構造と注意喚起の心理学的要因」兵庫教育大学学校教育学部附属実技教育研究指導センター 実技教育研究 1990.

二谷広二 「作業記憶と長期記憶との連携プレーによる第二言語処理・学習習得過程について」JELES 発表レジュメ, 1994.

Kurz, Ingrid. "'Shadowing' Exercises in Interpreter Training," *Teaching Translation and Interpreting: Training, Talent and Experience.* Papers from the First Language International Conference Elsinore, Denmark, 31 May – 2 June 1991, Cay Dollerup and Anne Loddegaard (eds.). Amsterdam/Philadelphia: John Benjamins Publishing Company, 1992.

Lambert, Sylvie. "A Human Information Processing and Cognitive Approach to the Training of Simultaneous Interpreters." In Hammond, Deanna L.(ed.) 1988.

大内茂雄（編）『講座・英語教育工学—研究と評価』, 研究社, 1973.

ロフタス, G. & R. ロフタス 『人間の記憶』, 大村彰道（訳）, 東京大学出版会, 1976.

佐藤史郎『クローズテストと英語教育』南雲堂, 1988, 15.

Schweickert, R & B. Boruff, "Short-Term Memory Capacity: Magic Number or Magic Spell?", *Journal of Experimental Psychology: Learning, Memory and Cognition,* Vol. 12, No.3, 1986, 419–425.

染谷泰正「通訳訓練手法とその一般語学学習への応用について」第47回通訳理論研究会報告要旨, 1996, 27–43.

玉井　健「"follow-up"の聴解力向上に及ぼす効果および"follow-up"能力と聴解力の関係」,『STEP BULLETIN』 Vol.4, 日本英語検定協会, 1992, 48–62.

玉井　健「シャドーイングの効果と聴解プロセスにおける位置づけ」『時事英語学研究』　vol.36, 1997.

Van Dam, Ine Mary. "Letter to the Editors." *The Interpreters' Newsletter* 3, 1990, 5–6.

Wainman, H. "Cloze Testing of Second Language Learners," *English Language Teaching* 33(2), 1979, 126–32.

吉富朝子, 荒井貴和　「通訳過程に関する基礎実験」　文部省助成研究『外国語教育の一環としての通訳養成のための教育内容方法の開発に関する総合的研究』 第5章, 1991.

Assessment of the Shadowing Skill in the Classroom

KEN TAMAI

Shadowing is defined as the act or task of listening in which the learner tracks the heard speech and repeats it as exactly as possible. The purpose of this paper is two-fold. The first is to discuss what it means to assess the shadowing skill and the difference in comparison with widely acknowledged listening tests such as TOEFL/TOEIC. The second is to propose two methods for assessing shadowing skills, and to examine their reliability, validity and practicality as measurement tools. The syllable method counts the number of the properly-uttered syllables, and the check points method counts the number of properly-uttered words which are listed beforehand. The results suggest the high level of reliability of the syllable method and the usefulness of the check-point method with its adequate level of reliability and convenience. They also indicates possible problems that need to be clarified in the future.