記述式英語テスト問題における コンピュータ支援評価法の吟味

東京大学教育情報科学研究室　池　田　英　子

Computer-Aided Scoring of English Constructed Response Test

Eiko IKEDA

A new technique of the computer-aided scoring method is proposed on the constructed free-response item formats of English-Japanese translation tests and Japanese-English translation tests. Counting of the spelling errors, length of the sentences and inclusion of the keywords are used for the computer scoring. The results are compared with expert judgment on the test answers obtained from 500 test takers. The responses with large difference between computer scoring and expert rating are also investigated by the additional raters. The nature and characteristics of test questions appropriate to computer-aided scoring are examined and usefulness of the proposed method is confirmed.

目　　次

Ⅰ．問題と目的

近年，学習の成果を評価するにあたり，学習者のパフォーマンスそのものに焦点をあてるテストの研究が盛んになってきている（Linn, Baker & Dunbar, 1991；Nickerson, 1989；Moss, 1992）。客観式テストでは表現力，構成力等の測定が不十分であると考えられるからである。その意味で，記述式テストが改めて注目されている（Frederiksen & Collins, 1989）。特に，英語の学習において，和文英訳や英文和訳等の記述式問題は，言語の習得状況を確認する上で重要である。

しかし，記述式テストの採点には採点者の主観がはいりやすく，受験者全体に統一のとれた信頼性の高い採点を行うことが難しい（田部，1987）。この問題を解決するひとつの手段として，コンピュータを人間の判断の補助手段として利用するハイブリッドな「コンピュータ支援型採点」が考えられる。記述式の答案であっても，それがワープロで入力されるような時代になると，コンピュータで判断可能な部分の評価はコンピュータが行うことも不可能ではない。解答者の自由思考に依存する部分が大きい論述式テストの場合は別として，英文和訳や和文英訳のようにある程度解答範囲に一定の枠組みが期待される課題であれば，十分コンピュータを導入できると思われる。この研究の目的は，英文和訳と和文英訳課題にコンピュータ支援型採点を試みた場合の差異を比較し，採点のしやすさや，測定内容の妥当性を検討することと，実際にコンピュータ支援採点を行う場合に適切な問題設定は何かということを考察することである。

II. 研究 1

A. 方 法

1. 使用する問題

題材は採点済みの英文和訳2題，和文英訳2題についての500名分の答案を利用できる機会があったのでそれを利用することにした。英文和訳の2題のうち英文和訳1は140語程度の英文の大意を要約するもので，英文和訳2は140語程度の英文中の35語程度を部分訳する課題である。和文英訳の2題（和文英訳1と和文英訳2）は，約300字からなるある評論の一節で，100字程度を部分訳する問題である。問題文の難度はかなり高く，大学生程度以上の英語学習上級者を対象としている。答案の採点は複数の英語専門家の分担で行われ，模範解答を基準に，0点から10点までの主観的評定点として与えられている。課題ごとに単数または複数の採点者が分担採点しているが，ひとつの答案についてみれば一人の採点者のみによって評定されている。答案はフロッピーに入力しておく。

2. 手続き

本研究の一つの目的は，人間による採点をできるだけ補助できるようにコンピュータの機能を利用する方法を考えることである。まず，人間が英文和訳や和文英訳の採点を行う場合について考えてみる。人間の場合は採点というひとまとまりの作業過程の中で，機械的に判断する部分から，総合的な判断を必要とする部分までを織りまぜながらこなしている。それらの判断作業にはコンピュータが可能と思われるものも含まれている。そこで英文和訳・和文英訳の採点を行う際に，コンピュータで測定する要素を次のように設定する。

a. コンピュータ測定値の設定

コンピュータで測定できる要素は，大きく次の3つのレベルに分類できる。

(1)機械的な判断でできるレベル

(2)ある程度人間の判断を加え，その判断ルールを定式化して測定するレベル

(3)人間の判断を必要とし，しかもその判断ルールを定式化することが困難なレベル

(1)まず，機械的に判断でき，採点者の主観的判断がはいらないレベルのものとして，スペルミス，文章やパラグラフの長さ，使用語彙の難易度の測定などの形式的な部分の採点が考えられる。スペルミスが文章の善し悪しの基本的な要素になることは明らかであるが，文章やパラグラフの長さ等も読みやすさの基本要素になると考えられる。読みやすい文章にはそれなりに適切な長さがあるのではないか，あるいは適切な解答はそれなりの長さを必要とし，不十分な解答や正解の見い出せない解答は十分な長さを持っていないのではないかということである。そうした要素をコンピュータで客観的に一義的に測定することは不可能ではない。同様に，使用語彙の難易度等も，たとえば語彙使用の頻度水準などを手がかりにその答案の質を決める重要な要素となろう。こうした要素はコンピュータで機械的に判断できる部分である。

(2)次の，ある程度人間の判断を必要とするレベルとは，正解として重要な言葉が何であるかの判断は人間がするとしても，それが答案に含まれているか否かの判定はコンピュータにしてもらうという考え方である。つまり採点のルールを人間が決めて，コンピュータがその判定を実施するのである。

人間が行う採点の場合を考えると，採点者の頭の中には，正答に当然含まれるべきであると期待するいくつかの用語があるはずであり，それが答案の中に書かれているかどうかを意識的または無意識的にチェックしていると考えられる。そこで，ルールとして，よい答案には当然含まれていると思われる基本用語(キーワードとよぶ)を設定し，それが含まれているかどうかをコンピュータで探索することを考える。設定すべきキーワードはあらかじめ複数の採点者の合意の上で決めることもできる。

そのほか，人間による採点には不注意による見落としや，一つの表現の過大評価や過小評価による採点の非一貫性も見られることがあるが，コンピュータによる採点ではそのような危険を避けることができる。そうした人間の注意の不完全性を補うほか，時制，単数・複数形，冠詞，文頭の大文字．文末のピリオドの有無など，必要に応じて基本的な文法ミスのチェックも可能であろう。

(3)さらに，人間の判断にとっていちばんレベルが高いと考えられる文章表現の洗練さや構成の美しさといった問題が残るが，これはコンピュータによる採点としても最高段階のものである。そうした問題の解決にはエキスパートシステムの完成を待たねばならないので，本研究では主題からひとまずはずすものとするが，そうした研究もすでに始まっていることを付記しておく(Johson & Soloway, 1985；Bennett, Gong, Kershaw, Rock Soloway & Macalalad, 1990；Braun, Bennett, Frye & Soloway, 1990；Bennett, Rock, Braun, Frye, Spohrer & Soloway, 1990；Sebrechts, Bennett & Rock, 1991；Bennett, Sebrechts & Rock, 1991)。

そこで本研究では手始めとして，これらの中から答案文の長さ，スペルミス，キーワードの3要素をもとにし

たコンピュータ採点を取り上げることにする。

文の長さは答案をフロッピーに入力しておけば，簡単なプログラムで容易に測定できる。またスペルミスの測定は，和文英訳については，市販のソフト Right Writer にあるスペルチェック機能を用いることにした。英文和訳については，１つ１つ日本語をチェックしていけば数えられないことはないが，大量の答案に対して一般化することが難しいので，今回は扱わないことにした。キーワードの作成にはいくつかの考慮するべき点があるので，次節にまとめる。

b．キーワード点の設定

(1)　キーワードの選び方

人間の判断でキーワードを読みとって採点していくとき，模範解答で要求されている言葉が使用されているときは当然採点の対象となるが，模範解答と同意の別の言葉が使われていても当然採点の対象となるはずである。そこで，コンピュータで測定する場合も，その状況にできるだけ近づけるため，模範解答で使われた単語以外にそれと同等と思われる単語が出現すれば，それを拾い上げるようにキーワードリストを作成した。

説明のために簡単な例を用いることにする。

（英文和訳問題例）

Nothing is more important than health.

（模範解答）

健康ほど大切なものはない。(小寺，1989)

（キーワードリスト）

模範解答以外に,「もっとも大切なものは健康です。」などの解答も可能なので，キーワードリストには

健康，大切，{いちばん，一番，もっとも，最も}等をあげる。

（和文英訳問題例）

私は英語がペラペラになりたい。

（模範解答）

I want to speak English fluently.

（キーワードリスト）

模範解答以外に，

I want to be a fluent speaker of English.

I want to be fluent in speaking English.

I want to have a good command of English.

(小寺，1989)

などの解答も可能なので，キーワードには

I, want, to, {speak, speaker}, English, {fluent, fluently}, a good command of, 等を設定する。

ここで模範解答の中の言葉と同等の意味を持つ単語をキーワードに含めるようにすると，キーワードにかなり多くの言葉を含めることになる。この場合，英文和訳問題の場合と和文英訳問題の場合では，キーワード作成上注意すべきポイントに違いがある。

英文和訳問題ではキーワードが日本語である。日本語では語尾が{健康だ，健康です}のように様々に変化するので，文節の語尾を含めるか含めないかによって，採点の規準を厳しくするか否かが決まってくる。ここでは語尾は含めないことにした。また,{いちばん,一番,もっとも，最も}などのような漢字と平仮名の使い分けについては，自然な表現をすべてキーワードに含めることとし，それらの間に差はつけないことにした。

接尾語，接頭語等の派生部をキーワードに含めるようにすると，それだけ指定した個々の組み合わせの出現頻度が少なくなり，その結果採点は辛くなる。逆に共通部分だけ（たとえば「健康」だけ）をキーワードに取り上げるようにすると，出現頻度は多くなり，実質的に採点を甘くすることと同等になる。その範囲をどの程度にするのが最適であるかは今後の重要な研究課題の一つであるが，キーワード設定の条件は余り細かくしない方がよいように思われる。派生部分の差異は日本語の言い回しの違いによるものが多いからで，本質的な誤りからくるものは少ないと思われるからである。

和文英訳問題ではキーワードが英単語句となる。英語では名詞の複数形が単数形を含むことがあり（例えば book は books の一部），また，規則動詞では過去形が現在形を含む（例えば play は played の一部）ことがある。したがって，book や play だけが含まれているかどうかを探していくと，同時に books や played もカウントされることになり,両者を区別しなければならないときは，それぞれを識別できるような判定プログラムを別に組まなければならない（単複や現在過去を問わないときには必要はない)。

冠詞については，それがとくに必要と考えられた場合にはキーワードに含める形にした。

(2)　キーワードの得点化

以上の方法でキーワードリストを作成した結果，キーワード数は Table 1 上段のようになった。このリストは全答案をざっと見通して重要と考えられる出現単語句を拾い上げて作成された。しかし，答案に同意語も含めて必要な単語句が含まれているかどうかという点が問題なのであるから，キーワードの得点化では，模範解答にある単語と同等と見なせるキーワードはひとつのグループにくくり，そのグループ内にあるいずれかの単語が答案に含まれていれば１点，含まれていなければ０点として採点することにした。例では{いちばん，一番，もっと

も，最も｝をひとつのグループとして扱うことになる。このようにするとグループの個数はTable 1 下段に見られるように整理された。ここで分かるように，英文和訳問題の方が，同一グループに属するキーワードの個数が多いことが分かる。つまり，同じことを表現する語彙が日本語の場合豊富であることを示している。

Table 1
Number of Keywords and Group of Keywords

	英文和訳1	和文英訳1	英文和訳2	和文英訳2
個別単語句ごと	168個	91個	72個	76個
キーワード群	42組	31組	21組	33組

ところで，いくつかのキーワードを同一内容のものとしてひとつのグループにまとめる場合，それぞれの単語句を同等に扱ってよいかどうか問題である。上の例でいうと｛いちばん，一番，もっとも，最も｝の中で，選んだ単語に差をつけなくてもよいかが問題である。ここでは最初の試みなので，英文和訳，和文英訳ともにそのような差を区別せず，グループ内の同意語は対等に扱うことにした。また，同じ単語が同一文の中で複数現れる場合があるが，その頻度も対象としないことにした。評価の本質的部分とはあまり関係がないと考えられることが多いからである。たとえば

「太郎は友達の大勢いる太郎のクラスが大好きである。」と

「太郎は友達の大勢いる彼のクラスが大好きである。」

とでは，

「太郎」をキーワードとするとき，前者では2回，後者は1回数えられ，前者の答案の方が得点が高くなってしまうからである。こうしたルールの細かい設定の違いが採点の誤差変動として現れてくることになるが，それが得点全体の信頼性にどう影響してくるかは今後の研究課題としたい。

このようにして各解答者の答案を数えて得られた該当キーワード（またはキーワード群）の個数をここでは「重みなしキーワード点」と呼ぶことにする。

c．コンピュータ測定値と採点者評定点の算出

以上のようにして得られたスペルミス（和文英訳の場合のみ），答案文の長さ，重みなしキーワード点からなる各答案のコンピュータ測定値は人間による評定点とどれだけ一致するものか，つまり，これらのコンピュータ測定値が人間の判断とどれだけ共有できるかを検討するため，各コンピュータ測定値と人間による評定点との関係を調べることにした。比較のために両者の単純相関係数，並びに各コンピュータ測定値（スペルミス，文の長さ，キーワード点）と評定点との重相関係数が求められた。

ところで，この重みなしキーワード点の算出にあたっては，答案に含まれたそれぞれのキーワード（またはキーワード群）の相対的な重要性の違いは考慮されていない。設定されたキーワード（またはキーワード群）が含まれていれば1点，含まれなければ0点を与えて合計するという単純な方法がとられたことになる。このような方法は他にキーワード相互の相対的重要さを判定する基準がないときに，実際にコンピュータ測定値を利用して採点するには便利な方法であるが，不自然な採点法であることも事実である。しかし，外部に何らかの手がかりを持たない以上，キーワード間の重要度の差を数値で特定することは難しい。

幸いなことに，ここでは採点者の評定点がすでに与えられているので，その採点者の判断を重要さの基準にとるという観点から，それと一番近くなるように，各キーワードの重みを決めることを試みた。つまり，各キーワード（またはキーワード群）と採点者のつけた評定点との重相関係数を求めることによって，各キーワード（またはキーワード群）の重みを決定するのである。こうして人間の判断を取り入れた場合のコンピュータ測定値と採点者評定点の相関も合わせて算出した。

なお，採点者の評定点には，採点者の評定点そのものをもとにする方法と，ロジット評定点をもとにする方法との両方が用いられ比較された。

一般に上限に満点を設定した（この場合は10点満点）テストの評定点は，基本的に満点に対する正解度の割合を意味するものであり，$-\infty$から∞まで変化する連続量とは考えられない。そのために床効果や天井効果が生じ，他の連続変量と曲線相関を持ちやすい。その欠点を補うため，0-1（0-100%）の範囲を持つ評定点のロジット値が$-\infty$から∞までの直線で表されることを利用して，ここにロジット点Yを定義することにする。つまり，Pを満点を1とした評定点とするとき，

$$Y=a\cdot\log\{P/(1-P)\}+b$$

で表される。ここでaとbはロジット値を標準化するための単位と原点を表す。こうして定義された標準化Yを「ロジット評定点」と名付けることにする。

d．評定者間の主観的評定点の不一致の調整

記述式テストにおいて，主観的評定点を利用する際に

常に問題となるのは，採点者間の評定不一致の問題である。つまり，各採点者の評定基準が異なっており，そのために評定平均値やばらつきが採点者によって一致しないことである。特に，一つの課題を複数の採点者が分担して採点する場合には，各採点者の評定基準のずれを調整することが望ましい。

一般にとられる方法のひとつは選択科目間の調整などにも応用される「アンカーテスト法」である。たとえば，すべての受験者が受験している客観テストがあれば，その結果を手がかりとして，採点者間の評定平均値やばらつき（標準偏差）の違いが客観テストにみられるずれと等しくなるように各採点者のつけた評定点を調整する方法である。

しかし，これはアンカーテストにとった客観テストの測定するものが，調整対象となるテスト得点が測定しようとしているものと同一である場合に正しい。もし，アンカーテストが意図した測定対象とは異なる特性を測定していると考えられる場合にはこの方法は適当ではない。事実この例の場合にも，同じ解答者に対して実施された客観式テストの情報（語彙，文法などを問う問題）があり，それと評定点との相関係数はTable 3 最下段に示すようになった。結果はそれほど高くなく，記述式テストの評定点が測ろうとしているものと，客観式テストが測ろうとしているものとが，同一の特性でないことを示唆するものである。しかし，測定対象となる答案を客観的に採点したコンピュータ測定値は，どの解答者に対しても共通の基準で評価したものであり，それをアンカーとして，異採点者間の評定のずれを調整することは今までの方法に比べてより優れた方法であると考えられる。

その目的のために，ここでは正準相関分析法の考え方を応用した一つの調整法を提案する。それはコンピュータ採点に利用した文の長さ（X1）およびキーワード点（X2）さらに和文英訳の場合は，スペルミス（X3）も含めて，それらの重み付き合成点と，各採点者のつけた評定点またはロジット評定点の調整点との相関が最大になるように，各採点者別にそれぞれのロジット評定点を調整するための尺度の重みと位置定数を決める方法である。つまり，採点者 j がつけたロジット評定点Yを採点者ごとに定められた定数 bj1，bj2 を用いて

$$Y' = bj1Y + bj2$$

の形に一次変換し，それと，コンピュータによって客観的に得られたコンピュータ測定値(X1，X2，X3)の合成得点

$$X' = a1X1 + a2X2 + a3X3$$

との相関が最大になるように，a1，a2，a3 および bj1，bj2 の値を決定する方法である。こうして各コンピュータ測定点を合成してもとめられた値X′を「コンピュータ基準点」，また，各評定者のロジット評定点を調整して得られた値Y′を「ロジット評定基準点」と名付け，両者の相関係数を求めることにした。なお，重み計算にあたっては，コンピュータ基準点X′の平均，標準偏差が0，1になるように規準化してある。

B．結果と考察

1．コンピュータ測定値による採点の効果

こうして求められた文章の長さ，スペルミスの測定値，キーワード点の評定点とロジット評定点との相関はTable 2 に示す。

Table 2
Correlation Coefficients of Computer Scores with Rating and Logit Rating Scores

	英文和訳1		英文和訳2	
項　目	評定点	ロジット評定点	評定点	ロジット評定点
文の長さ	0.492	0.619	0.268	0.332
重みつきキーワード点	0.660	0.713	0.660	0.705
重みなしキーワード点	0.531	0.589	0.553	0.613
	和文英訳1		和文英訳2	
項　目	評定点	ロジット評定点	評定点	ロジット評定点
文の長さ	0.810	0.828	0.705	0.763
重みつきキーワード点	0.811	0.856	0.719	0.763
重みなしキーワード点	0.751	0.773	0.693	0.732
スペルミス	0.113	0.137	0.206	0.226

コンピュータで測定した諸要素とロジット評定点との相関係数をみると，文の長さ，キーワード点ともに英文和訳より和文英訳の方が高い相関を示している。このことは，和文英訳の方がコンピュータ測定が人間の判断に近いことを示唆している。日本人が英語を学習する場合，

習得する内容が比較的型にはまっていて，答案に多様性が表れにくいという点が影響しているのではないかと思われる。外国語は母国語と違い，使える表現パターンが少なく，答案で正答とされるには，特定の語句や文法を知っているかどうかにかかっているのである。そして，その語彙を知っていれば答案が書けるために文の長さも長くなって点が上がるというように2つは結びついていると考えられる。和文英訳でスペルミスと評定点とにあまり関連性が見られなかったのは，解答者に基本的なミスが少なかったためか，採点者がスペルミスを重視していなかったためと考えられる。

コンピュータ測定値のロジット評定点に対する重相関係数は Table 3 に示す。これについても，英文和訳より和文英訳の方が高くなっている。このことからも和文英訳問題についてコンピュータ測定値を利用することは十分可能であると思われる。

Table 3
Multiple Correlation Coefficients of Computer and Multiple-Choice Test Scores with Logit Rating Score

	英文和訳1	和文英訳1	英文和訳2	和文英訳2
キーワード点が				
重みありのとき	0.739	0.885	0.706	0.822
重みなしのとき	0.668	0.856	0.664	0.804
客観式問題との相関	0.211	0.333	0.385	0.343

採点者の評定を取り入れた重みつきキーワード点を使用する場合と取り入れない重みなしキーワード点を使用する場合とでは，いずれの課題の場合も取り入れたほうが相関が高くなる。このことはコンピュータ採点にも全くコンピュータだけに依存するより人間の判断を残した方がよいという相互補完的な使用法を示唆するものである。

2. 採点者間の評定基準のずれの調整の効果

コンピュータ測定値を利用して採点者間の違いを調整し，「ロジット評定基準点」と「コンピュータ基準点」を推定したが，それぞれの基準点を出すための重み，および両者間の相関係数は Table 4 および Table 5 のようになった。

Table 4
Comparison of Adjusted Logit Rating Score and Adjusted Computer Score

英文和訳1		重みつきキーワード点		重みなしキーワード点	
		尺度単位	原点の単位	尺度単位	原点の単位
		bj1	bj2	bj1	bj2
採点者	1	0.8327	−0.1614	0.7153	−0.1469
	9	1.2315	0.0544	1.1190	0.0766
	11	0.5136	−0.2342	0.2850	−0.0724
	12	1.9031	−0.3583	1.9850	−0.4179
文の長さ	a1	0.4820		0.7203	
キーワード	a2	0.6118		0.3718	
和文英訳1		重みつきキーワード点		重みなしキーワード点	
		尺度単位	原点の単位	尺度単位	原点の単位
		bj1	bj2	bj1	bj2
採点者	3	1.0980	0.4279	1.0631	0.2611
文の長さ	a1	0.4279		0.6509	
キーワード	a2	0.6147		0.4020	
スペルミス	a3	0.0268		0.0486	
英文和訳2		重みつきキーワード点		重みなしキーワード点	
		尺度単位	原点の単位	尺度単位	原点の単位
		bj1	bj2	bj1	bj2
採点者	2	0.6289	−0.0218	0.5517	−0.0308
	5	0.9712	0.4596	0.8909	0.4305
	6	1.0013	0.1375	0.9217	0.1438
	8	0.4887	−0.0835	0.2851	−0.0010
文の長さ	a1	−0.0007		0.0466	
キーワード	a2	1.0003		0.9737	
和文英訳2		重みつきキーワード点		重みなしキーワード点	
		尺度単位	原点の単位	尺度単位	原点の単位
		bj1	bj2	bj1	bj2
採点者	4	0.9591	0.4215	0.9574	0.4161
	7	0.9875	1.5339	0.9461	1.4985
	10	0.9707	0.3528	0.9439	0.3363
文の長さ	a1	0.5833		0.6414	
キーワード	a2	0.4807		0.4159	
スペルミス	a3	0.0383		0.0429	

Table 5
Correlation Coefficients between Adjusted Logit Rating and Adjusted Computer Scores

	英文和訳1	和文英訳1	英文和訳2	和文英訳2
重みありのとき	0.798	0.884	0.746	0.872
重みなしのとき	0.749	0.856	0.661	0.858

各コンピュータ測定値を合成したコンピュータ基準点と，採点者の評定を調整したロジット評定基準点との相関は予想以上に高い。このことから，英文和訳・和文英訳問題の採点を行うとき，コンピュータ測定値として今回取り上げた採点のための3要素を用いることだけでも効果的であることがわかる。さらに他の要素を加えていくことで，一層効果が上がることが予想される。

ところで，全体的にみればコンピュータによって支援することは十分可能であることがわかったが，個々の答案についてみるとロジット評定基準点とコンピュータ基準点の間に大きなくい違いのある答案もいくつか見られた。これは人間の判断をコンピュータがうまく予測できなかったためと考えられる。くい違いの大きかった答案に対してその原因を検討することは重要である。

III. 補足実験　不一致ケースの検討

A. 方　法

1. 使用する答案

研究1で，各答案に対して，キーワードに重みをつけた場合と重みをつけない場合のそれぞれについてロジット評定点・ロジット評定基準点・コンピュータ基準点を算出した。そのなかで，ロジット評定点をステイナイン化した1点から9点までの各段階において，{ロジット評定点，ロジット評定基準点，コンピュータ基準点} の三者の間に大きな差がある答案，および，重みつきキーワード点をもとに計算した基準点と重みなし（単純和）キーワード点をもとに計算した基準点とで結果の差異が大きいものから各課題ごとに24個の答案を選んだ。ステイナイン得点とは，正規分布を0.5標準偏差ずつ区切った面積の割合に対応するように人数を9段階に割り当てて得点化したものである。1点から9点までの人数は4，7，12，17，20，17，12，7，4％に割り振られる。

2. 被験者

都内の大学で英語専門教育を受けている大学院生4名と，英語を専門に使用する会社の職員8名の計12名に協力を依頼した。

3. 手続き

この実験では，不一致の大きかった各答案について，それが採点者の評定点の単なる非一貫性によるものなのか，今回用いたルールではコンピュータで測定しきれない部分を人間が的確にとらえて評価したのかを調べることを目的とし，独立な複数の採点者の評定と比較した。答案の提示と採点はTable 6のように設定した条件下で，被験者をABCDの4つの場合にランダムに割当ててパソコン上で行った。その場合指示として，参考点としてステイナイン化したロジット評定点，ロジット評定基準点，コンピュータ基準点をどれがそれであるかを伏せておいて提示し，自分ならそのどれに近い評定をするか答えてもらう。参考点の表示は，被験者の判断のずれを防ぐ効果を確認するためである。一方別の条件では，参考点を提示せず，ただ答案をステイナインで評定してもらう。その際，自分がいままでにつけたステイナイン尺度の評定分布がわかるようにしておき，前に遡って評定し直すことも許されている。これも何の手がかりもないために被験者の判断が極端にずれていくのを防ぐためである。

Table 6
Allocation of Conditions of Experiment

実験群	A群	B群	C群	D群
英文和訳1				
参考点の表示	あり	あり	なし	なし
キーワードの重み	あり	なし		
和文英訳1				
参考点の表示	あり	あり	なし	なし
キーワードの重み	あり	なし		
英文和訳2				
参考点の表示	なし	なし	あり	あり
キーワードの重み			あり	なし
和文英訳2				
参考点の表示	なし	なし	あり	あり
キーワードの重み			あり	なし

4. 分析方法

各答案について，同一条件で評定した被験者の評点の平均点と，先に定めたロジット評定点，ロジット評定基準点，コンピュータ基準点との相関係数を計算し，依頼した被験者の評点平均がそのどれにいちばん近いかを比

較する。

B．結果と考察

同一条件での被験者のつけた評点平均と各基準点との相関係数はTable 7の通りである。その結果，英文和訳1だけはどの実験群でもロジット評定点Yとの相関がいちばん高かった。このことは，英文和訳1が大意を要約して述べる課題であったことによるものと考えられる。

その他の課題では，和文英訳1の条件Bを除き，キーワードに重みをつけるときはロジット評定基準点が被験者の判断にいちばん近くなり，キーワードをすべて同等に扱うときはむしろコンピュータ測定値による基準点の方が被験者の判断に近かった。つまり，キーワードの重要性の違いを考慮しないならば，人間の判断はコンピュータに劣るということになる。人間の判断の特徴はキーワードに重み付けをすることにあると思われる。

Table 7
Correlation Coefficients of Rater's Mean Score with Other Comparative Measures

	n	Y	Y′W	X′W	Y′N	X′N
英文和訳1						
A の平均	3	0.664	0.657	0.550		
B の平均	3	0.593			0.552	0.313
CDの平均	6	0.607	0.563	0.469	0.507	0.163
和文英訳1						
A の平均	3	0.755	0.815	0.612		
B の平均	3	0.765			0.809	0.750
CDの平均	6	0.615	0.663	0.540	0.663	0.690
英文和訳2						
ABの平均	6	0.315	0.458	0.428	0.380	0.365
C の平均	3	0.508	0.610	0.547		
D の平均	3	0.249			0.394	0.587
和文英訳2						
ABの平均	6	−0.002	0.110	0.637	0.110	0.752
C の平均	3	0.387	0.613	0.562		
D の平均	3	0.308			0.388	0.621

n：被験者数
Y：ロジット評定点
Y′W：ロジット評定基準点（重みつきキーワード使用）
X′W：コンピュータ基準点（重みつきキーワード使用）
Y′N：ロジット評定基準点（重みなしキーワード使用）
X′N：コンピュータ基準点（重みなしキーワード使用）

参考点を見せると基準点との相関を高める働きがあり，少なくとも採点者の気まぐれ的判定を防ぐ効果はあるものと思われる。参考点を作業の前半で見て評定するのと，見ないで評定するのとでは顕著な差異は見いだされなかった。

IV．結　論

英文和訳・和文英訳問題をコンピュータでの測定値から得られる情報を利用して採点することは可能である。特に，和文英訳問題の採点の方がコンピュータの支援効果があることがわかった。和文英訳の方がうまく推定できるだけでなく，キーワードを設定する際も，和文英訳のキーワードを作成する方が労力がかからなかった。英文和訳は解答が日本語なので表現の多様性が高く，キーワードによる正解への接近が難しい。逆に，和文英訳の場合は解答者が使える表現パターンが限られており，コンピュータによる正解への接近が比較的容易であると思われる。

課題の内容のわずかな違いによってもコンピュータによる補助のしやすさに差が見られた。英文和訳問題で大意を要約する場合は，単なる部分訳と違い，文章をまとめる力なども採点の対象になっていて，文の長さとキーワードの測定だけではとらえきれなかったと考えられる。

記述式答案といっても，目的によって課題の内容や採点の上で重視するポイントは様々である。コンピュータを利用して採点を補助する場合に適した記述式問題，出題形式をさらに明らかにしていきたい。

同時に，コンピュータ測定値のような補助手段を用いることにより，各問題の妥当性の検討も行えることが示唆された。テストを行うときには，そのテストで測定したい内容をあらかじめ決め，それに対して適切な問題を作成するはずである。そのプロセスがはっきりしていれば，それに沿ってコンピュータが測定できる内容もはっきりしてくる。もし，コンピュータ測定値と人間の採点との相関が低かったり，一連の問題の中ではずれた値が出るときは，それは両者のいずれかまたは両方が，本来測定したかった内容とは別のことを測定したことを意味するであろう。その意味で，問題の項目分析にも役立てることができる。

以上のように出題内容の違いがコンピュータ測定値に影響を及ぼすことから，今後さらに充実したコンピュータによる支援方法と，コンピュータ支援型採点に適した出題方法の研究を重ねていくことが必要である。コン

ピュータと人間の力をうまく嚙み合わせることが今後の発展につながるはずである。

（指導教官　渡部洋教授）

謝　辞

本論文は東京大学大学院教育学研究科に提出した修士論文（1992年度）の一部に手を加えたものです。本論文の作成にあたり御指導いただきました，東京大学教育学部の渡部洋先生に心から感謝いたします。また本研究の素材となる資料提供ならびに実験に協力いただいた日本英語検定協会に御礼申し上げます。

引用文献

小寺茂明（1989），「日本語の対比で教える英作文」，大修館書店。

田部定義（1987），「主観テストの採点上の「ゆれ」に関する統計的研究」，英語教育　1987　12　40-48。

Bennett, R. E., Gong, B., Kershaw, R. C., Rock, D. A., Soloway, E., & Macalalad, A. (1990). Assessment of an expert system's ability to grade and diagnose automatically student's constructed responses to computer science problems. In R. O. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp.293-320). Hillsdale, NJ : Lawrence Erlbaum Associates.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of constrained free-response to multiple-choice and open-ended items. *Applied Psychological Measurement*, 14, 151-162.

Bennett, R. E., Sebrechts, M. M., & Rock, D. A. (1991). Expert-system scores for complex constructed-response quantitative items : a study of convergent validity. *Applied Psychological Measurement*, 15, 227-239.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, 27, 93-108.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18, 27-32.

Johnson, W. L., & Soloway, E. (1985). PROUST : An automatic debugger for Pascal programs. *Byte*, 10 (4), 179-190.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment : expectations and validation criteria. *Educational Researcher*, 20, 15-21.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement : implications for performance assessment. *Review of Educational Research*, 62, 229-258.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher*, 18, 3-7.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Machine-scorable complex constructed-response quantitative items : agreement between expert system and human raters' scores. (*GRE Board Professional Report* No.88-07aP, ETS RR-91-11). Princeton, NJ : Educational Testing Service.