日本人学習者の英語語意知識測定テストの開発と検証

―正答率および応答自信度による評価―

東京都／東京大学大学院博士課程在籍　中田 達也

概要　本研究では，L2における語意知識（語の意味に関する知識）を測定するためのテストである再認式文章完成課題において，応答自信度を得点に反映することで，テストの1）信頼性，2）妥当性，3）実用性が改善されるかどうかを調査した。98人の日本人大学生を対象とした調査の結果，自信度を語意テスト得点に反映することで，信頼性と妥当性の構造的側面に関しては改善される可能性が示唆された。しかし，妥当性の一般化可能性および外的側面に関しては，差が見られなかった。また，テストの実用性に関しては，自信度を用いない従来の採点方式の方が優れていた。本研究の結果から，再認式の語意テストにおいて自信度を用いることは必ずしも必要ではないものの，高い信頼性が望ましい場合には，自信度を採点に取り入れることが1つの選択肢になり得ることが示された。

1 背景

第二言語の学習者は，“My mother is fond of *gathering* stamps”（cf. *collecting*; Maingay & Rundell, 1987, p.129），“When you visit us next time, please *take* your children”（cf. *bring*; Takahashi, 1984, p.142）といった誤用をすることが示されている。学習者が規範的ではない表現を使用する大きな原因の1つとして，L2における語意知識（語の意味に関する知識）の不足が挙げられる（Takahashi, 1984; Tanaka & Abe, 1985; Tanaka, Takahashi, & Abe, 1990; 今井, 1993他）。例えば，日本人学習者は英単語 wear を「着る」という和訳と結びつけて理解することが多いため，wear に対して持っている意味知識は限定的なものであることが多い。結果的に，wear a uniform / a dress / a jacket などの表現は使用できても，wear a hat / a ring / glasses / stockings などのコロケーションは使用できないことがある（今井, 1993）。

L2における語意知識は語彙運用能力の基盤となる重要な要素であるため，これまでにさまざまな語意知識測定テスト（以下，「語意テスト」と呼ぶ）が開発されてきた。しかしながら，従来の研究では語意テストの信頼性，妥当性，実用性について，十分な検証が行われていない。そこで，本研究ではL2語意テストとして広く用いられてきた再認式文章完成課題に焦点を当て，その信頼性，妥当性，実用性を検証・改善することをめざす。

2 先行研究

L2語の基本義を知っているにもかかわらず，その単語を誤用してしまうという現象は，外国語の学習者や教師にとって非常に身近な問題である（Ishii, 2005; Ishii & Schmitt, 2009; Jiang, 2004; Maingay & Rundell, 1987; Takahashi, 1984; Tanaka, 1983; 松田, 2004など）。したがって，学習者の語意知識を測定するという試みは，多くの研究者の関心を集めてきた。L2学習者の語意知識を測定するために，文章完成課題，容認性判断課題，類似性判断課題，誤文訂正課題，翻訳課題，自由作文，面接などさまざまな形式のテストが用いられてきた（Maingay & Rundell, 1987; Mochizuki, 2006; Tanaka, 1983;

Tanaka & Abe, 1985; Takahashi, 1984; 今井，1993; 中田, 2009; 松田, 2004など）。

中でも最も一般的に使われているのは，再認式文章完成課題である。再認式文章完成課題では，文脈に合う単語を選択肢の中から選び，文を完成させることが求められる。以下に，Nakata（in press）が用いた再認式文章完成課題の例を示す。

▶ 図 1：再認式文章完成課題の例

2つの選択肢の内，どちらか1つの単語を（　　）に入れ，日本語訳に合うように英文を完成させてください。わからなかったら推測せずに，「わからない」を選んでください。

1. When you visit us next time, why don't you（take / bring / わからない）your children with you?
今度私たちを訪ねる時は，あなたのお子さんを連れて来てください。
2. When you go out tonight, don't forget to（take / bring / わからない）an umbrella with you.
今夜外出する時は，傘を持って行くのを忘れないでください。
3. I'll（take / bring / わからない）you a cup of coffee.
コーヒーを一杯あなたに持って行ってあげましょう。

模範解答：1. bring,　2. take,　3. bring

（注）Takahashi（1984), Tanaka & Abe（1985), Nakata（in press）を基に作成。

図 1 に示したテストに正解するためには，take と bring の語意を正確に理解していることが必要になる。したがって，図 1 の文章完成課題は学習者の語意知識を測定していると見なすことができる。

再認式文章完成課題は，学習者の語意知識を測定するためのテストとして広く用いられてきた（Csabi, 2004; Ishii, 2005; Ishii & Schmitt, 2009; Jiang, 2004; Koya, 2005; Mochizuki, 2006; Nakata, in press; Okuno, 2008; Tanaka, 1983; 小屋, 2002; 田頭, 2007; 水口, 2002など）。しかし，再認式文章完成課題には，いくつかの問題点もある。最も大きな問題として，推測の影響が挙げられる。いくつかの例外（Koya, 2005; Mochizuki, 2006; Tanaka, 1983; 小屋, 2002）を除いて，再認式文章完成課題では二者択一形式をとるものが大半である（Csabi, 2004; Ishii, 2005; Ishii & Schmitt, 2009; Jiang, 2004; Nakata, in press; Okuno, 2008; 田頭, 2007; 水口, 2002）。二者択一式のテストでは当て推量で解答をしても50%の確率で正解できてしまうため，推測の影響を大きく受け，信頼性・妥当性に疑念が残る。

多肢選択式テストにおける推測の影響を取り除く1つの方法として，応答自信度（response confidence）を得点に反映させることが考えられる。すなわち，学習者が選択した答えにどのくらい自信があるかを申告させ，その自信度を得点に反映させるのである。例えば，自信度「高」で正解した場合は 3 点，「中」であれば 2 点，「低」であれば 1 点というように自信度に応じて得点を与えることで，従来の二値的な採点法では測定できないような微妙な知識の差も測定できると考えられる（Kamimoto, 2006; Shizuka, 1999, 2004）。

教育測定や実験心理学の分野では，自信度に関する数多くの研究がなされているにもかかわらず，外国語教育の分野では自信度はほとんど活用されてこなかった（Shizuka, 1999, 2004）。数少ない例外として，Shizuka（1999, 2004），Kamimoto（2006），Iso & Aizawa（2008, 2009）がある。Shizuka（1999, 2004）は多肢選択式読解テストの採点に自信度を反映させることで，信頼性および項目弁別力が高くなるという結果を得た。しかし，自信度を反映したとしても，妥当性に関しては十分な改善は見られなかった。多肢選択式の語彙サイズテストにおける自信度の効果を検証した Kamimoto（2006）では，自信度を採点に取り入れることでテスト得点の信頼性が改善されることが示された。Iso & Aizawa（2008, 2009）では自信度を採点に取り入れることで，多肢選択式の語彙サイズテストにおける推測の影響を統制できる可能性が示唆された。

しかし，再認式語意テストに自信度を取り入れることの効果に関しては，これまでに実証研究は行われていない。先行研究で用いられたテストの多くが二者択一形式を採用しており，推測の影響を大きく受けることを考慮すると，再認式文章完成課題における自信尺度の効果を検証することが必要であると思われる。

3 研究課題

再認式文章完成課題に関する従来の研究の限界を踏まえ，本研究では以下の研究課題について調査する。

「語意知識を測定するための再認式文章完成課題において，応答自信度を得点に反映させることで，テストの 1) 信頼性，2) 妥当性，3) 実用性に変化が見られるか」

4 研究方法

4.1 参加者

本調査の参加者は，日本人の大学 1 年生98人である。98人のうち，38人は A大学，60人は B大学の学生であった。両大学では習熟度別の授業が行われており，A大学の38人の学生は，13段階中一番上のクラスに在籍していた。B大学の学生60人のうち，33人はAレベル（5 段階中 1 番上のクラス），残りの27人は Bレベル（5 段階中 3 番目のクラス）に在籍していた。

参加者の英語熟達度の指標として，片桐テスト（Katagiri, 2002）を実施した。片桐テストは，32問からなる英語語彙サイズテスト（学習者がどのくらい多くの単語を知っているかを測定するためのテスト）である。片桐テストを使用したのは，片桐テストのスコアから受験者の総合的な英語熟達度を高い精度で予測できることが示されている（Katagiri, 2002）からである。なお，片桐テストを実施した日に欠席していた受験者もいたため，片桐テストのデータがあるのは参加者98人中87人であった。KR21により算出した片桐テストの信頼性係数は .72 であった。

片桐テストの平均得点（括弧内は標準偏差）は，A大学の学生で26.97（2.50），B大学 Aレベルの学生で24.43（3.73），B大学 Bレベルの学生で19.33（3.64）であった。一元配置分散分析の結果，3 グループの平均得点には有意差が見られた［F (2, 84) = 40.20, $p < .001$, $\eta^2 = .48$］。3 グループのどこに有意差があるのかを調べるために，テューキー HSD の多重比較を行った。多重比較の結果，いずれのグループの平均得点にも有意な差があることが示された（$p < .05$）。すなわち，片桐テストの得点が高い順に被験者グループを並べると，「A大学の学生＞ B大学 Aレベルの学生＞ B大学 Bレベルの学生」となる。

4.2 使用テスト

本研究では，1) 再認式文章完成課題，2) 再生式文章完成課題という 2 種類のテストを使用する。

4.2.1 再認式文章完成課題

再認式文章完成課題は，学習者の英語語意知識を

▶ 図 2：再認式文章完成課題の抜粋

答え方について

① (A)・(B) どちらかの単語を下線部に入れ，日本語訳に合うように英文を完成させてください。選択肢 (A) または (B) のいずれかに○をつけて答えてください。
答えがわからない場合は，「わからない」に○をつけてください。

② (A) か (B) のいずれかに○をしたら，答えの自信度を「低」・「中」・「高」の 3 つから選んでください。

「低」	答えに自信がない。
「中」	「低」と「高」の中間。
「高」	答えに自信がある。

「わからない」を選んだ場合は，自信度を選択する必要はありません。

英文	(A)	(B)	わからない	自信度
You'd better not buy those bananas. They are still _____. そのバナナは買わない方がいいですよ。 まだ**青い**からです。	green	blue	わからない	低　中　高

模範解答：green

(注) Takahashi (1984) を基に作成。

測定するためのテストである。本テストでは，文脈に合う単語を 2 つの選択肢から選び，文を完成させることが求められた。本研究で使用した再認式文章完成課題の抜粋を図 2 に示す（さらに詳細な説明に関しては，「資料」の項を参照のこと）。

再認式文章完成課題では二者択一の形式をとるものが大半であるため（「2 先行研究」を参照），本研究でもこれにならった。当て推量の影響を軽減するため，「わからない」という選択肢も用意した。学習者は解答を選択することに加えて，解答の自信度を 3 段階で示すことが求められた。具体的には，答えに自信がある場合は自信度「高」を，答えに自信がない場合は自信度「低」を，両者の中間の場合は自信度「中」を選ぶように指示された。「わからない」を選んだ場合は，自信度を選択することは求められなかった。自信度をどのように得点に反映したかという点については，4.3.1で詳述する。

なお，どのくらいの自信度を「低」・「中」・「高」と解釈するかは主観的なものであるため，自信度の使用方法には個人差が見られることが予測される。例えば，ある人の自信度「低」は別の人の「中」に相当し，ある人の自信度「中」は別の人の「高」に相当するかもしれない。本研究では，自信度の使い方に関する個人差を考慮した採点方式を採用することで，個人差が得点に影響しないように配慮した（詳しくは，4.3.1 の 4）COPS を参照）。

出題項目の候補として，50項目が作成された。テスト項目を作成する上では，先行研究（Ishii, 2005; Jiang, 2004; Koya, 2005; Nakata, in press; Takahashi, 1984; Tanaka, 1983; 小屋, 2002など），母語話者コーパス（BNC），辞書，市販の教材などを参考にした。テストで用いる正解および錯乱肢の単語は，学習者にとってすでになじみがあると思われる語（take, make, listen, high, loud, expensive など）から選んだ。その理由は，今回のテストの目的は「ある単語の意味をどのくらい深く理解しているか」を測定することであるが，基本義すら知らない単語の正確な意味理解は困難であると考えられるからである。本実験に先立って54名の日本人大学生を対象としたパイロット・スタディを行い，最終的な試験形式や出題項目を決定した。本実験での出題項目として，パイロットで使用した50項目の中から39項目を選定した。

正解および錯乱肢として使われている語が学習者にとってなじみがあるかどうかを確認するために，正解・錯乱肢となる英単語の JACET 8000（JACET 基本語改訂委員会, 2003）レベルを分析した。その結果，rent を除くすべての単語がレベル 1 〜 2 の語であった。rent のみがレベル 4 の語であったが，以下の理由により正解・錯乱肢として用いても問題がないと判断した。1）日本人大学生に英単語の親密度調査を行った横川（2006）では，rent の親密度平均は 7 段階中5.04と高かった。2）日本語の「レンタカー」，「レンタルビデオ」などの表現から，学習者は rent という語にある程度のなじみがあると考えられる。以上の理由から，rent を含むすべての正解・錯乱肢は，今回の実験の被験者である日本人大学生にとって既知語である可能性が極めて高いと判断した。

また，項目や選択肢の提示順序が学習者の解答に与える影響を相殺するため，提示順序を入れ替えた 4 つのバージョンを用意した。

4.2.2　再生式文章完成課題

再認式文章完成課題の妥当性を検証するための外的基準の 1 つとして，本研究では再生式文章完成課題を実施した（5.3.3を参照）。再生式課題の抜粋を図 3 に示す。

再生式課題では，再認式課題と全く同じ39項目が出題された（問題の出題順序は異なる）。

▶ **図 3：再生式文章完成課題の抜粋**

下線部に適切な英語を入れ，日本語訳に合うように英文を完成させてください。答えは「解答欄」に記入してください（答えがわからない場合は，解答欄は空欄のままにしてください）。

英文	解答欄
You'd better not buy those bananas. They are still _____. そのバナナは買わない方がいいですよ。まだ<u>青い</u>からです。	

模範解答：green

再生式課題と再認式課題の違いは，以下の２点である。

⑴ 再認式課題では選択肢の単語（正答１つと錯乱肢１つ）が提示されるのに対して，再生式課題では選択肢が提示されない。

⑵ 再認式課題では解答の自信度を示すことが求められたのに対して，再生式課題では自信度を示すことが求められない。

項目の提示順序が学習者の解答に与える影響を相殺するため，出題順序が異なる２つのバージョンを用意した。

4.3 手続き

データ収集は授業時間内に実施した。調査の手続きは以下のとおりである。

1. 再生式文章完成課題

再認式文章完成課題の妥当性を検証するための外的基準の１つとして，再生式文章完成課題（4.2.2）を実施した。再生式課題を再認式課題よりも先に実施したのは，後者で選択肢が提示されることで，前者における解答に影響を与える可能性があるためである。

全員が受験し終わったのを確認した後に，再生式課題を回収した。試験実施にかかった時間は約15分であった。

2. 再認式文章完成課題

再生式課題を回収した直後に，再認式課題（4.2.1）を実施した。全員が受験し終わったのを確認した後に，再認式課題を回収した。試験実施にかかった時間は約10分であった。テスト終了後に模範解答を配布し，解説を行った。

4.3.1 採点方法

1. 再認式文章完成課題

本研究では応答自信度を採点に取り入れることの効果を検証するために，応答自信度を考慮しない従来型の採点方法と，応答自信度を得点に反映させる新しい採点方法とを比較する。

具体的には Shizuka（2004）にならい，以下の４種類の採点方法による結果を比較する。

１）Dichotomous：応答自信度を点数に反映させない，従来型の採点方法である。この採点方式では，学習者の自信度にかかわらず，不正解であれば０点，正解であれば１点が与えられる。「わからない」を選んだ場合も，０点として扱う（表１）。

２）Polytomous（旧）：Polytomous（旧）では，正答・誤答ともに自信度を点数に反映させる（どのような方法で学習者に自信度を選択させたかについては，4.2.1を参照）。正解の場合，自信度が「高」であれば６点，「中」であれば５点，「低」であれば４点が与えられる。このようにすることで，従来の二値的な採点方法では測定できないような微妙な知識の差も測定できると考えられる（Kamimoto, 2006; Shizuka, 1999, 2004）。不正解の場合は，自信度が「高」であれば０点，「中」であれば１点，「低」であれば２点が与えられる。「わからない」を選んだ場合は，３点として扱う（表１）。

■ 表１：本研究における得点換算表

正誤	自信度	D	P旧	P新
正解	高	1	6	6
正解	中	1	5	5
正解	低	1	4	4
わからない	—	0	3	0
不正解	低	0	2	0
不正解	中	0	1	0
不正解	高	0	0	0

（注）Shizuka（2004）を基に作成。D ＝ Dichotomous；P旧＝ Polytomous（旧）；P新＝ Polytomous（新）

Polytomous（旧）では，自信度が高い誤答ほど低い点数が与えられる。その理由は，誤答であるかもしれないということを自覚しながら誤答をする学習者の方が，正答であると確信して誤答をする学習者よりも，高い能力を有していると考えられるからである（Shizuka, 1999, 2004）。

また，自信度が高い誤答に低い点数を与えることは，学習者が自信度を正確に申告することにもつながると考えられる。例えば，自信度が高い誤答も低い誤答も同じように扱われるのであれば，良い点をとろうという学習者は，すべての問題に対して自信度「高」を選択することであろう。このような学習者が多くいた場合，自信度のデータは学習者の本当の自信度を反映しているとは言い難く，もはや意味

をなさなくなってしまう（Shizuka, 2004）。したがって，自信度が高い誤答に低い点数を与えることは一種のペナルティとして働き，学習者が自信度を正直に申告することを促すと考えられる。

3）Polytomous（新）：正答のみ自信度を得点に反映させる採点方式。正解の場合，自信度が「高」であれば6点，「中」であれば5点，「低」であれば4点が与えられる。不正解の場合は，自信度に限らず一律0点として扱う。「わからない」を選んだ場合も，0点として扱う（表1）。

Polytomous（旧）では，自信度が高い誤答ほど低い点数が与えられた。その理論的背景となるのは，「誤答であるかもしれないということを自覚しながら誤答をする学習者の方が，正答であると確信して誤答をする学習者よりも高い能力を有している」という仮説であった（Shizuka, 1999, 2004）。しかし，Shizuka（2004）の行った調査では，この仮説を支持する結果は得られなかった。

そこで，Shizukaは不正解の際に自信度に応じて得点を与えるのは妥当ではないと考え，誤答は自信度に限らず一律0点として扱う採点方式を考案した。それが，Polytomous（新）である。誤答および「わからない」に得点が与えられない点を除いては，Polytomous（新）はPolytomous（旧）と全く同一である。

4）COPS：学習者ごとに各自信度における正答率を計算し，その正答率を得点に反映させる採点方式である。

自信度を得点に反映させるPolytomous（旧）およびPolytomous（新）には，1つの問題がある。それは，学習者によって自信度の使い方に個人差があるということである。例えば，ある人の自信度「低」は別の人の「中」に相当し，ある人の自信度「中」は別の人の「高」に相当するかもしれない。このような個人差を無視して，自信度「低」は4点，「中」は5点，「高」は6点と一律の得点を全受験者に与えると，測定誤差が増える危険性がある（Shizuka, 2004）。そこで，自信度の使い方に関する個人差に影響を受けない採点方法として，Shizuka（2004）はthe Clustered Objective Probability Scoring（以下，COPS）という採点方式を開発した。

ここでは，以下のような架空のデータを例に，COPS得点の算出方法を説明する。

■ 表2：架空のテストにおける自信度の使用頻度および自信度ごとの正答数

	低		中		高	
	頻度	正解	頻度	正解	頻度	正解
受験者 1	4	1	3	2	3	3
受験者 2	0	-	0	-	10	6
受験者 3	10	6	0	-	0	-
受験者 4	4	2	3	3	3	1
受験者 5	4	0	0	-	6	6

（注）Shizuka（2004, p.180）を基に作成。

表2は，5人の受験者が10問からなる4択式の試験を受けた場合の，自信度の使用頻度および自信度ごとの正答数をまとめたものである。例えば，受験者1は自信度「低」を4回使いそのうち1回正解，「中」では3回中2回正解，「高」では3回中すべて正解だったことを示す。この場合，COPS得点は以下の⑴〜⑶の手順で算出する（なお，以下の手順において，変数xはそのテストにおける当て推量による正答率を小数第2位で四捨五入した値とする。例えば，表2のデータは4択式の試験に関するものであるため，当て推量による正答率は1/4 = .25である。このとき，xは.3となる）。

⑴ 受験者ごとに各自信度における正答率を計算する。例えば，表2の受験者1の場合，自信度「低」における正答率は.25，「中」では.67，「高」では1.00となる。同様に，全学習者について自信度ごとの正答率を計算する（表3）。

■ 表3：架空のテストにおける自信度ごとの正答率

	低	中	高
受験者 1	.25	.67	1.00
受験者 2	—	—	.60
受験者 3	.60	—	—
受験者 4	.50	1.00	.33
受験者 5	.00	—	1.00

（注）Shizuka（2004, p.180）を基に作成。

⑵ 各自信度における正答率を小数第2位で四捨五入した値がx以下のとき，その正答率は0として扱う。例えば表3で，受験者1の自信度「低」

における正答率は .25，受験者 4 の自信度「高」における正答率は .33である。このとき，両者を四捨五入すると .3となり，x（.03）以下である。したがって，正答率 .25と .33はともに .00に置換する（表 4 の網掛け部分）。

■表 4：架空のテストにおける自信度ごとの正答率（補正後）

	低	中	高
受験者 1	.00	.67	1.00
受験者 2	—	—	.60
受験者 3	.60	—	—
受験者 4	.50	1.00	.00
受験者 5	.00	—	1.00

なお，小数第 2 位で四捨五入して x 以下となる正答率を 0 と見なすのは，当て推量により正解した項目に対して得点が与えられるのは妥当でないと考えられるためである（Shizuka, 2004）。

(3) 各学習者について，「ある自信度を選んだ場合の正答率×その自信度を選んだ場合の正答数」を自信度ごとに計算し，それらを合計する。この合計値が，その学習者の COPS 得点である。例えば，受験者 1 は自信度「低」における正答率が .00で 1 問正解，「中」では .67で 2 問正解，「高」では1.00で 3 問正解である。このとき，受験者 1 の COPS 得点は，「.00×1 ＋ .67×2 ＋ 1.00×3」となる。同じように，全受験者に対して COPS 得点を計算する（表 5 ）。

■表 5：架空のテストにおける COPS 得点

	計算式	COPS得点
受験者 1	.00×1 ＋ .67×2 ＋ 1.00×3	4.34
受験者 2	.60×6	3.60
受験者 3	.60×6	3.60
受験者 4	.50×2 ＋ 1.00×3 ＋ .00×1	4.00
受験者 5	1.00×6	6.00

（注）Shizuka（2004, p.180）を基に作成。

以上のように，ある自信度で選ばれた正答に対して，COPS では受験者ごとに異なる得点が与えられる。例えば，自信度「高」で正解した項目 1 つにつき，受験者 1 ・ 5 は1. 00点が与えられるが，受験者 2 は.60点しか与えられない（表 5 ）。このようにすることで，COPS では学習者によって自信度の使い方が異なるという問題の解決をめざしている。

本研究では，以上の計算処理を自動化するプログラムを Microsoft Visual Basic 6.5により作成し，COPS 得点を算出した。

なお，本研究の参加者に自信度の使い方を説明する際には，Polytomous（旧）に関してのみ説明を行い，他の採点方法については言及しなかった（資料）。その理由は，自信度が高い誤答にペナルティが与えられない Dichotomous や Polytomous（新）で採点すると予告した場合，学習者が自信度を正直に申告しない可能性がある（Shizuka, 2004）と考えたからである。また，COPS 得点の計算方法は複雑であり，学習者を混乱させる可能性があると考えたため，COPS についても事前の説明は行わなかった。

2. 再生式文章完成課題

先行研究（Ishii, 2005; Jiang, 2004; Koya, 2005; Nakata, in press; Takahashi, 1984; Tanaka, 1983; 小屋, 2002など）を基に正誤を判断した。本テストは学習者の意味知識を測定するものであり，スペリングや派生形，活用形の知識を問うものではない。したがって，スペリング，派生形，活用形の誤りがある解答も正答として扱った。模範解答と一致しない解答については，筆者と日本人英語教師 1 人が独立して正誤を判断した。2 人の評価者での一致度は 99.7%であった。採点結果が一致しないものは，協議によって解決した。

4.3.2 採点方法の比較

4 つの採点方法の中でどれが最も優れているかを判断する上では，大友（1994）の枠組みにならい，(a) 信頼性，(b) 妥当性，(c) 実用性という 3 つの観点から考察する。具体的には，以下のような分析を行う。

(a) 信頼性：テスト得点の信頼性とは，「（ある）集団に対して，同様な条件のもとでテストを繰り返すとき，どのくらい一貫したテスト得点が得られるかという度合い」（大友, 1996, p.175）のことである。本研究では，古典的テスト理論および項目応答理論における信頼性係数を検討する。

(b) 妥当性：テストの妥当性とは，「そのテストが測定しようとしていることを本当に測定しているかどうかという度合い」（大友, 1994, p.300）を意味する。妥当性は 1 つの統合された概念であるが，(1) 内容的側面，(2) 本質的側面，(3) 構造的側面，(4) 一般化可能性の側面，(5) 外的側面，(6) 結果的側面という 6 つに分けてとらえられることがある（Messick, 1995。和訳は平井, 2006 による）。これらの中で，採点方法を変えることで大きく変化すると考えられるのは，構造的側面，一般化可能性の側面，外的側面の 3 点である。したがって，本研究ではこれらの 3 側面から，4 つの採点方法の妥当性を比較する。具体的には，以下のような分析を行う。

（i）構造的側面とは，「得点の内的構造が構成概念の下位領域や次元性などの理論的構造に一致していることを示す証拠」（平井, 2006, p.30）のことである。本研究では，再認式課題における得点が項目応答理論に適合しているかどうかを調べることで，構造的側面について検討する。

（ii）一般化可能性の側面とは，「得点の意味や測定論的特性（平均や標準偏差，項目間の相関構造など）が，ある特定のデータセットだけでなく他の被験者集団，実施場面，実施時期，項目セットに対しても不変であるという証拠」（平井, 2006, p.30）と定義される。本研究では，学習者の熟達度が上がるにつれて再認式課題における得点も上がるかどうかを検討することで，一般化可能性の側面について考察する。

（iii）外的側面とは，「他の変数との間に理論上想定される相関パターンが実際にも示されるという証拠」（平井, 2006, p.30）を指す。本研究では，再生式課題との相関係数を比較することで，外的側面について検討する。

(c) 実用性：テストの実用性とは，「経済性，実施と採点の容易さ，結果の解釈の容易さ」（大友, 1994, p.302）などを指す。本研究では，実施の容易さ，採点の容易さ，および実施時間によって実用性を判断する。

テスト得点の分析には，SPSS 17.0.1，Winsteps 3.45.2，R 2.8.1 を使用した。

5 結果

5.1 記述統計

5.1.1 古典的テスト理論による記述統計

まず，古典的テスト理論による記述統計を示す。再認式および再生式文章完成課題に関する記述統計は，表 6 のとおりとなった（なお，再生式課題では用紙の裏面にも問題があることに気付かず，半分以上の問題に解答しなかった受験者が 1 人いた。したがって再生式課題では，この受験者 1 人を除いた97人を分析対象とした）。

次に，再認式課題における自信度ごとの正答率を検討する。4.2.1で述べたとおり，再認式課題において，学習者は解答の自信度を 3 段階（「低」・「中」・「高」）で選ぶことが求められた。自信度を反映した採点方法が意図した効果を上げるためには，学習者が自信度を適切に使用したことが前提となる。そこで，自信度の使用方法が妥当であったかどうかを検討するために，自信度の使用状況を表 7 にまとめた。

■ 表 6：再認式および再生式文章完成課題に関する記述統計

	再認式				再生式
	D	P旧	P新	COPS	
平均値	26.78	159.21	143.37	18.85	21.32
標準偏差	4.00	20.62	26.33	6.61	4.31
95%信頼区間下限	22.02	139.00	118.24	14.96	16.54
95%信頼区間上限	31.54	179.42	168.49	22.74	26.10
クロンバック α	.63	.75	.76	.91	.68
満点	39	234	234	39	39

（注）受験者は再認式課題で 98 人，再生式課題で 97 人。
D = Dichotomous；P旧 = Polytomous（旧）；P新 = Polytomous（新）

■ 表 7：再認式文章完成課題における自信度の使用頻度および自信度ごとの正答率

	頻度	%	正答	%	誤答	%
「低」	887	23.2%	510	57.5%	377	42.5%
「中」	1047	27.4%	674	64.4%	373	35.6%
「高」	1746	45.7%	1440	82.5%	306	17.5%
「わからない」	142	3.7%				

表7は，再認式文章完成課題における自信度の使用頻度および自信度ごとの正答率を示したものである。例えば，表7の2段目（「低」の段）は，学習者が自信度「低」を合計で887回使用し，これは全解答の23.2%を占めていたことを示す。そして，学習者が自信度「低」を選択した場合，その内の57.5%（510回）は正答であり，残りの42.5%（377回）は誤答であった。

3つの自信度における平均正答率を比較すると，自信度「低」を選択した場合は57.5%，「中」を選択した場合は64.4%，「高」を選択した場合は82.5%であった。カイ二乗検定の結果，3つの自信度における正答率には有意差が見られた［x^2 (2) = 213.68, p < .001, Cramer's V = .24］。3つの自信度レベルのどこに差があるのかを調べるために，ライアン法による多重比較を行った。その結果，いずれの自信度における正答率にも有意差が見られた（p < .05）。以上の結果は，自信度とともに正答率が上がっていることを示唆するものであり，学習者が応答自信度を適切に使用したことを示している。

また，自信度「低」を選択した場合の正答率である57.5%は，当て推量による正答率である50%を上回るものであった［x^2 (1) = 19.94, p < .001, Cramer's V = .11］。つまり，自信度「低」で解答を選んだ際にも，学習者は完全な当て推量で答えていたわけではないということである。以上の結果から，本研究の参加者は自信度を適切に使用していたと考えられる。

5.1.2 項目応答理論による記述統計

次に，項目応答理論の1パラメータ・モデル（ラッシュ分析）を用いて，受験者の能力推定値を計算した。項目応答理論を用いたのは，テストの測定精度に関して古典的テスト理論よりも詳細な情報が得られるためである（大友, 1994, 1996; 小泉, 2005; 靜, 2007）。

項目応答理論は，テストの一元性（unidimensionality）を前提としている（大友, 1996; 靜, 2007）。一元性とは，(1) そのテストの項目すべてが同じ「何か」を測定している，(2) テスト項目を易しい項目から難しい項目まで，1つの直線上に（一次元上に）並べられることを指す（靜, 2007）。

学習者がどのような語意を習得しているかは，どのようなインプットに接してきたかに依存する。したがって，あらゆる学習者に共通する普遍的な語意の習得順序はないと考えられる。また，先行研究から，語によってはU字型の発達が見られることも示されている（Kellerman, 1979; Koya, 2005; Tanaka, Takahashi & Abe, 1990）。すなわち，ある用法に関しては，初級者では正答率が高く，中級者では正答率が低く，上級者では正答率が再び高くなるのである。このように，語意習得は複雑なプロセスを経るため，一元性を満たさない項目・受験者も少なからず出てくると予測される。ただし，テスト項目および受験者が多くなれば誤差は相殺され，ある程度の一元性が見られるであろう。

以上の議論を踏まえると，今回の語意テスト得点に項目応答理論を適用し，一元性を満たすかどうかを調査することは，意味があると考えられる。そこで，項目応答理論による能力推定値を求めた。なお，COPS得点から能力推定値を求める際には，各学習者のCOPS得点を小数第2位で四捨五入して10倍するという変換（Shizuka, 2004）を行った。

能力推定値に関する記述統計は，表8のようになった。

■ 表8：項目応答理論による能力推定値

	再認式				再生式
	D	P旧	P新	COPS	
平均値	1.16	0.43	0.15	-0.18	0.43
標準偏差	0.74	0.22	0.18	0.21	0.75
標準誤差の平均値	0.43	0.10	0.08	0.06	0.44
標準誤差の標準偏差	0.05	0.01	0.01	0.04	0.03
受験者信頼性	.64	.72	.75	.88	.65
受験者分離指数	1.32	1.62	1.74	2.68	1.37

（注）受験者は再認式課題で98人，再生式課題で97人。D＝Dichotomous；P旧＝Polytomous（旧）；P新＝Polytomous（新）

表8から得られる示唆については，5.2および5.3において述べる。

5.2 信頼性の比較

5.1.1において学習者が自信度を適切に選択したことが示唆されたため，次に再認式課題における4つの採点方式の比較を行う。4.3.2で述べたとおり，(a)

信頼性，(b) 妥当性，(c) 実用性という 3 つの観点から，どの採点方式が最も優れているかを検討する。

まず，信頼性に関しては，古典的テスト理論および項目応答理論における信頼性係数を比較する。古典的テスト理論における信頼性係数であるクロンバック α は，Dichotomous で.63，Polytomous (旧) で .75，Polytomous (新) で .76，COPS で .91となった（表 6)。一般的に，.7〜.8以上の信頼性があるとき，そのテスト結果には十分な内的一貫性があると解釈される (Nunnally, 1978)。したがって，Dichotomous 以外の 3 つの採点方式ではある程度の内的一貫性が確保できたと考えられる。多値的な採点方式の中では，COPS における信頼性が最も高かった。

次に，項目応答理論における信頼性について検討する。具体的には, Shizuka（1999, 2004）にならい，能力推定値の(1) 標準誤差および(2) 信頼性係数を比較する。

能力推定値の(1) 標準誤差とは，受験者の真の能力と推定値との誤差を指す（大友, 1996; 靜, 2007)。よって，この値が小さいほど，真の能力を正確に推定しているということになる。表 8 のとおり，能力推定値の標準誤差は Dichotomous で0.43，Polytomous (旧) で0.10，Polytomous (新) で0.08，COPS で0.06となった。この結果は，Dichotomous 以外の採点方式ではある程度の信頼性が確保できたことを示唆している。

次に，能力推定値の(2) 信頼性係数を比較する。項目応答理論における信頼性係数は，Dichotomous で .64，Polytomous (旧) で .72，Polytomous (新) で .75，COPS で .88となった（表 8)。この結果も，Dichotomous 以外の採点方式ではある程度の内的一貫性が確保できたことを示している。

以上の結果を総合すると，テストの信頼性に関しては，「COPS ＞ Polytomous (新) ＞ Polytomous (旧) ＞ Dichotomous」の順に優れていると考えられる。この結果は，1）自信度を得点に反映させることでより信頼性の高い測定が可能になる，2）多値的な採点方式の中では COPS の信頼性が最も高い，という 2 点を示唆するものである。

5.3　妥当性の比較

次に，4 つの採点方式の妥当性について検討する。4.3.2で述べたように，本項では(1) 構造的側面，(2) 一般化可能性の側面，(3) 外的側面の観点から比較を行う。

5.3.1　構造的側面

構造的側面に関しては，項目応答理論（大友, 1994, 1996; 靜, 2007など）への適合度を調べることで検討する。具体的には，一元性の前提を満たさないミスフィットと見なされる受験者および項目の数を比較する。ミスフィットは，Infit Mean Square の値によって判別する。どのくらいの値をミスフィットと見なすかという点に関してはさまざまな基準があるが（靜, 2007)，本研究では「Infit Mean Square が平均＋2×標準偏差以上」（小泉, 2005, p.69）をミスフィットと見なすこととする。

どのくらいのミスフィットが許容されるかという点に関しては，小泉（2005）と同じく，「テストではミスフィット項目が全体の10%未満」，「受験者ではミスフィットと判断された受験者の割合が 2 %未満」(p.69）であれば，データに一元性があると判断することとした。

表 9 は，ミスフィットと見なされた受験者およびテスト項目の割合を，採点方式別にまとめたものである。

■ 表9：各採点方式におけるミスフィットの基準値と割合 (注1)

		Infit Mean Square			Misfit の割合
		平均	標準偏差	Misfit 基準	
D	項目	1.00	0.08	1.16	5.3% (2/38)
	受験者	1.00	0.21	1.42	4.1% (4/98)
P旧	項目	1.01	0.11	1.23	7.7% (3/39)
	受験者	1.04	0.43	1.90	4.1% (4/98)
P新	項目	1.00	0.11	1.22	5.1% (2/39)
	受験者	1.03	0.29	1.61	4.1% (4/98)
COPS	項目	1.02	0.19	1.40	5.1% (2/39)
	受験者	1.09	0.53	2.15	2.0% (2/98)

(注) 受験者は 98 人。受験者＝受験者の能力；項目＝テスト項目の難易度

まず，受験者のミスフィットについて検討する。いずれの採点方式においても，ミスフィットと判断された受験者は全体の2.0%以上を占めており，一元性が十分に満たされているという証拠は得られなかった。これは，総得点が低い受験者でも難しい項

目に正解し，総得点が高い受験者でも易しい項目に誤答をするなど，応答パターンにある程度のランダム性が見られたことを示している。4 つの採点方式について比較すると，COPS においてミスフィットと判断された受験者は2.0%であり，他の採点方式（いずれも4.1%）よりも少なかった。

次に，項目のミスフィットについて検討すると，いずれの採点方式においてもミスフィット項目は全体の10%未満であり，テスト項目の難易度に関しては一元性が満たされていることが示された。4 つの採点方式を比較すると，Polytomous（旧）ではミスフィット項目が全体の7.7%を占めており，それ以外の採点方式（5.1%または5.3%）よりもやや多かった。

5.3.2 一般化可能性の側面

次に，一般化可能性の側面から 4 つの採点方式を比較する。先行研究から，学習者の総合的な熟達度と語意知識には，ゆるやかな相関があることが示されている（Ishii, 2005; Koya, 2005; Okuno, 2008; 小屋, 2002など）。英語熟達度が高くなるにつれて語意知識も豊かになるのであれば，英語熟達度が高い学習者ほど再認式課題における得点も高くなると考えられる。

4.1で述べたとおり，本調査への参加者は英語熟達度によって 3 つのグループに分けることができる。すなわち，1）A大学の学生（n = 38），2）B大学 Aレベルの学生（n = 33），3）B大学 Bレベルの学生（n = 27）の 3 グループである。英語熟達度とともに語意知識も発達すると仮定すると，再認式課題の平均得点は，「A大学＞ B大学 Aレベル＞ B大学 Bレベル」の順になるであろう。

上の仮説が 4 つの採点方式において満たされているかどうかを検証するため，それぞれの採点方式におけるグループごとの平均得点を集計した（表10）。

各グループの平均得点に有意差があるかどうかを検定するため，一元配置の分散分析およびテューキー HSD の多重比較を行った。その結果を表11に示す。

表11のとおり，いずれの採点方式においてもグループ間の差は有意であり，大きな効果量が見られた（$\eta^2 > .14$。Cohen, 1988）。多重比較の結果，採点方式にかかわらず，「A大学の学生＞ B大学 Aレベルの学生＞ B大学 Bレベルの学生」という仮説が支持された。この結果から，一般化可能性の側面に関しては，いずれの採点方式も妥当であると考えられる。

■ 表10：再認式文章完成課題における平均値と標準偏差（採点方法・レベル別）

		平均	*SD*	95%信頼区間	
				下限	上限
A大学	D	29.87	3.09	28.91	30.83
	P旧	175.58	15.32	170.78	180.37
	P新	162.95	21.30	156.62	169.27
	COPS	24.10	0.76	22.59	25.61
B大学 Aレベル	D	26.09	3.52	25.06	27.12
	P旧	155.82	17.31	150.67	160.96
	P新	140.61	21.95	133.82	147.39
	COPS	18.00	0.82	16.38	19.61
B大学 Bレベル	D	23.26	1.89	22.12	24.40
	P旧	140.33	10.34	134.65	146.02
	P新	119.19	13.09	111.68	126.69
	COPS	12.51	0.90	10.72	14.30

（注）受験者は 98 人。D ＝ Dichotomous；P旧＝ Polytomous（旧）；P新＝ Polytomous（新）；*SD* ＝標準偏差

■ 表11：3 つのグループにおける再認式文章完成課題得点の差の検定

		検定結果
ANOVA	D	$F(2, 95) = 40.20, p < .001, \eta^2 = .46$
	P旧	$F(2, 95) = 45.53, p < .001, \eta^2 = .49$
	P新	$F(2, 95) = 39.67, p < .001, \eta^2 = .46$
	COPS	$F(2, 95) = 49.15, p < .001, \eta^2 = .51$
多重比較	D	A大学＞ B大学 A レベル＞ B大学 B レベル
	P旧	A大学＞ B大学 A レベル＞ B大学 B レベル
	P新	A大学＞ B大学 A レベル＞ B大学 B レベル
	COPS	A大学＞ B大学 A レベル＞ B大学 B レベル

（注）多重比較において，「＞」は 5 %水準で有意差があったことを示す。

5.3.3 外的側面

最後に外的側面の検討を行う。具体的には，再生

式課題（4.2.2）との相関係数を比較することで，外的側面について検討する。

選択肢が提示されない再生式課題では，当て推量により正解する確率は再認式課題よりも低いと考えられる。したがって，もし自信度を点数に反映させることで当て推量の影響が軽減されるのであれば，Polytomous（旧）・（新）およびCOPSにより算出した得点の方が，Dichotomousで算出した得点よりも再生式課題における得点との相関係数が高くなるはずである。

この仮説を検証するため，4つの採点方式による得点と再生式課題得点とのピアソンの積率相関係数を計算した（表12）。

■ 表12：再認式文章完成課題と再生式文章完成課題における得点との相関係数

	相関係数	95%信頼区間	
		下限	上限
Dichotomous	.79	.70	.85
Polytomous（旧）	.78	.69	.85
Polytomous（新）	.78	.69	.85
COPS	.76	.66	.83

（注）いずれも1％水準で有意。受験者は97人。

表12のとおり，いずれの採点方式においても.76以上の有意な相関が見られた。4つの採点方式における相関係数に有意差が見られるかどうかを調べるため，それぞれの相関係数の差を*t*検定により検定した（Glass & Hopkins, 1996）。第1種の過誤を減らすために，ボンフェローニの修正を用い，有意水準は.008（.05 / 6）とした。結果は表13のようになった。

■ 表13：再認式課題と再生式課題における得点との相関係数の差の検定

	D	P旧	P新
P旧	t（94）= 0.44, p = .662		
P新	t（94）= 0.23, p = .820	t（94）= -0.31, p = .760	
COPS	t（94）= 1.25, p = .215	t（94）= 0.71, p = .477	t（94）= 0.79, p = .433

（注）D＝Dichotomous；P旧＝Polytomous（旧）；P新＝Polytomous（新）

表13のとおり，いずれの採点方式の間にも有意な差は見られなかった。この結果から，再生式課題における得点との相関という観点ではいずれの採点方式も妥当であり，採点方式による違いは見られないと考えられる。

5.3.1〜5.3.3の分析結果をまとめると，次のようになる。(1) 妥当性の構造的側面に関しては，COPSが最も優れている，(2) 妥当性の一般化可能性および外的側面に関しては，4つの採点方式の間に差はない。

5.4　実用性の比較

最後に，4つの採点方式の実用性について検討する。大友（1994）にならい，採点方法ごとの実施の容易さ，採点の容易さ，および実施時間を比較する。

5.4.1　実施の容易さ

実施の容易さに関しては，Dichotomousが最も優れていると考えられる。自信度を反映させる採点方式では，解答の選択肢に加えて自信度も選択することが求められるため，Dichotomousよりも受験者の負担が大きい（Kamimoto, 2006）。また，Dichotomous以外の採点方式では，自信度の選び方や採点方法について受験者に説明する必要があるため，テスト実施者の負担も大きくなる。以上の理由から，受験者および実施者の双方にとって，Dichotomousの方がそれ以外の採点方法よりも実用的であると考えられる。

なお，実施の容易さには実施時間も含まれるが，実施時間については5.4.3で検討するため，本項では扱わない。

5.4.2　採点の容易さ

採点の容易さに関しては，「Dichotomous ＞ Polytomous（新）＞ Polytomous（旧）＞ COPS」の順に優れていると考えられる。

まず，Dichotomousでは，解答の正誤のみで得点を計算することができる。一方で，Dichotomous以外の採点方式では，自信度によって異なる点数が与えられるため，採点作業により多くの時間と労力が必要である。以上の理由から，Dichotomousによる採点が最も容易であると考えられる。

最も採点作業が複雑なのは，COPSであろう。COPS得点を算出するには，自信度ごとの正答率を

計算する必要があるためである（4.3.1を参照）。本研究ではCOPS得点を計算する自作プログラムを使用したため，得点の計算自体にそれ程の手間はかからなかった。しかし，データをコンピュータに入力するのに相応の時間が必要であった。

Polytomous（旧）およびPolytomous（新）による採点の容易さは，DichotomousとCOPSの中間に位置するであろう。Polytomous（旧）・（新）では，COPSのように自信度ごとの正答率を計算する必要はないが，Dichotomousとは異なり，自信度を得点に反映させる必要があるためである。Polytomous（旧）・（新）を比較すると，後者の方が採点が容易であると考えられる。Polytomous（旧）では誤答でも自信度に応じて点数が与えられるのに対して，Polytomous（新）では誤答は一律0点として扱えばよいからである。

以上の理由から，採点の容易さに関しては，「Dichotomous ＞ Polytomous（新）＞ Polytomous（旧）＞ COPS」の順に優れていると考えられる。

5.4.3 実施時間

最後に，採点方式の違いがテストの実施時間に与える影響について検討する。PolytomousおよびCOPSでは，学習者は解答の自信度を示すことが求められるため，その分テストの実施時間が長くなる（Kamimoto, 2006）。ゆえに，実施時間を考慮すると，Dichotomousの方が多値的な採点方式よりも実用的と言えるだろう。

それでは，自信度を学習者に尋ねることで，テストの実施時間はどのくらい長くなるのであろうか。本実験では紙上でテストを実施したため，厳密な解答時間を測定することはできなかった。そこで，自信尺度を導入することで試験時間がどのくらい長くなるかを調査するために，追実験としてコンピュータ上で再認式文章完成課題を実施した。

追実験の詳細は以下のとおりである。

1）参加者

調査対象者は5名の日本人英語学習者であった。調査は被験者ごとに個別に実施した。

2）手続き

追実験の参加者は，再認式課題をコンピュータ上で受験した。コンピュータ版のテストは，Microsoft Visual Basic 6.5を用いて筆者が自作した。コンピュータ版テストでは紙版と同じく，二者択一形式で答えの入力が求められた（図4）。

▶ 図4：コンピュータ版再認式文章完成課題の画面（解答の選択）

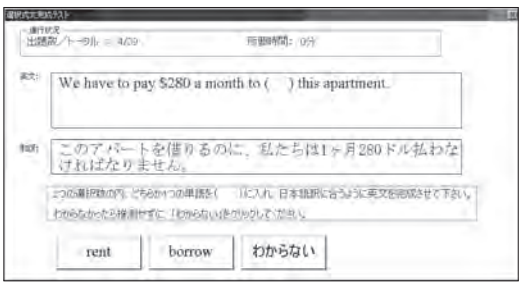
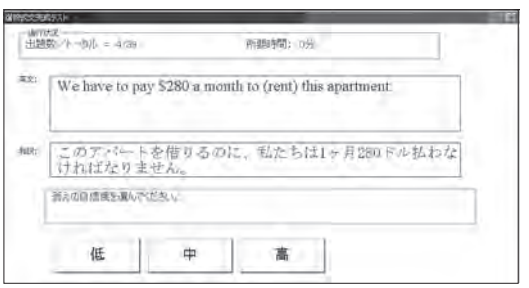


図4の画面で「わからない」以外のボタンが押された場合，自信度を3段階で選択することが求められた（図5）。

▶ 図5：コンピュータ版再認式文章完成課題の画面（自信度の選択）



コンピュータ版テストでは，解答の選択にかかった時間（図4）および自信度の選択にかかった時間（図5）が自動的に記録された。それ以外の点は，紙版と同一であった。

なお，厳密な反応時間を測定する際には，SuperLabやDMDXなどの専用ソフトおよび入力装置を使用することが望ましい。しかし，今回の調査はミリ秒単位の反応時間を比較する心理言語学実験ではないため，自作のプログラムでも問題はないと判断した。入力装置にはマウスを使用した。

3）結果

コンピュータ版テストにより測定した1問あたりの平均所要時間（括弧内は標準偏差）は，解答を決めるまでにかかった時間が6.21秒（1.29），自信度を決めるまでにかかった時間が1.36秒（0.50）であった。この結果は，多値的な採点方式を使用することで，実施時間がDichotomousと比較して約22.0％長くなることを示唆するものである。

なお，追実験参加者のコンピュータ版再認課題における得点（括弧内は標準偏差）をまとめると，Dichotomous で29.80（4.45），Polytomous（旧）で177.60（24.94），Polytomous（新）で172.20（24.63），COPS で24.19（6.40）であった。本実験の参加者と比較すると，追実験の参加者の得点の方が若干高かったことがわかる。

5.4.1〜5.4.3の分析結果を総合すると，実施の容易さ，採点の容易さ，および実施時間のいずれに関しても，Dichotomous の方がそれ以外の採点方法よりも優れていると考えられる。

6 結論

本研究では，(a) 信頼性，(b) 妥当性，(c) 実用性という3つの観点から，1）Dichotomous，2）Polytomous（旧），3）Polytomous（新），4）COPS という 4 つの採点方法を比較・検討した。調査の結果，自信度を語意テスト得点に反映することで，信頼性と妥当性の構造的側面に関しては改善される可能性が示唆された。しかし，妥当性の一般化可能性および外的側面に関しては，有意な差は見られなかった。また，テストの実用性に関しては，自信度を用いない従来の採点方式の方が優れていた。今回の研究結果は，多肢選択式の読解（Shizuka, 1999, 2004）および語彙サイズテスト（Kamimoto, 2006）に関する先行研究の結果とほぼ一致するものであった。

次に，本研究から得られる示唆について述べる。本研究では，自信度を反映した多値的な採点方法は妥当性の大幅な改善にはつながらず，さらに実用性では従来の採点方法に劣ることが示された。したがって，語意テストにおいて自信度を用いることは必ずしも必要ではないと考えられる。しかし，実用性を多少犠牲にしたとしても高い信頼性が望ましい場合には，自信度を採点に取り入れることが 1 つの選択肢になり得るだろう。

多値的な採点方法の中では，信頼性の高さやミスフィットの少なさから COPS が最も望ましいと考えられる。ただし，COPS は他の採点方法と比較して，採点により多くの手間がかかるという欠点もある。したがって，COPS 得点の計算に必要な人的および時間的な資源がある場合のみ，COPS による採点を行うのが現実的であろう。

実用的な理由から COPS を使用することが困難である場合は，COPS の次に信頼性が高くミスフィットが少ない Polytomous（新）を使用するのが好ましいと考えられる。ただし，Polytomous（新）には 1 つの大きな欠点がある。それは，誤答は自信度にかかわらず一律 0 点として扱われるため，常に自信度「高」を選んでいれば最高の得点が得られてしまうという点である（Shizuka, 2004）。この欠点を補うために，試験前には Polytomous（旧）で採点をすると予告し，実際には Polytomous（新）で採点をするという方法（Shizuka, 2004）が考えられる。

ただし，事前の説明と異なる方法で採点を行うことは，倫理的・実用的な理由から望ましくない場合もあるであろう（Shizuka, 2004）。その際には，信頼性や妥当性の構造的側面という点では少し劣るものの，Polytomous（旧）を採用することが現実的であると思われる。

最後に，本研究の限界並びに今後の課題を述べる。第 1 に，本研究では項目分析は行わず，全体得点のみを分析した。より測定精度の高い語意テストを開発するためには項目分析を行い，どのような項目でミスフィットが多く，その理由は何であるかを検討することが必要であろう。

また，項目分析は語意指導に関する有益な示唆にもつながると考えられる。例えば，先行研究では take / make の軽動詞的用法は上級者でも誤用が多いという結果が得られているが（Koya, 2005; Tanaka, Takahashi, & Abe, 1990; 小屋, 2002など），本調査でも軽動詞的用法の正答率は低かった(注2)。また，先行研究で報告されているとおり（Tanaka & Abe, 1985; Takahashi, 1984），本研究でも「bring ＋生物」の用法（例：When you visit us next time, please *bring* your children. Takahashi, 1984, p.142）は，「bring ＋無生物」の用法（例：When you visit us, please *bring* a bottle of wine with you. Tanaka & Abe, 1985, p.115）よりも正答率が低かった。このような項目分析を行うことで多くの学習者に共通する傾向が明らかになり，より効果的な語意指導への具体的な示唆を得ることが可能になるであろう。

第 2 に，当初の計画では正答率・自信度に加えて，反応時間も得点に反映する予定であった。同じ正答であっても，語意知識のある学習者ほど反応時間が短くなり，反応時間によって語意力のある学習者と

そうでない学習者とを弁別できると考えたからである。しかし，今回の研究では，反応時間を得点に反映させることの効果については残念ながら検証することができなかった。この点については，今後の課題としていきたい。

第3に，本研究における追実験参加者の語意テスト得点は，本実験参加者のそれよりも高かった(5.4.3)。したがって，追実験の参加者に関して得られた結果は，本実験の参加者には必ずしも一般化できない可能性があることに留意が必要である。

謝　辞

この研究の機会をくださった（財）日本英語検定協会と選考委員の先生方，特に貴重なアドバイスをくださった大友賢二先生に厚く御礼申し上げます。また，ご指導をいただきました岡秀夫先生に感謝いたします。

さらに，研究にご協力くださった友田路さん，Hywel Evans 先生および研究参加者の皆様に感謝申し上げます。

注

(1) Dichotomous ではすべての受験者が正解した項目が1つあった。項目応答理論では全受験者が正解した項目の難易度を計算することはできない（大友，1996）ため，Dichotomous においてはこの項目を除いた38項目のみを分析対象とした。したがって，Dichotomous におけるミスフィット項目の割合は，5.1％（2/39）ではなく5.3％（2/38）となる。

(2) 軽動詞的用法とは，動詞が本来の意味をほとんど持たずに使われている用法のことである（Koya, 2005）。例えば，take / make の軽動詞的用法には，take a shot, take an approach, take responsibility, make a copy, make contact, make an appointment などがある。

参考文献（＊は引用文献）

＊Cohen, J.(1988). *Statistical power analysis for the behavioral sciences*(2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

＊Csabi, S.(2004). A cognitive linguistic view of polysemy in English and its implications for teaching. In M. Achard & S. Niemeier(Eds.), *Cognitive linguistics, second language acquisition, and foreign language teaching*(pp.233-256). New York: Mouton de Gruyter.

＊大学英語教育学会(JACET)基本語改訂委員会(編).(2003).『大学英語教育学会基本語リスト』. 東京：大学英語教育学会.

＊Glass, G.V. & Hopkins, K.D.(1996). *Statistical methods in education and psychology*(3rd ed.). Boston, MA: Allyn & Bacon.

＊平井洋子.(2006).「測定の妥当性からみた尺度構成―得点の解釈を保証できますか―」. 吉田寿夫(編著).『心理学研究法の新しいかたち』.(pp.21-49). 東京：誠信書房.

＊今井むつみ.(1993).「外国語学習者の語彙学習における問題点―意味表象の見地から―」.『教育心理学研究』, 41, 243-253.

＊Ishii, T.(2005). *Diagnostic test of vocabulary knowledge for Japanese learners of English*. Unpublished PhD thesis, University of Nottingham, England.

＊Ishii, T. & Schmitt, N.(2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, 40, 5-22.

＊Iso, T. & Aizawa, K.(2008). Revisiting learners' vocabulary size estimation: The effects of randomization and confidence. *KATE Bulletin*, 22, 13-22.

＊Iso, T. & Aizawa, K.(2009). The interrelationship among word frequency, learner behavior in a vocabulary size test, and teachers' perception of difficult words. *ARELE*, 20, 141-150.

＊Jiang, N.(2004). Semantic transfer and development in adult L2 vocabulary acquisition. In P. Bogaards & B. Laufer(Eds.), *Vocabulary in a second language: Selection, acquisition, and testing*(pp.101-126). Amsterdam: Benjamins.

＊Kamimoto, T.(2006). Guessing and vocabulary tests: Looking at Vocabulary Levels Test. JACET 英語語彙研究会第3回大会口頭発表.

＊Katagiri, K.(2002). The ten-minute vocabulary tests for quick and approximate estimates of general English ability of Japanese EFL learners IV. *JLTA Journal*, 5, 111-128.

＊Kellerman, E.(1979). Transfer and non-transfer: Where we are now. *Studies in Second Language Acquisition*, 2, 37-57.

＊小泉利恵.(2005).「日本人中高生における発表語彙知識の広さと深さの関係」. *STEP BULLETIN*, vol.17, 63-80.

＊Koya, T.(2005). *The acquisition of basic collocations by Japanese learners of English*. 早稲田大学大学院未公刊博士論文.

＊小屋多恵子.(2002).「日本人学習者の英語コロケーション能力に与える日本語の影響」.『日本実用英語

学会論叢』, 10, 63-77.
＊Maingay, S. & Rundell, M.（1987）. Anticipating learners' errors: Implications for dictionary writers. In A. Cowie（Ed.）, *The dictionary and the language learner*（pp.128-135）. Tubingen: Max Niemeyer Verlag.
＊松田文子.（2004）.『日本語複合動詞の習得研究―認知意味論による意味分析を通して―』. 東京：ひつじ書房.
＊Messick, S.（1995）. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
＊水口里香.（2002）.「類義語の使い分けにおけるメタ言語知識の役割」.『日本文化学報』, 14, 259-277.
＊Mochizuki, M.（2006）. Acquisition of different senses of prepositions and its implications for lexicography. JACET 英語辞書研究会（編）. *English Lexicography in Japan*（pp.262-273）. 東京：大修館書店.
＊Nakata, T.（in press）. The effects of positive evidence and metalinguistic information on L2 lexico-semantic development. *JACET Journal*.
＊中田達也.（2009）.「第二言語語意習得における肯定証拠の役割―プロトタイプ・項目学習・体系学習の観点から―」. 東京大学外国語教育学研究会（編）.『外国語教育学研究のフロンティア：四技能から異文化理解まで』.（pp. 2-14）. 東京：成美堂.
＊Nunnally, J.C.（1978）. *Psychometric theory*（2nd ed.）. New York : McGraw-Hill.
＊Okuno, M.（2008）. *The relationship between the vocabulary size and the knowledge of lexical choice of Japanese learners of English*. 東京大学大学院未公刊修士論文.
＊大友賢二.（1994）.「言語テストと第二言語習得」. 小池生夫（監）・SLA 研究会（編）.『第二言語習得研究に基づく最新の英語教育』.（pp.300-312）. 東京：大修館書店.
＊大友賢二.（1996）.『項目応答理論入門―言語テスト・データの新しい分析法―』. 東京：大修館書店.
＊Shizuka, T.（1999）. Combining response correctness and confidence level rating to produce polychotomous data from dichotomous items.『高専教育』, 22, 243-252.
＊Shizuka, T.（2004）. *New horizons in computerized testing of reading*. 大阪：関西大学出版部.
＊靜哲人.（2007）.『基礎から深く理解するラッシュモデリング―項目応答理論とは似て非なる測定のパラダイム―』. 大阪：関西大学出版部.
＊田頭憲二.（2007）.「日本人英語学習者の L2意味範疇の再構築における L1語彙特性の役割」.『広島大学大学院教育学研究科紀要．第二部，文化教育開発関連領域』, 56, 147-154.
＊Takahashi, T.（1984）. *A study on lexico-semantic transfer*. Ann Arbor: University Microfilms International.
＊Tanaka, S.（1983）. *Language transfer as a constraint on lexico-semantic development in adults learning a second language in acquisition-poor environments*. Ann Arbor: University Microfilms International.
＊Tanaka, S. & Abe, H.（1985）. Conditions on interlingual semantic transfer. In P. Larson & E.L. Judd & D.S. Messershmitt（Eds.）, *On TESOL '84: A brave new world for TESOL*（pp.101-120）. Washington, DC: TESOL.
＊Tanaka, S., Takahashi, T. & Abe, H.（1990）. Acquisition of the lexeme MAKE by Japanese learners of English.『英語英文学新潮』, 5, 406-422.
＊横川博一（編著）.（2006）.『日本人英語学習者の英単語親密度　文字編―教育・研究のための第二言語データベース―』. 東京：くろしお出版.

資　料　再認式文章完成課題の抜粋

答え方について

① (A)・(B) どちらかの単語を下線部に入れ，日本語訳に合うように英文を完成させてください。選択肢 (A) または (B) のいずれかに○をつけて答えてください。
答えがわからない場合は，「わからない」に○をつけてください。

＜例＞

	英文	(A)	(B)	わからない	自信度
1	He can _____ the piano. 彼はピアノを**演奏**できます。	play	sing	わからない	低　中　高
2	I want to _____ the USA. 私はアメリカを**訪問し**たいです。	come	visit	わからない	低　中　高

② (A) か (B) のいずれかに○をしたら，答えの自信度を「低」・「中」・「高」の 3 つから選んでください。

「低」	答えに自信がない。
「中」	「低」と「高」の中間。
「高」	答えに自信がある。

「わからない」を選んだ場合は，自信度を選択する必要はありません。

得点は以下のように計算されます。

	自信度		
	「低」	「中」	「高」
正解	＋ 1 点	＋ 2 点	＋ 3 点
不正解	－ 1 点	－ 2 点	－ 3 点
「わからない」	0 点		

＊ 自分の答えに**自信がある**場合は，自信度「高」を選択します。自信度「高」で正解すれば，最高の得点 (3 点) につながります。
＊ 自分の答えに**自信がない**場合は，自信度「低」や「中」を選択します。自信がないのに自信度「高」を選択して不正解だと，表のようにペナルティで点数が大きく引かれてしまう (－ 3 点) 可能性があります。
＊ 答えが**わからない**場合は，「わからない」を選択してください。「わからない」を選ぶとその問題に関して点数は入りませんが，ペナルティで点数が引かれることもありません (「わからない」を選んだ場合は，自信度を選択する必要はありません)。

＜1番易しかった問題＞

英文	(A)	(B)	わからない	自信度
_____ a picture 写真を撮る	take	make	わからない	低　中　高

模範解答：take

＜平均的な難易度の問題＞

英文	(A)	(B)	わからない	自信度
You can _____ CDs at that store for only 50 yen. あの店では CD をたった50円で借りることができます。	rent	borrow	わからない	低　中　高

模範解答：rent

＜1番難しいかった問題＞

英文	(A)	(B)	わからない	自信度
The snow lay 3 feet _____ on the street. 雪が道に 3 フィートの高さに積もりました。	high	deep	わからない	低　中　高

模範解答：deep

(注) Takahashi (1984) を基に作成。