英語能力テストに関する研究

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

茨城県／筑波大学大学院博士課程在籍　小泉 利恵

概要 本研究の目的は，コミュニケーション能力の一部である発表語彙知識に焦点を当て，発表語彙知識の広さ（中核的な意味を知っている単語がどのくらいあるか）と深さ（ある単語の１つの意味に加え，連想・接辞の知識など他の側面をどの程度知っているか）の２つの関係がどの程度あるかを調べることである。結果は，発表語彙知識の広さの3000語レベルまでの幅広い学習者層で見ると，広さと深さの関係は強く，広さレベル（1000語単位）ごとに見ると，中程度の関係があった。これらの点から，発表語彙知識を測定する際，対象者の発表語彙知識の広さの範囲が幅広い場合には，広さから深さがある程度予測できるが，範囲が狭い場合には，広さと深さの両方を測る必要があることなどが示唆された。

1 研究の背景

現在の英語科の学習指導要領（文部省，1999a, 1999b）では，コミュニケーション能力の育成が大きな目標である。コミュニケーション能力を育成するためには，コミュニケーション能力とは何か，どのような知識・能力からなるか，各構成要素はどのように関係しており，どのように習得されるかを知ることが重要である。それらを知ることにより，コミュニケーション能力とその発達を包括的にとらえながら，長期的な視野で指導・評価を行うことが可能になるからである。

コミュニケーション能力のモデルは Canale & Swain (1980)，Bachman & Palmer (1996) などさまざまあるが，どのモデルにおいても，ある一定の役割を占めているのが語彙知識である。また，語彙知識は実際にコミュニケーションをする際にも重要である（野呂＆島本，2003など）。

語彙知識は「広さ」と「深さ」に分けることができる（中西＆島本，2003など）。広さとは，語彙サイズとも言われる量的な側面で，「中核的な意味を知っている単語がどのくらいあるか」である。深さとは，質的な側面で，「ある単語について，１つの意味を知っているだけでなく，その語をどの程度知っているか」である（中西＆島本，2003）。深さに含まれる例としては，語の使用頻度・連想・接辞の知識・統語的特徴などがある（Nation, 2001）。広さだけでなく深さの点を見ることで，語彙知識をより体系的に見ることが可能になる。

では，(1) 広さと深さははどのように関係しており，(2) コミュニケーション能力の中で，これら２つはどのような役割を占めているのだろうか。

本研究では，２つの問いの１つ目を扱う。特に，語彙知識の中でもスピーキングとライティングで用いられる「発表語彙知識」に焦点を当て，中学生・高校生における，発表語彙知識の「広さ」と「深さ」の関係を吟味することを目的とする。

2 先行研究

語彙知識は聞く・読む際に使う「受容」の側面と，話す・書く際に使う「発表」の側面に分けられる（Nation, 2001; Schmitt, 2000など）。また，受容語彙と発表語彙は明確に分けることはできず，連続帯の端と端に存在すると考えられる（Melka, 1997）。過去の研究では，受容・発表という用語の定義が研究者によって異なり問題があった（Read, 2000）。そこで Laufer, Elder, Hill, & Congdon（2004）は，語彙知

識を２つの観点から分類し，より明確な定義を提案した（表１）。第１の観点は，語彙知識を使用する際に，何を取り出すのか（意味から形式を取り出すか，形式から意味を取り出すか）で，第２の観点は，形式・意味の取り出し方（再生か認知か）である。

例えば，書く際に取り出すのは，語彙知識の形式であり，その方法は再生である（例：「私は英語が好き」と書くとき，意味「英語」から形式 “English” を再生する）。この分類法は明確であり，評価できる。

■ 表１：語彙知識の種類

	再生（recall）	再認（recognition）
形式の取り出し	テスト問題の例） Turn into water *m*　　（答え：melt）	例）*Turn into water* a. elect　b. blame c. melt*　d. threaten
意味の取り出し	例）When something *melts* it turns into （答え：water）	例）*Melt* a. choose　b. accuse c. make threats d. turn into water*

（注）Laufer et al.（2004）を修正。*は答えを表す。

この２つの観点に基づき，第二言語の語彙知識の広さと深さの関係を調べた研究を分類すると表２になる。

表２の先行研究での相関係数を見ると，広さと深さの関係は，ほぼないものから強いものまでさまざまである。この理由としては，(1) 取り出す対象とその方法（再生か認知か）の組み合わせが異なること，(2) 各研究で使用したテストの妥当性が異なること，の２点が考えられる。(1) については，組み合わせと関係の強さに一定のパターンが見られないことから，(2) の可能性があると思われる。

例えば，妥当性の重要な１局面である信頼性について言及している研究は，12研究中５つのみ（Aizawa, 2005; Nurweni & Read, 1999; Qian, 1999, 2002; 島田，2004）で，他の研究には対象者の得点分布が狭くテストの信頼性が低い可能性があるものも見られる。なお，Bachman（1990）によると，信頼性はテストの質だけでなく，受験者の回答によっても変化するため，研究対象者の得点の信頼性は常に推定する必要がある（p. 171）。本研究では，(2) の

■ 表２：広さと深さの関係を調べた先行研究と本研究

	広さ		深さ		相関	対象者（人数）
	取り出す対象	再生か認知か	取り出す対象	再生か認知か		
Schmitt & Meara（1997）	形式	認知	接尾辞・連想	認知・再生	.27 to .62	高大（88）
Nurweni & Read（1999）	意味	再生	P 知識・S 知識	認知	.18 to .81	大（324）
Qian（1999）	形式	認知	P 知識・S 知識	認知	.82	大（74）
Mochizuki & Aizawa（2000）	形式	認知	接辞	認知	.54 to .65	高大（403）
Shimamoto（2000）	形式	認知	P 知識・S 知識	認知	.15 to .81	大（134）
	意味	再生	P 知識・S 知識	認知	.55 to .88	
Noro（2002）	形式	認知	P 知識・S 知識・形態素	認知	.36	大（90）
Qian（2002）	形式	認知	P 知識・S 知識・同意語	認知	.70 to .80	大（217）
福地（2004）	形式	認知	P 知識・S 知識	認知	-.02 to .63	高（128）
島田（2004）	意味	認知	P 知識・S 知識	認知	.47 to .63	大（91）
Aizawa（2005）	形式	認知	多義語・接辞	認知	.54 to .55	大（229）
Akase（2005）	形式	認知	P 知識・S 知識・接辞	認知	.41 to .66	大（86）
Shimamoto（2005）	形式	認知	P 知識・S 知識	再生	.32 to .47	大（93）
本研究	形式	再生	派生語・反意語他	再生		中高（909）

（注）中 = 中学生；高 = 高校生；大 = 大学生；P 知識 = パラディグマティックな知識（例：cute の同意語を知っている）；S 知識 = シンタグマティックな知識（例：cat につながる語を知っている）

妥当性への対策として，対象者を幅広く集め，使用するテストの妥当性検証を行った後に，広さと深さの関係を調べる。

また先行研究では，検討されていない点が２点ある。第１に，広さと深さの測定をともに「再生」の組み合わせで行った研究は見られない。第２に，対象者がすべて高校生以上であり，中学生を含めた研究はない。語彙知識の構成要素間の関係を，学習初期の段階から包括的にとらえるためには，広さと深さの測り方をさまざまな組み合わせで調べ，かつ，中学生を含めた幅広い対象者で吟味することが必要だろう。したがって，本研究では先行研究で検討されていない上記２点を扱う。さらに，発表語彙知識の広さのレベルごとに，発表語彙知識の広さと深さの関係がどの程度強いかを調べることで，語彙知識や心的辞書の構造モデルや語彙習得モデルの構築に向けた第一歩となると考えられる。

3 研究目的・定義・特徴

本研究の目的は，以下の２点である。

(1) 発表語彙知識の広さと深さの関係を調べること
(2) 広さと深さの関係について，広さレベルごとに検討すること

本研究では，先行研究で見られる広さと深さの関係の強さに幅があり仮説が作りにくいために，以下のリサーチクエスチョン（RQ）を挙げる。

RQ１：発表語彙知識の広さと深さの関係はどの程度あるか
RQ２：広さが大きくなるにつれて，広さと深さの関係は変化するか

なお，本研究では，Laufer et al.（2004）と中西＆島本（2003）に基づき，以下のように定義する。

(a) 広さ（語彙サイズ）：中核的な意味と関連付けて，書記形式を筆記できる語がどのくらい多くあるか
(b) 深さ：書記形式と関連付けて，派生語・反意語・コロケーションを筆記できる語がどのくらい多くあるか。なお，Read（2004）は深さを①意味の正確さ，②包括的語彙知識，③ネットワーク的知識に分けたが，本研究ではそのうちの②と③を扱う。
(c) 発表語彙知識：語彙の形式を再生する知識。ある意味に対する形式を再生するものと，派生語・反意語・コロケーションについて再生するものとが含まれる。なお，本研究では文脈のない場合の再生は「知識」の中に属し，「知識」と文脈のある言語運用は異なるものと考える。

本研究の特徴は以下３点である。

(x) 研究目的に関して，先行研究で行われていない，発表語彙知識（形式の再生）での広さ・深さの関係を扱い，構造方程式モデリングと相関を用いて検討する点
(y) 対象者に関して，先行研究で対象とされていない中学生を含め，多くの中学生・高校生を対象者とする点
(z) テスト作成に関して，本研究が扱う発表語彙知識の広さと深さを測るテストがまだ開発されていないことから，テストを独自に作成し，その妥当性の確認をさまざまな観点から行う点

4 研究方法

4.1 参加者

茨城県・東京都・岐阜県・広島県の公立・国立の中学生・高校生913人（中３：７校330人；高１～３：８校583人）。幅広い英語能力の英語学習者を対象とした。当初は1066人の参加者を得たが，海外滞在期間が１年以上の者，白紙だったセクションがあった者，問題を完全に勘違いした者は除いた。

なお，人数については Cohen（1992）に基づき，検定力を .80，有意水準を .05とし，中程度の相関関係（r = .30）があると仮定した場合，有意になるのに必要な85人（p. 158）を，レベルごとに分けたときにも超えるように意図した。

4.2 使用テスト

4.2.1 発表語彙知識テスト

発表語彙知識の定義により，すべて形式を再生させる方法を用いた。なお，以降はテスト問題の１問１問を「項目」と呼ぶ。

4.2.1.1　発表語彙知識の広さテスト

日本語の意味から英語の形式を引き出すテスト方法を用いた（資料１）。意図した語を引き出し，同意語を排除する（Laufer et al., 2004）ため，最初の１文字を提示した。Laufer らは英語（L2）で意味を提示したが，本研究では日本語（L1）で意味を提示した。Laufer & Nation（1995, 1999），Meara & Fitzpatrick（2000）を使用しなかった理由は，難易度が高いために，日本人中高生には適さず，文を理解できるかなどの意図しない能力も測ってしまう（Koizumi, 2003）などの問題点があるためである。

作成手順は２段階に分けられる。第１に，JACET8000（JACET 基本語改訂委員会，2003）の Level 1〜3 の中から，各レベル32語をランダムに選択した。それに加え，他の調査のために別語12語を加えたが，本稿では言及しない。第２に，選択した語の第１義を英英辞典（Summers et al., 2003，他２冊）で調べ，英和辞典（小西＆南出，2002，他２冊）を見てわかりやすい訳をつけた。正解が２つ以上あるなど問題があった場合は，語を再度ランダムに選んで作成した。広さ（語彙サイズ）は，全体の中で正解できた数の割合から推定した（Laufer, 1992）。

4.2.1.2　発表語彙知識の深さテスト

作問が可能で，採点が難しすぎない深さの要素３つを選んだ。各セクションとも，広さテストと共通の語と異なる語の両方を入れた。

① 派生語セクション（20問）

対象語に派生接尾辞を付加して派生語を作る項目（例：work から worker を書く；12問）と，対象語から派生接尾辞を取って逆成語を作る項目（例：comfortable から comfort を書く；８問）を出題した。Bauer & Nation（1993）のリストの Level ２〜４の派生接尾辞の中から，より多くの派生接尾辞が使われるように配慮した。また，「テストに出題する語」と「その派生語または逆成語」の両方が JACET8000 の Level 1〜3 に入るようにした。テスト時間が限られていたために Schmitt & Meara（1997）の方法はとらず，本研究では品詞を指定することで，１語につき派生語・逆成語を１語引き出そうとした（資料２(a)）。

② 反意語セクション（18問）

反意語は，(1) 相補関係（例：alive and dead），(2) 反対関係（例：buy and sell），(3) 反義関係（例：hungry and full），(4) 非両立性（例：spring, summer, fall, and winter）の４種類に分けられる（望月，2003, p. 55）。ここでは，(1)〜(3) を出題した。(4) 非両立性を出題しなかった理由は，中高生にとって難しいと予想されたためである。答えの候補の１つが最低3000語内のものを出題した。分析時に同綴異義語の like を削除し，17問で分析した。

なお，反意語とコロケーションのセクション両方とも，Shimamoto（2005）に基づき作成したが，ある語に関連する語を「できるだけ多く」書く Shimamoto（2005）の方法に対し，本研究では，「１つ」書くよう求めた（資料２(b)）。その理由は，第１に，受験者が最も適していると判断する連想を知ることができ，第２に，ある語の深さを見る際に，たくさん書かせることで，中核的な意味に対する形式を書くことも増えるため，広さの側面を測る可能性が増えるという問題をより少なくでき，第３に，実施時間がより短くて済むという利点があるためだった。

③ コロケーション・セクション（18問）

コロケーションは「偶然を超えた確率で互いに共起する語の連鎖」（横川＆島本，2003, p. 246）である。コロケーションは，(1) 文法的・統語的コロケーション（例：access to）と (2) 意味的・語彙的コロケーション（例：spend money）と (3) 文法的・意味的パターンを持たないコロケーション（例：[at でなく] on Monday）の３つに分けられる（Schmitt, 2000, p. 77）。本研究では (2) のみを対象とした。また，Schmitt（2000）はコロケーションの定義に (a) 共起することと，(b) さまざまな程度の排他性があることを挙げているが，本テストでは，(a) のみでもコロケーションと考え，より広い範囲の組み合わせを認めた。さらに，意味の透明度が高くても低くてもコロケーションと考えた。使用したテスト方法は，Shimamoto（2005）に基づき，ある語につながる語を，品詞を指定して引き出した（資料２(c)）。

4.2.2　発表語彙知識テストに関するアンケート

海外経験などの情報を得るためと，テストの表面的妥当性などの検討のために実施した（資料３）。回答には，Cumming, Grant, Mulcahy-Ernt, & Powers

(2004) に基づき「わからない」という選択肢も設けた。その理由は，テストの妥当性などをあまり意識していない場合もあるため，答えをどれかに選択させるよりも，「わからない」という選択肢を設けたほうがよいと考えたためだった。

4.3　手順

テストは，英語授業中または放課後に実施した。広さテストは25分間で行い，終了しても先に進めないようにした。深さテストは20分間で行い，前の広さテストには戻れないようにした。これは，広さテストの答えの一部が，後の深さテストに提示されているためだった。

実施後に，広さテストに関して，同じスペルで始まり意味がほぼ同じ語が存在した12語を除外した。Level 1〜3に属する数をそろえるために，さらに6語を，ランダムに選んだ順番の最後のほうから除外し，各レベル26問，全体で78問にした。広さ（語彙サイズ）の推定は，正答数 / 78×3000の式で算出した。

その後，次に述べる3段階の分析を行った。第1に，高い信頼性を確保するために，4.3.1節の方法で採点した。第2に，作成したテストの妥当性検証を4.3.2節の方法で行った。第3に，発表語彙知識の広さと深さの関係について，構造方程式モデリングとピアソンの積率相関を用い，全体と広さレベルごとに検討した。分析では SPSS（SPSS, 2003），Amos（Arbuckle, 1999），Facets（Linacre, 2003）を使用した。

4.3.1　採点方法

すべての項目が1題1点だった。広さテストについては，完全に正しい綴りと小西＆南出（2002）に載っている綴り（例：socks / sox, work / wark）を正解とした。

深さテストについては，Shimamoto（2005）に基づき，以下の手順で行った。まず，辞書（小西＆南出，2002，他7冊）を参照し可能性のあるすべての品詞・語義について模範解答を作り，採点した。模範解答にない答えは，小西＆南出（2002）で語として載っているかを確認し，載っていない語は誤りとした。載っている語については，派生語の採点では，指定した品詞と意味が同じもののみ正解とした。

反意語・コロケーションのセクションの採点では，語として載っていて，答えと完全に異なる解答以外は3人の評価者（英語母語話者2人と日本人英語上級者1人）に評価を依頼した。評価者は回答が適切かどうかを判断し，間違っている，容認可能でない，冗長だ，変だと感じた場合は0点とした。特定の文脈（おとぎ話など）のみで適切な場合は誤りと判定した。

3人の評価者での一致度と内的一貫性は，十分高いとは言えなかったが，ある程度の一致が見られた（反意語の3人全員で一致した割合76.16%，内的一貫性クロンバックの α = .65；コロケーションの一致度52.38%，内的一貫性 α = .63）。この理由は，模範解答にある語と完全に間違いの語は抜いた，残りの語での判定であること，また，特に低い一致度になっているコロケーションは，経験や文化的背景に影響されやすいことが考えられる。そこで，3人全員で一致しなかった項目については，別の英語母語話者1人と日本人英語上級者1人に依頼し，5人中3人以上が正解としたものを正答ととした。

4.3.2　妥当性検証とその方法

妥当性とは，一般に，「テストが測ろうと意図した特性（構成概念）を測っている度合い」（Henning, 1987など）であるが，厳密に言うと，「テスト得点に基づいた解釈と使用の適切さが，実証的証拠と理論的根拠によって裏付けられる程度」（Messick, 1989）である。妥当性は，言語テストを作り，使用する際に最も重要であり（Bachman, 1990など），テスト作成・使用において常に検証すべきもの（Chapelle, 1999）である。妥当性は，テストが持つ特性ではなく，テスト目的と連動したテスト得点の解釈についての性質（Messick, 1996）である。よって，厳密に言うと本研究での妥当性検証は「テスト得点の解釈の妥当性」について行う。しかし，記述の容易さを考え，以降は「テストの妥当性」として記述する。

妥当性は全体として1つの概念であるが，Messick（1996）は6つの別個の要素に分けた。その妥当性の要素と，妥当性検証に含めるべき側面などについてのまとめが表3である。

妥当性検証をどのように実行するかについて，Chapelle（1999）は，妥当性検証の3段階のステップを挙げている。それは，第1に，仮説を作る，第2に，仮説を検証するための関連した証拠を提示する，第3に，証拠と理由付けを提示・統合するような妥当性に関する主張を行う方法である。

■ 表 3：妥当性の 6 つの要素と分析手法

要素	■ 妥当性検証で吟味する側面 ・チェック項目の例	具体的な分析手法の例
内容的要素	■ テストの内容と測定領域の内容が一致しているかの検証 ・構成概念の定義が明確か ・構成概念とテスト項目・タスクが一致しているか ・内容が関連しているか ・内容に代表性があるか ・技術的質が保たれているか （例：指示が明確か，適切な難易度か）	・タスク・カリキュラムの分析 ・専門家の判断［5.1.1節］ ・分析的枠組みの使用
実質的要素	■ 受験者が項目・タスクにどう反応しているかの検証 ・理論的に予測されるプロセスが実際のテスト中に見られるか ・専門家でない人にとって，テストが意図した構成概念を測っているように思えるか（このチェック項目の位置付けは難しいが，本研究では Fulcher（1999）に基づきここに位置付けた）	・観察・質問紙［5.1.2節］ ・面接・プロトコル分析による質的分析 ・受験者が用いる方略・言語の報告 ・談話分析
構造的要素	■ テストの内的構造（タスク・項目・設問間の関係）の検証 ・得点尺度や採点法・採点基準やテストの内的構造と，構成概念の内的構造が一致しているか ・データの次元が，仮定された構成概念の次元と一致（適合）するか	・項目応答理論 ・信頼性［5.1.3節］ ・因子分析 ・構造方程式モデリング［5.2節］
一般化可能性的要素	■ 時間・グループ・受験状況・タスク・評価者の変化などにより，テストのプロセスと構造が変化するかの検証 ・得られた得点の解釈が，構成概念の領域に一般化できるか	・信頼性 ・一般化可能性理論 ・差異的項目機能（DIF） ・分散分析［5.1.4節］
外的要素	■ テスト得点とテストの外的構造（他の測定結果・背景的変数）の関係の検証 ・構成概念によって，相関の外的パターンを合理的に説明できるか ・外的基準との関係（関係のなさ）が得点の意味と一致するか	・相関［5.1.5節］ ・多特性・多方法の方法 ・因子分析 ・構造方程式モデリング
影響的要素	■ テスト得点を解釈・使用する際に社会的な影響があるかの検証 ・テスト得点を解釈・使用する際に，意図した影響と意図しなかった影響が短期的・長期的にあるか	・観察・インタビュー ・質問紙［5.1.2節］ ・事前・事後テストの分析

（注）Bachman & Eignor（1997），Banerjee & Luoma（1997），Chapelle（1999），Cheng & Watanabe（2004），Messick（1989, 1996）をもとに作成。下線部 = 本研究で使用；［ ］= 結果が述べられる節。

本研究でも，可能な限りこの 3 段階のステップに沿って妥当性検証を行う。その際，Messick（1996）が言うように，特定の妥当性の側面にだけ偏って行わず，より包括的に行うことに注意する。

各分析については原則として結果の前で述べるが，以下の 2 点についてはここで述べる。第 1 に，本研究では 3 グループ以上間（例：英検の 2 級・準 2 級・ 3 級・ 4 級・ 5 級取得者間）で違いがあるかを見る際に，一元配置分散分析を用いた。分散分析の前提である「等分散性」が満たされなかった場合，Brown-Forsythe の修正を行った。従属変数が複数あり，それぞれ相関があったが，正の相関だったため，多変量分散分析（MANOVA）を使わず（Tabachnick & Fidell, 2001），（一変量の）一元配置分散分析を用

いた。そして，第１種の誤りを減らすために，ボンフェローニの修正を用い，有意水準を .01（.05 / 4）として検定した。分散分析で有意になった場合にはボンフェローニの多重比較を用いた。また，最近は独立変数が従属変数と関連するか否かしか見ない有意性検定だけでなく，関連の程度（効果量）も見るべき（Wilkinson & Task Force on Statistical Inference, 1999など）との考えが広まってきている。

本研究では効果量の１種の，独立変数が従属変数の分散を説明する割合である η_G^2（generalized eta squared：「主効果の平方和/全体の平方和」により算出；Olejnik & Algina, 2003）を報告し，解釈は Cohen（1988）に基づいた（.010 < 小 <.059；.059 < 中 < .138；.138 < 大）。

第２に，２つの変数間の関係を調べる際には，ピアソンの積率相関係数を用いた。２つの相関係数間に差があるかを吟味する際には，Glass & Hopkins（1996）の式を用い，対応のある場合とない場合の検定のどちらかを行った。第１種の誤りを減らすために，ボンフェローニの修正を用い，有意水準を下げて（.05/ まとめて行う検定の数）検定した。さらに，後の研究の統合のために，青木（2003）の Excel シートを用い，相関係数の95％信頼区間も挙げた（信頼区間については Smithson, 2003 を参照）。

5 結果と考察

5.1 発表語彙知識テストの妥当性検証

5.1.1 内容的要素の検討

構成概念の定義に基づき，テスト方法が選ばれた。広さテストについては，JACET8000の Level 1～3 の語からランダムに選んで作成したことで，内容の関連性と代表性は保たれていると考えられる。深さテストについては，JACET8000の Level 1～3 の語から選んで作成したことから内容の関連性はあると考えられる。さらに２つのテストとも予備調査を行い，指示や難易度の調整を行ったことから，技術的な質も十分であると考えた。よって，内容的要素はかなり満たされていると考えられる。

5.1.2 実質的要素・影響的要素の検討

質問紙で「わからない」と答えた受験者の解答を除いた結果が表４である。ここでは，実質的要素について尋ねた２問（Q2と Q3）と影響的要素について尋ねた１問（Q4）について検討した。

■ 表４：発表語彙知識テストに対する受験者の反応

	Q2（n = 775）	Q3（n = 824）	Q4（n = 833）
M（*SD*）	3.85（1.07）	3.64（1.17）	3.55（1.19）

（注）*M* = 平均値; *SD* = 標準偏差; n = 人数

表面的妥当性（Q2），指示の明快さ（Q3），テスト受験が今後の学習に影響するか（Q4）についての質問では，平均値はすべて3.5以上であった。受験者は，本テストをある程度妥当なテストととらえたようである。

5.1.3 構造的要素の検討

第１に，広さテストと深さテストの３セクションについて，４つの観点から吟味した。第１に，項目応答理論のミスフィット項目の観点から，第２に，信頼性の観点から，第３に，広さテストの JACET8000レベル間の関係の観点から，第４に，構造方程式モデリングの観点からである。第４点目は，5.2節で検討する。

第１の分析で使用する項目応答理論は，受験者の能力とテスト項目の難易度を独立に推定でき，テストの精度について細かな情報を得られる有用なテスト理論である（大友，1996など）。本研究では，１パラメータ・モデル（ラッシュ分析）を用い，以下の仮説を検証した。

妥当性の仮説１：本研究のテスト項目はテストのモデルに適合している（ミスフィットの項目がない）

もし，モデルに本テスト項目が十分適合しているならば，項目応答理論の前提である，テストの一次元性（同時に分析されるテスト項目が同じ１つの能力を測っていること）が満たされると考えられる（McNamara, 1996）。ミスフィットの基準は「Infit Mean Square が平均 ＋２×標準偏差以上」（McNamara, 1996）とした。

一次元性については，テストではミスフィット項目が全体の10％未満ならばよく（Stansfield & Kenyon, 1995），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％未満ならばよい（McNamara, 1996）と考えた。なお，この節以外の分析では，満点と０点の受験者がいるために，項目応答理論の能

■表5：ミスフィットの基準値・割合と信頼性

		Infit Mean Square の *M* (*SD*) [ミスフィットの基準値]	ミスフィットの割合（基準を超えた数 / 全体）	信頼性 [Separation]
広さ1回目（n = 913）	受験者	1.0（0.3）[1.6]	2.41％（22/913）	.96 [4.85]
	項目	1.0（0.2）[1.4]	0.00％（0/84）	1.00 [20.03]
広さ2回目（n = 910）	受験者	1.0（0.3）[1.6]	2.53％（23/910）	.95 [4.60]
	項目	1.0（0.2）[1.4]	0.00％（0/78）	1.00 [19.71]
派生語（n = 910）	受験者	1.0（0.4）[1.8]	3.63％（33/910）	.87 [2.55]
	項目	1.0（0.2）[1.4]	5.00％（1/20）	1.00 [18.89]
反意語（n = 910）	受験者	1.0（0.4）[1.8]	3.63％（33/910）	.81 [2.10]
	項目	1.0（0.2）[1.4]	0.00％（0/20）	1.00 [19.33]
コロケーション（n = 910）	受験者	1.0（0.5）[2.0]	3.63％（33/910）	.78 [1.91]
	項目	1.0（0.1）[1.2]	0.00％（0/18）	1.00 [19.49]

（注）*M* = 平均値；*SD* = 標準偏差；n = 人数；受験者 = 受験者の能力；項目 = テスト項目の難易度

力推定値でなく，テストの素点を用いた。

最初に広さテストを分析したところ，ミスフィットの受験者（22人）が，全受験者の2％を超えていた。そのうち広さテストの1ページ以上が空白であり，誤って抜かした形跡があった3人は除外して再分析した。

表5を見ると，受験者のミスフィットについては，各セクションとも2～4％のミスフィットがあった。その理由を検討したところ，特にテスト項目に関する問題は見当たらず，除外しないことにした。

項目のミスフィットについては，派生語セクションでの1語（5％）以外はなく，仮説1はほとんど支持されたため，テストの一次元性は満たされていたと考えられる。1つあったミスフィット項目（資料2(a)）はわずかなミスフィット（Infit Mean Square が1.41）だったため，除外せず分析を進めた。

妥当性の第2の観点である信頼性については，「高い（妥当性の仮説2）」ものが求められる。結果は，項目と受験者の信頼性ともに，すべてのセクションにおいて高く，仮説2は支持された。

第3の観点である，広さテスト内でのJACET8000レベル間の関係については，個人で見た場合「Level 1の正答数が最も多く，Level 3が最も少ない（妥当性の仮説3）」と予想される（例：Level 1で15問正解 > Level 2で8問正解 > Level 3で3問正解ならば仮説が満たされる）。この仮説に一致した受験者は多く（93.96％；855 / 910），仮説3はほぼ支持された。仮説に一致しなかった受験者（55人）には中高の幅広い層の受験者が含まれていた。そのうち46人（83.64％；46 / 55）はある高校の生徒で，先生によると，対象者（高2）に難しめの単語テストを毎回行っているとのことだった。その単語テストに出た語と本研究の広さテストとの共通語は10語で，46人中45人はその共通語により仮説からずれた分布の説明が可能だった。

5.1.4　一般化可能性的要素の検討

ここでは，受験者のグループ間（英検取得級間）で違いがあるかについて検討した。英検はコミュニケーション能力全般を測るため，取得級が高いと語彙知識も高いと予想されるため，以下の仮説が立てられる。

妥当性の仮説4：英検取得級が上がると，本テスト得点は高くなる（2級 > 準2級 > 3級 > 4級 > 5級）

表6によると，級の違いによる効果量は大きかった。英検2級・準2級・3級の間では常に有意差があったが，3級・4級・5級間では，全体的に有意差がなかったため，仮説は一部支持された。3級・4級・5級の級間に差が見られなかった理由は2点考えられる。第1に，本テストが，3級・4級・5級程度の学習者の弁別力が低いためである。第2に，各級の語彙レベルが2級で約5100語，準2級で3600語・3級で2100語・4級で1300語・5級で600語（英検，2004）というところからもわかるように，3級・4級・5級のレベルの差は2級・準2級・3級

■ 表 6 ：英検取得級による語彙知識テスト得点の違い

	級	2	準2	3	4	5
	人数	49	164	274	82	20
広　さ	*M*	1985.10	1365.15	1002.24	864.89	898.10
	SD	427.03	418.35	325.10	376.23	519.77
	F (4, 120.41) = 75.86** ［2 > 準2 > 3 = 4 = 5］; η_G^2 = .40					
派生語	*M*	14.63	10.29	7.27	5.70	5.70
	SD	3.73	3.96	3.63	3.63	4.52
	F (4, 584) = 63.16** ［2 > 準2 > 3, 4, 5; 3 > 4; 3 = 5; 4 = 5］; η_G^2 = .30					
反意語	*M*	11.67	8.38	5.83	5.39	4.60
	SD	2.24	2.94	2.47	2.76	3.42
	F (4, 584) = 73.20** ［2 > 準2 > 3 = 4 = 5］; η_G^2 = .33					
コ　ロ	*M*	14.41	12.15	10.34	9.78	7.65
	SD	2.13	2.59	2.49	2.56	4.45
	F (4, 68.07) = 33.80** ［2 > 準2 > 3 = 4 = 5］; η_G^2 = .24					

（注）*M* = 平均値；*SD* = 標準偏差；［ ］= 多重比較の結果；コロ = コロケーション。**p < .01

■ 表 7 ：CASEC と発表語彙知識テストの相関関係

	CASEC1	CASEC2	CASEC3	CASEC4	CASEC合計
広　さ	.58（.45 to .68）	.61（.49 to .71）	.59（.47 to .69）	.76（.68 to .82）	.77（.69 to .83）
派生語	.52（.38 to .63）	.54（.41 to .65）	.50（.36 to .62）	.67（.56 to .75）	.68（.58 to .76）
反意語	.50（.36 to .62）	.53（.40 to .64）	.55（.42 to .66）	.71（.61 to .79）	.69（.59 to .77）
コ　ロ	.47（.33 to .59）	.38（.22 to .52）	.46（.31 to .58）	.50（.36 to .62）	.54（.41 to .65）

（注）133人で分析。すべて 1 %水準で有意。CASEC1 = CASEC のセクション 1 ；コロ = コロケーション；（ ）= 95%信頼区間

のレベルの差よりもなく，3 級・4 級・5 級保持者の語彙知識はもともとそれほど差がないためとも考えられる。

5.1.5　外的要素の検討

本節では，ある中学校における，本テストと他のテスト（CASEC: Computerized Assessment System for English Communication）の相関を調べた。CASEC は，項目応答理論とコンピュータ適応型テストシステムを用いた，コミュニケーション能力判定テストである（教育測定研究所，2005）。セクション 1 ～ 4 から成り，それぞれ「語彙の知識」，「表現の知識・用法」，「聞いた内容の大意を理解する能力」，「具体的な情報を聞き取る能力」を測定するとされる。テスト方法は，セクション 1 ～ 3 は多肢選択問題で，セクション 4 は書き取りである。テスト内容やテスト方法が近いもの同士は遠いもの同士より相関が高いという前提（Henning, 1987など）に立ち，以下の 3 つの仮説を立てた。

妥当性の仮説 5 ：（語彙知識はコミュニケーション能力の一部であり，2 つの相関はある程度あると考えられるため），本研究での広さ・深さテストと「コミュニケーション能力を測る」CASEC の総合得点は中程度以上の相関がある

妥当性の仮説 6 ：「語彙知識を測る」セクション 1 は，他の要素を測るセクション 2 と 3 より，本研究での広さ・深さテストとの相関は高い

妥当性の仮説 7 ：（テスト方法が近いため）セクション 4 はセクション 2 と 3 より，本研究での広さ・深さテストとの相関は高い

結果（表 7 ）は，仮説 5 については，CASEC 合計と 4 セクションの相関はすべて中程度以上の相関が見られ，支持された（資料 4 参照）。

仮説 6 については，相関係数で見ると一部仮説が

満たされた（派生語について，セクション１ ＞ ３；コロケーションについて，１ ＞ ２，３）が，相関係数の差は小さく，５％水準でも（有意水準を下げた1.67［＝５/３］％水準でも）有意な差はなかった（派生語とセクション１と３間で t（130）＝0.33，［最も差があった］コロケーションとセクション１と２間で t（130）＝1.19）。その理由として，テスト問題を見るとセクション１と２の主な違いは選択肢の長さ（１語提示かそれ以上か）で，測る要素がより近く，また本テストで測っている語彙知識がセクション２と３で測る知識・能力とも関連していることが考えられる。

仮説７については，相関係数で見るとすべてが満たされたが，0.63（＝５/８）％水準で有意な差があったのは４箇所（広さとセクション２と４間・３と４間，派生語とセクション３と４間，反意語とセクション２と４間。それぞれ t（130）＝3.67, 3.35, 3.16, 3.44）で，仮説は一部支持された。

5.1.6　発表語彙知識テストの妥当性のまとめ

今まで６つの妥当性の要素の観点から，本テストを検討してきた。仮説に一致しなかった点は，妥当性に対する否定的な証拠であり，今後さらに改善が必要と思われる。しかし，仮説に支持された点は妥当性に関する肯定的な証拠であり，かなりの肯定的な証拠が見られたことから，本研究で作成したテストの妥当性はある程度あり，研究で使用する意味があると考えられる。

5.2　発表語彙知識の広さと深さの関係

広さと深さの関係を見る際に，受験者全体の分析では構造方程式モデリングを，広さレベルごとに見る場合は，人数が少ない群があるためピアソンの積率相係数を用いた。

第１の分析で用いる構造方程式モデリングは，テスト得点を観測変数として扱えるだけでなく，直接測定することのできない構成概念を潜在変数としてモデルに組み込み，誤差（図１の円）を取り除いた形で変数間の関係が吟味できる分析手法である（狩野＆三浦，2002など）。

本研究では，「深さテストの３セクションの素点」と「広さの推定値（78問から語彙サイズを推定した値）」を観測変数（図１の四角）とし，「深さ」を潜在変数（図１の楕円）とした。

まず，構造方程式モデリングを行う前に，Kunnan（1998）に基づき２つの前提を確認した。１つ目の「変数ごとの正規性」については，歪度・尖度ともすべて|±2| 内だったため満たしていた（資料５，６参照）。２つ目の「変数を組み合わせたときの正規性（多変量正規性）」については，910人の分析で多変量正規性の値が1.20で，５％水準で有意に正規性が満たされていなかった。そこで，最大の多変量外れ値を持つ（外れ値の基準：マハラノビスの距離 $\chi 2$ = 13.82以上，df = 2, p < .001; Tabachnick & Fidell, 2001）の受験者１人を除き，再分析したところ，多変量正規性の値が0.64になり，多変量正規性が満たされた。よって，後は909人のデータを使用した。除いた１人は，他の受験者と比べると，広さテスト得点が非常に高い（推定語彙サイズ2231語）のに対し，深さテストセクションでは普通の点（各８点，６点，12点）を取っていた。よって，本研究での結果は，そのような学習者には一般化できないことがわかる。

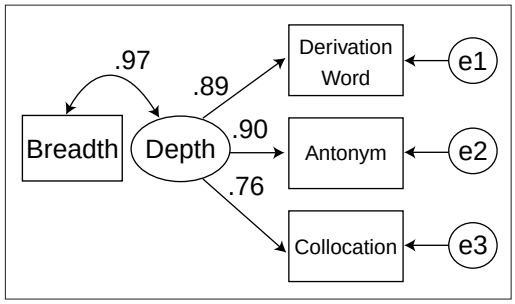
■ 表８：図１のモデルの適合度の結果

	χ^2 (df), p	χ^2/df	CFI	GFI	RMSEA (90%CI)
基準	p > .05	< 2.0	> .90	> .90	< 0.05
図１	8.84 (2), .01	4.42	1.00	1.00	0.06 (0.02 to 0.11)

（注）909人で分析。CFI = Comparative Fit Index；GFI = Goodness of Fit Index；RMSEA = Root Mean Square Error of Approximation；CI = 信頼区間（Confidence Interval）。基準は Arbuckle & Wothke（1995）を参照。

表８を見ると，χ^2 と χ^2 / df と RMSEA では基準が満たされていないものの，CFI と GFI では満たされており，モデルに適合したと考えた。図１によると，広さと深さの相関は .97で非常に強かった。つまり，全体で見ると，広さのある人は深さもある傾向が強い。これは，Mochizuki & Aizawa（2000）が述べるように，広さが増えれば深さも広がり，深さが大きくなれば広さも広がる傾向にあるためだろう。ただし，構造方程式モデリングを使用しない場合の相関（資料７）は r = .73〜.88であり，１つのテストから予測できる割合は53.79〜77.44％にすぎないことにも注意する必要がある。なお，相関が構造方程式モデリングでの相関よりも低くなっている理由は，

▼ 図１：構造方程式モデリングによる広さと深さの関係



（注）Breadth = 広さ；Depth = 深さ；Derivation Word = 派生語；Antonym = 反意語；Collocation = コロケーション

構造方程式モデリングが誤差を取り除いた分析をしているためである。

さらに，深さの観点から見ると，３つのセクションすべてが Depth の因子に高い負荷があり，構造的な妥当性の証拠と考えられる。

次に，909人を1000語単位の広さレベルごとにグループ分けをし，広さと深さの関係を検討した（表９）。全体での分析よりは範囲が限られるために相関が低くなるのは予想できたが，結果は3000語レベルの広さとコロケーションの相関（r = .25）以外は中程度の相関（r = .50〜.68）があった。低い相関が見られたのはコロケーションセクションの信頼性が低かった（α = .51；表10）ためかもしれない。全体的に見ると，1000語単位で区切る場合は中程度の相関があると考えられる。つまり，広さのある人は深さもある傾向が中程度ある。また，最も説明できる割合でも，相関を２乗して46.24％で，53.76％は説明できないことになる。つまり，説明できない割合が多いため，深さには広さから予測できない部分が多くあり，したがって，狭い能力範囲を調べるときは特に，広さだけでなく，深さを検討する必要もあると考えられる。

その後，1000語レベルごとの結果（表９）を縦に見た。0.56（= 5/9）％水準で有意な差があったのは１箇所で，派生語で1000語レベルと2000語レベル（1000 < 2000）だった（z = 4.10）。信頼性α = .70以上を高いと考えるならば，派生語テストの信頼性は基準を満たしているため，1000語レベルと2000語レベルと広さが増えると，広さと派生語の知識の関係は強くなると考えられる。全体的に見ると，レベルが異なってもそれほど広さと深さの関係は変わらないと考えられる。

■ 表９：1000語の広さレベルごとの広さと深さの相関

[人　数]	派生語	反意語	コロ
1000語レベル [409]	.50** (.42 to .57)	.64** (.58 to .69)	.50** (.42 to .57)
2000語レベル [412]	.68** (.62 to .73)	.68** (.62 to .73)	.51** (.43 to .58)
3000語レベル [88]	.60** (.45 to .72)	.56** (.40 to .69)	.25* (.04 to .44)

（注）コロ = コロケーション；（ ）= 95％信頼区間
1000語レベル = 1〜999語の推定値だった受験者群；2000語レベル = 1000〜1999語の群；3000語レベル = 2000〜3000語の群。*p < .05. **p < .01.

■ 表10：広さレベルごとの記述統計

[人数]	広さ <3000語>	派生語 <20点>	反意語 <17点>	コロ <18点>
1000語レベル [409]	711.71 (190.60, .73)	5.01 (2.92, .72)	4.13 (2.10, .59)	8.60 (2.81, .69)
2000語レベル [412]	1352.86 (274.82, .83)	9.96 (3.49, .76)	7.98 (2.51, .62)	11.80 (2.46, .63)
3000語レベル [88]	2266.58 (218.95, .71)	17.15 (2.05, .53)	12.92 (1.83, .34)	15.26 (1.98, .51)

（注）< > = 満点；平均値（標準偏差，クロンバックのα）；コロ = コロケーション

また，各広さレベルで「広さ」と「深さのどの要素」の関係が強いかについて，表９を横に見て，0.56（= 5/9）％水準で有意な差があるものを調べた。結果は，(1) 1000語レベルでは，広さと反意語の関係が広さと派生語，広さとコロケーションの関係よりも高かった（それぞれ t (406) = 3.71, 3.82）。(2) 2000語レベルでは広さと派生語，広さと反意語との関係が広さとコロケーションとの関係よりも高かった（それぞれ t (409) = 4.91, 5.02）。(3) 3000語レベルでは広さと派生語の関係が広さとコロケーションとの関係よりも高かった（それぞれ t (85) = 3.30）。しかし，(1)〜(3) とも組み合わせのどれかのセクションにおいて信頼性が低めで誤差が大きいため，今後さらに検討が必要だろう。

最後に，各レベルでの，語彙サイズが増えるに従い，深さの要素がどのように伸びるかについて検討した。

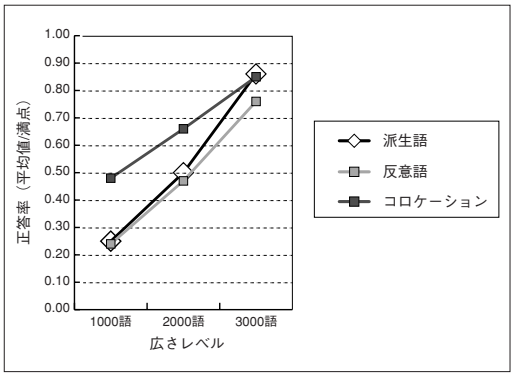
表11・図２によると，広さレベルが高くなるにつれて，すべての深さ得点は有意に伸び，効果量も大きかった。つまり本研究からは，Shimamoto (2000,

■ 表11：広さレベルによる深さテスト得点の違いの検定

セクション	F 値	η_G^2
語彙サイズ	F（2, 409.94）= 1962.53**	.80
派生語	F（2, 698.79）= 820.18**	.58
反意語	F（2, 532.11）= 759.03**	.59
コロケーション	F（2, 577.10）= 369.60**	.41

（注）多重比較の結果，すべて1000語レベル < 2000語レベル < 3000語レベル。

▼ 図 2：広さレベルの変化による深さテスト得点の伸び



2005）, Mochizuki & Aizawa（2000）と同様に，各グループとも，広さが増えるにつれて，深さの 3 要素が伸びていることがわかる。なお，深さ 3 セクションのもともとの難易度が等しいとの保証はないため，その間での比較はできない。

6 結論

本研究でわかったことは以下 2 点である。まず，リサーチクエスチョン RQ1（発表語彙知識の広さと深さの関係はどの程度の強さか）に関連して，発表語彙知識の広さの3000語レベルまでの幅広い学習者層で見ると，広さと深さの関係は強く，1000語単位で見ると，中程度の関係がある。

RQ2（広さが大きくなるにつれて，広さと深さの関係は変化するか）については，広さと派生語の相関で1000語レベルが2000語レベルより低いがそれ以外は違いがあるとは言えなかった。また，広さが増えるに従い，深さの 3 つの要素はすべて伸びる。

教育的示唆として以下の 3 点が挙げられる。第 1 に，全体的には強い相関が見られたことから，発表語彙知識を測定する際，対象者の発表語彙知識の広さの範囲が幅広い場合には，広さから深さがある程度は予測できるとわかる。しかし，発表語彙知識の広さのレベルごとでは中程度の関係で，予測できる値も半分以下だったことから，対象者の発表語彙知識の広さの範囲が狭い場合には，両方の測定をしないと，包括的に語彙をとらえたことにはならないことがわかる。

第 2 に，広さレベルごとに見ると，全体的には関連の強さにあまり違いがなく，レベルが上がるにつれて深さの局面も一貫して伸びていたことから，語彙サイズが小さい英語学習初期の段階から，広さだけでなく深さの知識も伸びているかを評価することが必要だと示唆される。また指導に関しては，混乱させない程度に，深さについても指導していくことが必要と思われる（望月，2003）。

第 3 に，語彙知識のさまざまな側面を測るテストが開発されてきていない（Aizawa, 2005）が，本研究で作成したテストを教育や研究で使うこともできると考えられる。また教育面では，深さテストを実施することで，広さ以外の語彙知識の要素に目を向けさせることも可能だろう（Aizawa, 2005）。

今後は，以下の 6 点を特に研究し，より一般化につなげるべきだろう。第 1 に，テスト方法の影響（Bachman, 1990）を減らすために，深さテストだけでなく広さテストにも複数のテスト方法（セクション）を用いる必要がある。

第 2 に，広さレベルごとの分析で十分な信頼性を保ちつつ解釈を行うために，母集団を反映した形で受験者の抽出を行い，参加者をより多く確保し，難易度が低い項目と高い項目もさらにテストに加えての再検証が必要である。

第 3 に，幅広い層の受験者を対象にする場合には，より易しいテストの版と，より難しいテストの版のように複数の版を用意するなど，受験者の能力に適したテストの実施が求められるだろう。通常の分析では別のテストを行うと比較可能でないが，項目応答理論を用いてテストの等化（大友，1996など）を行えば，同じ尺度上で解釈が可能になる。

第 4 に，妥当性の要素それぞれについて，よりさまざまな観点から調べることが必要である。特に，本研究での妥当性検証は，理論に基づくよりは，一般に考えて仮説が導けるものを挙げて検証するもので，Messick（1989）の言う「弱い妥当性検証」に当たる。今後理論が発達すれば，その理論に基づい

て検証を行う「強い妥当性検証」を行うことが望ましい。さらに，Messick（1996）の枠組み以降にも，妥当性や妥当性検証法についての議論がなされており（Borsboom, Mellenbergh, & van Heerden, 2004; Kane, 2001など），今後もより良い方法を検討していくべきである。

第５に，理論構築につなげるために，テストの妥当性・信頼区間などの観点から先行研究を再解釈し，先行研究と本研究の結果を統合してメタ分析（芝＆南風原, 1990など）を行い，今後も研究を重ねていくことが求められる。

第６に，本研究の対象は，発表語彙知識の広さと深さであり，それらが実際の言語運用とどのように関係するかについては，今後の課題である。

謝　辞

この研究の機会をくださった（財）日本英語検定協会，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，ご指導いただきました望月昭彦先生，磐崎弘貞先生，卯城祐司先生，印南洋さん，古賀功さんに感謝いたします。さらに，小菅敦子先生，伊佐地恒久先生，久保野雅史先生，植木明美先生，山本良一先生，佐野賢一先生，櫻井友裕先生，上西幸治先生，川島智幸先生，瀧口均先生，柴山聖徳先生，斉田智里先生，山内逸美先生，廣澤嘉成先生，折原史康先生，根本章子先生，武山晃子先生をはじめ，研究にご協力くださいました，たくさんの先生方にも深く感謝いたします。

参考文献（＊は引用文献）

＊Aizawa, K.（2005）. Assessing the depth of receptive vocabulary knowledge. *Facilitating vocabulary acquisition—JACET summer seminar proceedings No. 4*（pp. 18-28）. Tokyo: Japan Association of College English Teachers.

＊Akase, M.（2005）. The roles of breadth and depth of vocabulary knowledge in EFL reading comprehension: With a focus on English major students. *ARELE*, 16, 141-150.

＊青木繁伸.（2003）.「検定および推定」. Retrieved April 24, 2005, from http://aoki2.si.gunma-u.ac.jp/lecture/stats-by-excel/work-sheets/tests.xls

＊Arbuckle, J.L.（1999）. Amos（Version 4.02）[Computer software]. Chicago: SmallWaters Corporation.

＊Arbuckle, J.L., & Wothke, W.（1995）. *Amos 4.0 user's guide*. Chicago: SmallWaters Corporation.

＊Bachman, L.F.（1990）. *Fundamental considerations in language testing*. Oxford University Press.

＊Bachman, L.F., & Eignor, D.R.（1997）. Recent advances in quantitative test analysis. In C. Clapham, & D. Corson（Eds.）, *Encyclopedia of language and education Vol. 7: Language testing and assessment*（pp. 227-242）. Dordrecht, the Netherlands: Kluwer Academic.

＊Bachman, L.F., & Palmer, A.S.（1996）. *Language testing in practice*. Oxford University Press.

＊Banerjee, J., & Luoma, S.（1997）. Qualitative approaches to test validation. In C. Clapham & D. Corson（Eds.）, *Encyclopedia of language and education Vol. 7: Language testing and assessment*（pp. 275-287）. Dordrecht, the Netherlands: Kluwer Academic.

＊Bauer, L., & Nation, P.（1993）. Word families. *International Journal of Lexicography*, 6, 254-279.

＊Borsboom, D., Mellenbergh, G.J., & van Heerden, J.（2004）. The concept of validity. *Psychological Review*, 111, 1061-1071.

＊Canale, M., & Swain, M.（1980）. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.

＊Centre for Applied Language Studies（CALS, University of Wales, Swansea）.（n.d.）. KR21 Calculator. Retrieved August 12, 2005, from http://www.swan.ac.uk/cals/calsres/encyclopedia/kr21.htm

＊Chapelle, C.A.（1999）. Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

＊Cheng, L., & Watanabe, Y. with Curtis, A.（Eds）.（2004）. *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

＊Cohen, J.（1992）. A power primer. *Psychological Bulletin*, 112, 155-159.

＊Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D.E.（2004）. A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, 21, 107-145.

＊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.

＊福地美奈子.（2004）.「語彙知識と EFL リーディングとの関係についての一考察」.『第30回全国英語教育学会長野研究大会発表要綱』.（pp. 720-725）.

＊Fulcher, G.（1999）. Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics*, 20, 221-236.
＊Glass, G.V., & Hopkins, K.D.（1996）. *Statistical methods in education and psychology*（3rd ed.）. Boston, MA: Allyn & Bacon.
＊Henning, G.（1987）. *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle & Heinle.
＊Kane, M.T.（2001）. Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.
＊狩野裕＆三浦麻子.（2002）.『Amos, EQS, CALIS によるグラフィカル多変量解析―目で見る共分散構造分析―（増補版）』. 京都：現代数学社.
＊Koizumi, R.（2003）. A productive vocabulary knowledge test for novice Japanese learners of English: Validity and its scoring methods. *JABAET Journal*, 7, 23-52.
＊小西友七＆南出康世（編）.（2002）.『CD-ROM 版ジーニアス英和・和英辞典』. 東京：大修館書店.
＊Kunnan, A.J.（1998）. An introduction to structural equation modelling for language assessment research. *Language Testing*, 15, 295-332.
＊教育測定研究所.（2005）.「英語コミュニケーション能力判定テスト（CASEC）」. Retrieved April 24, 2005, from http://casec.evidus.com/index.html
＊Laufer, B.（1992）. How much lexis is necessary for reading comprehension? In P.J.L. Arnaud & H. Bejoint（Eds.）, *Vocabulary and applied linguistics*（pp. 126-132）. London: Macmillan Academic and Professional Ltd.
＊Laufer, B., Elder, C., Hill, K., & Congdon, P.（2004）. Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21, 202-226.
＊Laufer, B., & Nation, P.（1995）. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307-322.
＊Laufer, B., & Nation, P.（1999）. A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 33-51.
＊Linacre, J.M.（2003）. Facets: Rasch-measurement computer program（Version 3.45.2）[Computer software]. Chicago: MESA Press.
＊McNamara, T.（1996）. *Measuring second language performance*. Essex, U.K.: Addison Wesley Longman Limited.
＊Meara, P., & Fitzpatrick, T.（2000）. Lex 30: An improved method of assessing productive vocabulary in an L2. *System*, 28, 19-30.
＊Melka, F.（1997）. Receptive vs. productive aspects of vocabulary. In N. Schmitt, and M. McCarthy（Eds.）, *Vocabulary: Description, acquisition and pedagogy*（pp. 84-102）. Cambridge University Press.
＊Messick, S.（1989）. Validity. In R.L. Linn（Ed.）. *Educational Measurement*（3rd ed., pp. 13-103）. New York: National Council on Measurement in Education/American Council on Education.
＊Messick, S.（1996）. Validity and washback in language testing. *Language Testing*, 13, 241-256.
＊望月正道.（2003）.「単語を知っているとはどういうことか」. 望月正道，相澤一美，投野由紀夫（編）.『英語語彙の指導マニュアル』.（pp. 31-70）. 東京：大修館書店.
＊Mochizuki, M., & Aizawa, K.（2000）. An affix acquisition order for EFL learners: An exploratory study. *System*, 28, 291-304.
＊文部省.（1999a）.『中学校学習指導要領（平成10年12月）解説―外国語編―』. 東京：東京書籍.
＊文部省.（1999b）.『高等学校学習指導要領解説　外国語編　英語編』. 東京：開隆堂出版.
＊中西義子＆島本たい子.（2003）.「語彙知識とその測定」. 門田修平（編）.『英語のメンタルレキシコン：語彙の獲得・処理・学習』.（pp. 31-62）. 東京：松柏社.
＊Nation, I.S.P.（2001）. *Learning vocabulary in another language*. Cambridge University Press.
＊Noro, T.（2002）. The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *ARELE*, 13, 71-80.
＊野呂忠司＆島本たい子.（2003）.「英語の語彙知識と言語運用」. 門田修平（編）.『英語のメンタルレキシコン：語彙の獲得・処理・学習』.（pp. 123-150）. 東京：松柏社.
＊Nurweni, A., & Read, J.（1999）. The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, 18, 161-175.
＊大友賢二.（1996）.『項目応答理論入門』. 東京：大修館書店.
＊Olejnik, S., & Algina, J.（2003）. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8, 434-447.
＊Qian, D.（1999）. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, 56, 282-307.
＊Qian, D.（2002）. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52, 513-536.
＊Read, J.（2000）. *Assessing vocabulary*. Cambridge University Press.
＊Read, J.（2004）. Plumbing the depths: How should the construct of vocabulary be defined? In P. Bogaards & B. Laufer（Eds.）, *Vocabulary in a second language: Selection, acquisition, and testing*（pp. 209-227）. Amsterdam, the Netherlands: John

Benjamins.
＊Schmitt, N.（2000）. *Vocabulary in language teaching*. Cambridge University Press.
＊Schmitt, N., & Meara, P.（1997）. Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, 20, 17-36.
＊芝祐順＆南風原朝和.（1990）.『行動科学における統計解析法』. 東京：東京大学出版会.
＊島田勝正.（2004）.「EAP 語彙の広さと深さの関係」. 日本言語テスト学会第 8 回全国研究大会口頭発表資料.
＊Shimamoto, T.（2000）. An analysis of receptive vocabulary knowledge: Depth versus and breadth. *JABAET Journal*, 4, 69-80.
＊Shimamoto, T.（2005）. Exploring lexical network systems of Japanese EFL learners through depth and breadth of word knowledge. *ARELE*, 16, 121-130.
＊Smithson, M.（2003）. Confidence intervals. Thousand Oaks, CA: Sage Publications.
＊Stansfield, C.W., & Kenyon, D.M.（1995）. Comparing the scaling of speaking tasks by language teachers and by the ACTFL Guidelines. In A. Cumming & R. Berwick（Eds.）, *Validation in language testing*.（pp. 124-153）. Clevedon, U.K.: Multilingual matters.
＊Statistical Package for the Social Sciences（SPSS）.（2003）. SPSS（Version 12.0.1）[Computer software]. Chicago: SPSS Inc.
＊Summers, D. et al（Eds.）.（2003）. *Longman dictionary of contemporary English*（4th ed.）. Essex, U.K.: Pearson Education Ltd.
＊Tabachnick, B.G., & Fidell, L.S.（2001）. *Using multivariate statistics*（4th international student ed.）. Needham Heights, MA: Allyn & Bacon.
＊Wilkinson, L., & Task Force on Statistical Inference.（1999）. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
＊横川博一＆島本たい子.（2003）.「コロケーション，チャンク，語彙フレーズと外国語教育への応用」. 門田修平（編）.『英語のメンタルレキシコン：語彙の獲得・処理・学習』.（pp. 245-264）. 東京：松柏社.
＊財団法人日本英語検定協会（英検）.（2004）.「検定のご案内　各級概要」. Retrieved April 24, 2005, from http://www.eiken.or.jp/info/level/index.html

資料 1：発表語彙知識広さテストの例

［1］単語を書く問題A

（　　）内に，日本語にあたる英単語を 1 語書いてください。はじめの文字は書いてあります。
例：　　町（t　　）　答えは（town）または（towns）
　　ふれる，さわる（t　　）　答えは（touch）

< 1番易しかった問題 >
1.　犬　（d　　）　　答え：dog（s）

< 平均的な難易度の問題 >
53.　～に食べ物・えさを与える　　（f　　）　答え：feed

< 1番難しかった問題 >
54.　所有すること，入手すること　　（p　　）　答え：possession

資料 2：発表語彙知識深さテストの例

(a) 派生語セクション

［2］単語を書く問題 B
英単語を，［　　］の中に書かれた品詞にしてください。（　　）内には 1 語書いてください。ing と ed をつけた語は書かないでください。
問題例 1：kind［名詞：～なこと］　答えは（kindness）
　＊「名詞」とは，主に，人や物の名前（例：ネコ）を表す言葉です。例：This is my cat. →cat が名詞

問題例 2：introduction［動詞：～する］　答えは（introduce）
　＊「動詞」とは，主に，動き（例：走る）や状態（例：住んでいる）を表す言葉です。例：I live in Japan. →live が動詞

問題例 3：wonder［形容詞：～（の状態）の］　答えは（wonderful）
　＊「形容詞」とは，人・物・出来事などについて述べる言葉（例：かわいい）です。例：I have a cute cat. →cute が形容詞

< 1番易しかった問題 >
1.　work［名詞：～する人］　（　　）　答え：worker / workman

< 平均的な難易度の問題>
5.　inform［名詞：～したもの］　（　　）　答え：information

< 1番難しかった問題 >
3.　destroy［名詞：～すること］　（　　）　答え：destruction

< ミスフィット項目 >
4.　criticize［名詞：～したもの］　（　　）　答え：criticism

(b) 反意語セクション

［3］単語を書く問題 C
（　　）の中に入る，反対の意味の英単語を 1 語書いてください。
例：good　（　　）　答えの例は（bad）

< 1番易しかった問題 >
1.　open　（　　）　答えの例：close, closed, shut

< 平均的な難易度の問題 >
7.　nonsense　（　　）　答えの例：commonsense, meaningful, reasonable, sense, sensible, valid

< 1番難しかった問題 >
11.　supporter　（　　）　答えの例：attacker, critic, destroyer, dissident, nonsupporter, opposer

(c) コロケーション・セクション

［4］単語を書く問題 D
（　　）の中に入る，適した英単語（形容詞か名詞）を 1 語書いてください。good, better, best, bad, nice, wonderful, great, big, small, old は書かないでください。同じ語が入ってもかまいません。
例：形容詞が入る例：（　　）air　答えの例は（clean）air
　　名詞が入る例：（　　）cream　答えの例は（ice）cream

<［4］の中で 1 番易しかった問題 >
4.　（　　）birthday　答えの例：16th, Christ's, funny, happy, impressive, memorial, pleasant

<［4］の中で平均的な難易度の問題 >
3.　（　　）art　答えの例：abstract, beautiful, body, cheap, classical, cute, difficult

<［4］の中で 1 番難しかった問題 >
5.　（　　）tragedy　答えの例：air, ancient, another, awful, brother's, cat, deep

［5］単語を書く問題 E
（　　）の中に入る，適した英単語（名詞）を 1 語書いてください。同じ語が入ってもかまいません。
例：clean（　　）　答えの例は clean（air）

<［5］の中で 1 番易しかった問題＆［5］の中で平均的な難易度の問題 >
2.　wet（　　）　答えの例：blanket, box, boy

<［5］の中で 1 番難しかった問題 >
4.　eager（　　）　答えの例：animal, boy, collector

［6］単語を書く問題 F
（　　）の中に入る，適した英単語（名詞）を 1 語書いてください。同じ語が入ってもかまいません。
例：play（　　）　答えの例：play（baseball）
注：（a / an / the）は，a, an, the がついても，つかなくてもいいという意味

<［6］の中で１番易しかった問題>
2.　buy（a / an / the）（　　）　　答えの例：anything, apples, bag, cameras, candies, comic, computer

<［6］の中で平均的な難易度の問題>
6.　pay（a / an / the）（　　）　　答えの例： attention, bills, cash, coins, debt, fare, fee, 100 yen

<［6］の中で１番難しかった問題>
3.　destroy（a / an / the）（　　）　　答えの例：animal, baseball, boy, bridge, cake, car, card

資料３：アンケート項目

あなたの気持ちに一番近いものを，下の0〜5から選んで，［　］に書いてください。

0. わからない	1. 全然そうでない
2. ほとんどそうでない	3. どちらともいえない
4. 少しそうだ	5. とてもそうだ

(2-2)　このテストは英語の単語力をみていると思いますか？
(2-2)［　　］

(2-3)　テストのやり方はよくわかりましたか？
(2-3)［　　］

(2-4)　このテストを受けて，今後もっと英語の勉強をしようと思いましたか？
(2-4)［　　］

資料４：CASEC と本テストの記述統計

	CASEC1	CASEC2	CASEC3	CASEC4	CASEC合計
M	89.41	99.26	104.65	104.41	397.73
SD	23.59	22.41	23.90	30.92	84.67
KR21	0.95	0.96	0.97	0.98	1.00

	広さ	派生語	反意語	コロケーション
M	1279.38	8.68	7.60	10.65
SD	412.12	3.60	2.67	3.13
α	.93	.79	.70	.77

（注）133人で分析。CASEC1 = CASECのセクション１；M = 平均値；SD = 標準偏差；α = クロンバックのα。KR21 = Kuder-Richardson の公式21による信頼性係数（CALS, n.d., を使用）

資料５：テストの記述統計

	M	*SD*	最小値	最大値	歪度	尖度	α
広　さ	1152.83	530.21	38	2962	0.79	0.15	.95
派生語	8.43	4.84	0	20	0.43	-0.46	.88
反意語	6.73	3.55	0	17	0.45	-0.38	.82
コ　ロ	10.69	3.35	0	18	-0.33	0.16	.79

（注）909人で分析。*M* = 平均値；*SD* = 標準偏差；α = クロンバックのα；コロ = コロケーション

資料６：中高生・大学生の推定語彙サイズ（広さ）

	人数	*M*	*SD*	最小値	最大値	α	信頼区間
中学校平均	327	1032.72	414.84	38	2231	.93	817.60 to 1247.84
高校平均	583	1222.06	575.58	231	2962	.96	996.43 to 1447.69
大学生（院生含む）平均	17	1859.82	440.85	1077	2808	.93	1631.21 to 2088.43

（注）大学生・大学院生の結果は，予備調査で得られた，ある１大学（英語専門・他専門を含む）の結果。*M* = 平均値；*SD* = 標準偏差；α = クロンバックのα；95%信頼区間は Bachman（1990）に基づき，*M*±1.96×（SD×√（1-α）で算出。

資料７：広さと深さの相関

	派生語	反意語	コロケーション
広　さ	.86**（.84 to .88）	.88**（.86 to .89）	.73**（.70 to .76）
派生語	—	.79**（.76 to .81）	.69**（.65 to .72）
反意語		—	.70**（.67 to .73）

（注）909人で分析。**$p < .01$.