TOEICと「科学技術英語」統一テスト実施結果の分析
—妥当性と信頼性の観点から—

小　山　由紀江

This paper provides an analysis of the TOEIC and unified tests which were implemented at Nagoya Institute of Technology in the school year of 2005. The students took the TOEIC in April, 2005 and February, 2006 - the beginning and end of the school year. A unified test for the first year required English class was given as the final examination each semester - in July, 2005 and February, 2006. Discussing the definitions of reliability and validity in terms of test construction and implementation, the reliability and the validity of these two kinds of tests are examined. Finally, the future implications of improvement to the unified tests and changes in the role of the TOEIC in the curriculum are presented.

1．始めに

名古屋工業大学では2003年から1年次の学生に年間2回のTOEICを実施してきた。第一回目は入学時、第二回目は1年後期の期末（通常2月の初め）である。このTOEICはプレースメントテストとしての役割を持ち、1年生のクラス分けは第一回TOEICの結果をもって行っている。しかしTOEIC導入1年後の2004年に、カリキュラム全体が見直され、本学の英語科目は一般的な英語を内容とする「英語演習Ⅰ」から「科学技術英語Ⅰ」に変わった。この「科学技術英語Ⅰ」に関しては、「統一テスト」（以下、統一テスト）がアチーブメントテストして使用されている。学生の成績の50%はこのテストの結果が反映されている。

このように英語教育のターゲットが科学技術英語に定められたという現在の状況において、ビジネスの場面を主とする一般的なコミュニケーション能

力を測るTOEICがプレースメントテストとして使われることに関しては、英語教員の中でもこれまでに何回か論議がなされてきた。だがしかし、TOEICの社会的認知度は年々高まり、学生が卒業までに受験の機会を持ち、TOEICによる英語力測定を経験しておくことの意義が高まっていることも事実であり、この問題に関しては今後の英語カリキュラム運営の点からもさらに考察を重ねる必要がある。

2．本研究の目的

以上述べたような背景の下に、本学の1年生は年2回のTOEICテストを受験し、前期と後期の期末テストとして統一テストを受けている。本研究では既存の標準テストTOEICと科学技術を内容とする統一テストの結果を比較することにより、統一テストの信頼性と妥当性について考察する。またこのテスト結果の分析とアンケート調査の結果に基づき、TOEICテストと統一テストをカリキュラムにどのように組み込んで行くべきかという点に関して考察し提言を行う。以上が本研究の目的である。

3．実施している試験について

3．1．TOEICの現状

TOEICについては日本では既に社会的に広く認知され、その教育現場への波及効果（washback effect）も相当大きいものになっているが、ここではまずテストの概要を説明し、全体的な受験の現状に関して簡単な説明を行っておきたい。

TOEICはTest of English for International Communicationの略称で、英語による国際コミュニケーション能力を測定するテストであり、TOEFLなどを開発しているアメリカの非営利テスト開発機関Educational Testing Service (ETS) によって開発・制作されている。TOEICの日本公式サイト（http://www.toeic.or.jp/）によると、1979年に第一回の試験が実施されて以来、これまでに延べ1300万人以上が受験しており、我が国では受験者数の最も多い標準テストの一つと言えよう。開始時には日本、韓国などが中心で受験者も日本では3000人ほどであったが、2004年には日本だけで

143万人余が受験している。現在では世界約60ヶ国で実施され年間の総受験者は約450万人に登る。結果は合否ではなくTOEFL同様スコアで示され、その範囲は100点から990点の間である。テストの構成は以下の通りである。

＜TOEIC の構成＞

Section Ⅰ	Part 1-4	Listening Comprehension	100問	45分	計120分
Section Ⅱ	Part 5-7	Reading	100問	75分	

ただし公開テストに関してはより信頼性と妥当性の高いテストにするために、リスニングで使用される英語の種類をアメリカ英語のみから英国、オーストラリアの英語など幅を広げる等、2006年5月より一部の問題形式や内容が変更される予定である。大学等で団体受験を行っているTOEIC IPに関しては、変更の時期は遅れて2007年以降になる模様である。

3．2.「科学技術英語Ⅰ」の統一テスト

次に「科学技術英語Ⅰ」の統一テストであるが、これは名古屋工業大学の一年生対象に前期・後期の期末試験として実施している。つまりこのテストはアチーブメントテストとして位置づけられ、前記のように学生の成績の半分はこの統一テストに基づいてつけられている。統一テストは当然ながら統一教材の使用を前提とするが、2004年に施行された新カリキュラムの「科学技術英語ⅠおよびⅡ」（必修科目）においては一年生も二年生も統一の教材を使い、その授業内で学んだことを測定する意図で統一の期末試験を複数の教員が担当し、毎年新しいバージョンを作成している。

前期の統一テストは全部で100問、マークシートに解答する多肢選択問題で、1問－30問がリスニングセクション、31問－100問がリーディング、文法＆語彙のセクションである。全て前期に授業で使用した科学技術英語のテキストに基づき、関連した問題が出題されている。問題形式としては100問中4肢選択が61問、その他は表中の数カ所の空所に同数の語句群から適切な語を選んで入れるなどの形式が29問である。

後期の統一テストもほぼ同じような構成であるが、リスニングセクションの問題数は全部で32問、リーディング、文法＆語彙のセクションが68問である。前期も後期も統一テストのリスニングの問題はTOEIC形式に従い、問

題のインストラクションは全て英語で行われ、問題文は一回だけしか聴くことができない。また試験時間は全部で90分である。

＜統一テストの構成＞

前期統一テスト	1～ 30問	Listening Comprehension	15分	計90分
	31～100問	Reading, Vocabulary, Grammar	75分	
後期統一テスト	1～ 32問	Listening Comprehension	20分	計90分
	33～100問	Reading, Vocabulary, Grammar	75分	

4．良いテストとは何か

既存の標準テスト TOEIC はテスト作成の専門家集団が作成したテストとして、「良いテスト」であることは十分検証されているはずである。従ってここでは、「統一テスト」は果たして「良いテスト」と言えるのであろうかということが問題となる。

「良いテスト」とは何か、という問題は、そのテストの置かれる社会的状況、受験の目的、テストのスコアの付け方や結果の用いられ方、など、多くの視点から論じられている。例えば Bachman & Palmer (1996) はまずテストの有用性と質について論じているが、その前提として Target Language Use domain (TLU domain「言語使用領域」) に論拠を置くべきだとしている。つまりテストは TLU domain を念頭において作成されるべきであって、作成されたテストの有用性は抽象的に評価されるべきものではない。ある特定の状況における特定のテストに関して論ずるときに初めて、有用性の有無に関する議論が可能となるわけであり、一般的抽象的に「良いテスト」というものが存在するわけでなない。

さらに、Bachman & Palmer (1996) はテストの質 (quality) を決める要素として reliability (信頼性)、construct validity (構成概念妥当性)、authenticity (真正性)、interactiveness (相互性)、impact (影響)、practicality (実用性) の 6 点を挙げている (p19)。Authenticity とは、例えばある問題に正解したときに測定された能力は実際の場面 (TLU domain) でも同じように発揮できる能力と一致しているかどうかということであり、interactiveness はテストの受験者の個々人の特質がテストの課題をやり遂

げる際にどの程度どのような形で関わって来るのかということ、impact はテストが社会や教育制度、あるいはその制度の中にいる個人に与える影響のことであり、試験の washback effect（波及効果）もこれに含まれる。practicality はこれまで説明された要素とは異なり、テストがどのように実施されるのか、あるいはそもそも実施されうるのかどうかということに関わっている。しかし "Two of the qualities − reliability and validity − are, however, critical for tests, and are sometimes referred to as essential measurement qualities." と述べ、テストの質を決める本質的な要素は信頼性（reliability）と妥当性（validity）であると明言している（P17-42)。では次に、テストの信頼性と妥当性とは何であるかを論ずることにしたい。

5．信頼性について

5．1．信頼性とは

テストの信頼性に関する先行研究は数多あるが、代表的な信頼性の定義として以下その幾つかを列挙する。Bachman & Palmer（1996）は信頼性とは "consistency of measurement" 「測定の一貫性」であり、"a function of the consistency of scores from one set of tests and test tasks to another" (p19) 一つのテストからもう一つのテストに渡る得点の一貫性という機能である、と述べている。 McNamara（2000）は同様に、信頼性とはテストによって測定される個々人の測定の一貫性、これは通常、信頼性係数で表される、と述べている。 "Consistency of measurement of individuals by a test, usually expressed in a reliability coefficient" (p136)

最後にAlderson et al. (1995) を引用する。Alderson et al. は、"the extent to which test scores are consistent: if candidates took the test again tomorrow after taking it today, would they get the same result (assuming no change in their ability)? " (p294) と述べている。つまり、テストの信頼性とはテスト得点の一貫性の程度であって、信頼性が高いテストは、被験者がもし今日受けたテストを明日再び受けたとしても同じ結果を得られるということなのである。

以上をまとめると、テストの信頼性とは、あるテストによってある受験者を測定する際の測定の一貫性、と定義できる。ある受験者のある能力を測定

する場合に、同じものを測定するように作成されたテストであれば、別の状況下で行っても同じ結果を得られる、という意味での一貫性を持つことがテストの信頼性である。

5．2．信頼性の測定

この信頼性を測定する方法については多くの研究者が様々に分類しているが、代表的なものは以下の4つの方法である。

①再テスト法：同一のテストを期間をあけて同一の被験者に実施し、その二回の結果の相関をとる方法

②平行テスト法：同等同質な二つのテストを同一の被験者に実施し、その相関をとる方法③折半法：二回テストをせずに、一つのテストを予め決めた方法で半分にして、二つの部分の相関をとる方法

④内部一貫法：一つのテストの内的な整合性を見る方法であるが、Hughes (1989) のように③をこれの一部とする考え方もある (p41)。

実際的には、①は同一のテストを行うために、一回目のテストの記憶が二回目のテストの結果に影響を及ぼす可能性、またその影響を少なくするために一定の期間をあけた場合は、その間に被験者の能力に変化が生じる可能性、などの問題がある。また②に関しては、完全に同等同質なテストを作成することが困難である、という問題があり、③は二分割されたテストの一つの部分が他のもう一つの部分と等質性をどうやって保障するかという問題が生じる。これらの問題を解決する方法として④の内部一貫法が考案されたが、これは全ての分け方に対して相互の相関をとりそれを平均化するというもので、これが信頼性係数のクロンバックαと呼ばれる係数である。この式は信頼性係数としては最も一般的に使用されるものであるので、計算式を参考までに以下に示す。

クロンバックα係数

＝項目数／(項目数－1)×(1－(各項目の分散の合計／合計点の分散))

5．3．信頼性と妥当性

テストの信頼性と妥当性は深く関わりあっており、一方を欠いて他方を論ずることはできない。まず、信頼性と妥当性に関して代表的な研究者達がどう述べているかを見ることにする。

Hughes (1989) は信頼性と妥当性について "To be valid a test must provide consistently accurate measurements. It must therefore be reliable. A reliable test, however, may not be valid at all." (p42) と述べている。すなわち、テストの信頼性というのは妥当性を有するための必要条件であって、十分条件ではない。あるテストが信頼性はあっても、妥当性がない場合もありうるわけである。

Backman (1990) は同様に信頼性を高めることが妥当性を持つことの必要条件であるとしている。"When we increase the reliability of our measures, we are also satisfying a necessary condition for validity: in order for a test score to be valid, it must be reliable." (p160) そして両者は "complementary aspects of a common concern in measurement — identifying, estimating, and controlling the effects of factors that affect test scores" (p160) とあるように、テストの得点に及ぼす影響を見つけ、推定し、コントロールするという共通の利害を持った相補的な視点である、と述べている。

6．妥当性について

では、次に、良いテスト作成のための大きな二つの柱の一つ、妥当性について考察を進めることにしよう。信頼性はテストの結果が受験者の能力をどれだけ正確に測定しているかを表す指標であった。しかし、「何を測定するのか」という測定の中身は、ある意味では信頼性以上に重要な観点であり、測定するべきものを測定していなければいくら正確に測定してもテストを実施する意味がないとさえ言える。

6．1．妥当性とは

妥当性とは何をさすのだろうか、これまでの代表的研究者（研究機関）による妥当性の定義について年代順に整理してみることにしよう。まず第一にアメリカのテスト作成のガイドラインと言われるAERA, APA, & NCME. (1985)のStandards for educational and psychological testing では第一章に以下のように記されている。

"Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. .... Validity is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself." (p9) (下線部筆者)

ここでは、妥当性は、テストの点数から引き出されるある特定の推論の「適切さ、有意味性、有用性」であると言われている。その妥当性は様々な証拠(evidence)の積み重ねによって示されるものであるが、妥当性そのものは単一の概念(unitary concept)であり、内容的妥当性、構成概念妥当性、基準関連妥当性という分類は証拠を分類したものと考えるべきである。妥当性の有無が問われるのはテストそのものではなく、ある特定のテスト結果の推論（解釈）である。

Henning (1987) は妥当性に関して簡単に以下のように述べている。

"Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term valid when used to describe a test should usually be accompanied by the preposition for. Any test then may be valid for some purposes, but not for others. (p89) (下線部筆者)

妥当性とは、行ったテストあるいはその構成要素が測定しようと意図したものを適切に測定することである。測定しようとする対象を確実に測定しているか、それが重要である。

Messic (1988) は次のように記している。"an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (p33) つまり妥当性とは経験的な証拠と理論的な正当性に基礎を

持つものであり、テスト結果に基づく推論とその結果としての行為がどの程度十全かつ適切に行われたかを示すものである。

Hughes (1989) によると"…,in recent years the term construct validity has been increasingly used to refer to the general, overarching notion of validity." （下線部筆者）とあるように、構成概念妥当性が妥当性の中心概念となっていると述べられている。また、この後に続く箇所で、構成概念妥当性があると言うためには経験に基づいた証拠（empirical evidence）が必要であることが説明されている。

Hughes (2000) はここで、構成概念妥当性（construct validity）の部分として内容的妥当性（content validity）と基準連関妥当性（criterion-related validity）を挙げ、基準関連妥当性の下位概念として並存的妥当性（concurrent validity：同時期に受けた他のテストとの比較に基づく。）と予測的妥当性（predictive validity：後に受けるテストと比較。予測力があるかどうかを問う。）を説明している。又、内容的妥当性を高めるためには、スキルのスペック（specification：細目の決定）が必要であり、そのスペックとテスト内容とを比較することによって、妥当性を検証する方法を提案している（p26-27)。

6．2．構成概念妥当性

では、構成概念妥当性とは何を示すのであろうか。これについてはかなり複雑かつ多岐にわたる定義が様々な研究者によりなされているが、ここでは簡単に以下Alderson, et al. (1995)の引用をしておくことにする。

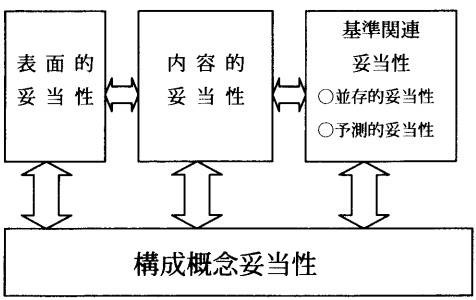
> "A shorter explanation is provided by Gronlund 1985, who describes construct validation as measuring 'How well test performance can be interpreted as a meaningful measure of some characteristic or quality' " (p58).

ここではGronlund（1985）が引用されている。構成概念妥当性とは「テストがある特質あるいは性質の有意味な測定として、いかにうまくperformするかを測定すること」である。構成概念とは、具体的なものではなく、心理学的に構成された抽象的概念であり、多くのテストを行い、その分析によって得られる結果と予測された結果を合わせることによって明らかになってくる測定対象のある特性である。あるテストが、測定しようとする構成概念を

適切に測定していると判断される場合、構成概念妥当性の高いテストであると言われるわけである。

清水（2005）は妥当性の定義の変遷に関して大変詳細な議論を行っているが、その中でも述べられているように、妥当性の概念は様々な種類に細分化される時期を経て、現在は大きな一つの概念としてまとまった形で捉える考え方が主流になっている。それは本章の冒頭で触れた、AERA, APA, & NCME.(1985) にある "Validity is a unitary concept." (p11) という言葉に端的に言い表されている通りでああるが、これは換言すれば、一度分類された様々な妥当性の概念の根底に構成概念妥当性が常に色々な形で関連していることを示しているに他ならない。

ここまでの妥当性に関する議論をまとめて図示すると以下のようになる。



6．3．妥当性の検証

さて、前章で妥当性の概念が一つの大きなまとまり「構成概念妥当性」として捉えられるようになったことを述べたが、それではあるテスト（あるいはテスト実施）の妥当性というものはどのようにして検証するのであろうか。清水（2005）は妥当性の検証に関して次のように述べている。

「妥当性をどのように捉えるにしても、その検証には唯一の方法というのがあるのではなく、複数の観点から検証しなければならないのが，多くの研究者に共通した見解と言えよう。具体的な方法としては、相関関係や因子分析、事前－事後テスト間の差の有意性の検証や性別や言語背景などをもとにしたグループ変数間の有意性の検証などを行うことになろう。」（p251）

このように、妥当性の検証は AERA, APA, & NCME.(1985) でも示されたように、多面的な「証拠の蓄積」によって初めてなされると言える。

7．TOEICと統一テスト

7．1．　TOEICと統一テストの位置づけ

テストの分類の仕方には様々な視点があるが、テスト結果の使用目的による分類には、①能力テスト（proficiency test)、②アチーブメントテスト、③診断テスト（diagnostic test)、④プレースメントテストの4種類がある。能力テストはそれまでに受けた教育やトレーニングに拘わらず、ある時点でのある課題をこなす能力を測定するものであり、アチーブメントテストは能力テストとは反対に、授業やトレーニングコースに直接関わり、被験者がその中でどの程度コースの目的に到達したかを測定するものである。これは各被験者の到達度の評価測定であるのみならず、コースそのものの評価としても使用することができる。アチーブメントテストには、さらに、コースの途中で行われる進捗情況テスト（progress achievement test）と最後に行われる最終テスト（final achievement test）の二種類がある。診断テストは、被験者の学習上の弱点や問題点を特定するために使用され、プレースメントテストは被験者を各人の能力に最も適切な学習レベルに振り分けるために使用される。

TOEIC はもともと能力テストとして開発されたものであるが、現実的には多くの大学や企業で色々な目的で用いられている。その一つがプレースメントテストとしての使用である。本学では現在 TOEIC は一年生の最初にプレースメントテストとして用いられ、また一年の最後にはある種のアチーブメントテストとして実施されている。しかし問題は、一般的コミュニケーション能力を測定する TOEIC を使って、その結果をもってレベル別に編成される授業（コース）は科学技術英語を内容とするという点である。この言わば「ズレ」は TOEIC がプレースメントテストとして導入された2003年に置かれていた必修の一般英語科目が、2004年の新カリキュラムの開始とともに無くなり、英語の必修科目は科学技術英語のみとなったことに起因する。現実的な問題としては一度大学として導入した TOEIC を一年のみで中止することは難しく、また学生の英語力の経年的な変化を測定するためには TOEIC のような標準的テストを長期的に実施することには大きな意義がある。しかし、TOEIC をプレースメントテストとして用いることに関しては依然として問題が残ると言えるだろう。

統一テストは当然ながら、「科学技術英語Ⅰ」のアチーブメントテストとして作成され、実施されている。このテストの作成には数人の英語担当教員が当たり、統一教材の内容に基づいて出題内容と出題形式を決定し、最終的には英語担当教員全員で検討を行っている。回答は全てマークシート方式で行い、問題用紙も全て回収するなど、テストセキュリティーの点でも配慮されている。

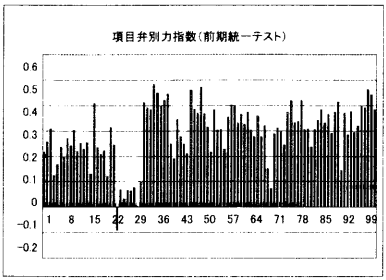
TOEIC はテスト作成の専門家集団が作成したテストであり、能力テストとしての信頼性や妥当性は十分検証されていると思われるため、以下、統一テストに関して信頼性に関する統計的な検証を行い、その後、TOEIC を含めてこれら4回のテストの妥当性についてカリキュラム全体を視野に入れて論ずることにしたい。

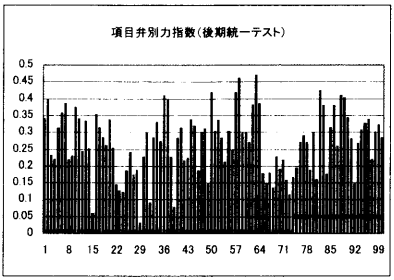
7．2．TOEIC と統一テストの結果

まず、統一テストの信頼性であるが、信頼性係数の一つであるクロンバック α に関しては、統一テストは前期0.8807、後期0.8650であり、いずれも高い値を示している。これは、項目の数が100と大きく、また多肢選択問題のみで採点者による差異がなく、選択肢の数も4つの問題が大半を占めるなど、信頼性を高める要素が多いためと考えられる。

また、ある課題に関して能力の高い受験者と低い受験者を正確に識別できるということは、テストの信頼性の一つの重要な要因であるが、その項目弁別力指数（item discrimination index）はテスト項目が適切なものであるか、改良すべきものであるかを示す指数となる。池田（1992）によると「良好と思われる問題」は0.4－1.0、「準良好な問題」は0.3－0.4、「要検討な問題」は0.2－0.3、弁別力指数が0.2以下の問題に関しては、「不良な問題」としてテスト項目から除外されるべきだとされる。項目弁別指数とは、受験者の合計点の数列と各項目の正解不正解の相関を取ったもので、前期後期の統一テストの、各項目の指数は以下のグラフのようになった。

このグラフに示されるように、前期後期ともに統一テストの項目弁別力指数は0.2未満のものが後期に関しては25問、前期に関しては17問あった。これらの項目に関しては今後見直しの対象とすべきであろうが、項目の弁別力という観点から見ると平均して80％ほどが一応の基準を満たしていたことになる。





また項目の難易度であるが、前期テストは100項目中、正解率が0.2未満のものが2問、0.8以上のものが33問、後期は同じく3問と29問あり、全体として正解率の高いものが多く、それが平均点の高さ（以下の基本統計表参照）となって現れている。正解率の高さは正規分布を目指す集団基準準拠テスト（Norm-referenced test）の場合は問題となるが、目標基準準拠テスト（Criterion-referenced test）の場合は、必ずしも問題とはならない。正解率が高い問題が多いということは、適切に作成された目標基準準拠テストの場合は、設定した基準を超える被験者が多いということを示す。即ちアチーブメントテストの場合はコース内で設定した目標に到達した学生が多いということになるからである。また、標準偏差はTOEICの方が極めて大きく、一回目二回目、また上中下位のレベルを問わず、点数のばらつきが大きいことを示している。

この研究の対象となった、4つのテストの基本統計を以下の表に示す。

＜トータルスコアの基本統計＞

	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
実施	05年4月	06年2月	05年7月	06年1月
最大値	965.00	975.00	94.000	94.000
中央値	400.00	420.00	65.000	71.000
最小値	150.00	105.00	24.000	21.000
平均	402.47241	415.94923	64.633554	69.459161
標準偏差	86.659694	99.225641	11.011111	11.261678
平均の標準誤差	2.8790755	3.2965511	0.3658197	0.3741442

＜レベル別基本統計＞

	上　位			
	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
平　均	497.40066	481.62252	75.738411	71.821192
標準偏差	55.976221	89.639458	7.668015	8.2180307
平均の標準誤差	3.2210695	5.1581712	0.4412447	0.4728945
	中　位			
平　均	398.75828	419.88411	69.738411	66.006623
標準偏差	20.850628	70.942187	8.0427294	7.4641973
平均の標準誤差	1.1998188	4.0822641	0.4628071	0.4295163
	下　位			
平　均	311.25828	346.34106	62.900662	56.072848
標準偏差	40.108733	85.617396	13.230266	10.678696
平均の標準誤差	2.3079982	4.9267275	0.7613162	0.6144899

これら4つのテストの相関は以下の表の通りである。これらの試験については、二回のTOEICは実施時期が、統一テストは実施時期も内容も異なるものであるが、相関係数を見る限りでは全体として比較的高い相関を示している。一番相関の高いのは予想に反してTOEIC ⅠとTOEIC Ⅱの0.6779ではなく、後期統一テストとTOEIC Ⅰとの0.6796、それに次いで高いのはやはり後期試験とTOEIC Ⅱとの0.6793であった。

4テスト相関	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
TOEIC Ⅰ	1.0000	0.6779	0.5730	0.6796
TOEIC Ⅱ	0.6779	1.0000	0.5913	0.6793
前期統一テスト	0.5730	0.5913	1.0000	0.6704
後期統一テスト	0.6796	0.6793	0.6704	1.0000

次に第一回目のTOEICの得点により、学生を三グループに等分して、得点の高い順に上位、中位、下位のグループに分け、それぞれの相関を出してみた。その結果は以下の三つの表の通りである。係数の下線の引いてあるものは、0.5以下の相関の低いものである。

上位相関	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
TOEIC Ⅰ	1.0000	0.6028	0.3796	0.4673
TOEIC Ⅱ	0.6028	1.0000	0.4267	0.5114
前期統一テスト	0.3796	0.4267	1.0000	0.5687
後期統一テスト	0.4673	0.5114	0.5687	1.0000

中位相関	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
TOEIC Ⅰ	1.0000	0.1579	0.1961	0.2150
TOEIC Ⅱ	0.1579	1.0000	0.3542	0.3980
前期統一テスト	0.1961	0.3542	1.0000	0.4239
後期統一テスト	0.2150	0.3980	0.4239	1.0000

下位相関	TOEIC Ⅰ	TOEIC Ⅱ	前期統一テスト	後期統一テスト
TOEIC Ⅰ	1.0000	0.4830	0.5434	0.5568
TOEIC Ⅱ	0.4830	1.0000	0.5460	0.6205
前期統一テスト	0.5434	0.5460	1.0000	0.6135
後期統一テスト	0.5568	0.6205	0.6135	1.0000

この三グループの相関係数を見ると、上位グループはやはり第一回目のTOEICの得点により分けられたためか、TOEIC二回目との相関が0.6028と一番高く、次が前後期の統一テストの相関0.5687であった。これは似たテスト間の相関が高いという比較的予想できる結果であった。しかし中位グループの相関はかなり低く、全ての相関が0.5を下回った。最も高い統一テスト同士の相関係数でも0.4239で、最も低いTOEIC同士は0.1579という値であった。これは中位の学生が試験ごとにパフォーマンスの変動が激しく、一定の能力を発揮していないということを示している。

次にリスニング（L）とリーディング（R:文法、語彙を含む）部分の相関をとってみたのが以下の表である。まずリーディングの相関であるが、全体に0.5以上と高い相関を示している。最も高かったのは、前期後期の統一テストであり、相関係数は0.6587とTOEIC同士よりも高い相関であった。リスニングに関しては、リーディングと比較して全体的に相関が低く、0.5以下のものがほとんどである。最も高いTOEIC同士の相関係数でも0.5720

にとどまっている。後期の統一テストに関しては、TOEICとリスニングの相関が比較的高かったのが特徴的である。

Rの相関	TOEIC I	TOEIC II	前期統一テスト	後期統一テスト
TOEIC I	1.0000	0.5845	0.5327	0.5964
TOEIC II	0.5845	1.0000	0.5175	0.6337
前期統一テスト	0.5327	0.5175	1.0000	0.6587
後期統一テスト	0.5964	0.6337	0.6587	1.0000

Lの相関	TOEIC I	TOEIC II	前期統一テスト	後期統一テスト
TOEIC I	1.0000	0.5720	0.3627	0.5349
TOEIC II	0.5720	1.0000	0.4523	0.5315
前期統一テスト	0.3627	0.4523	1.0000	0.4088
後期統一テスト	0.5349	0.5315	0.4088	1.0000

なお2004年から2005年のこれら4テストの結果についてはQuinn(2006) に詳しく論じられているので、これを参照されたい。

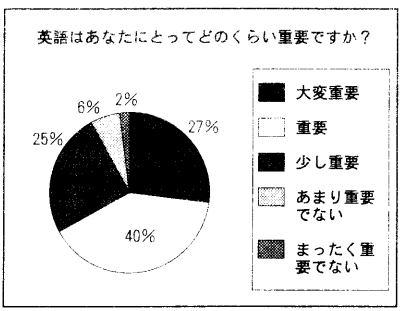
8．アンケートの結果

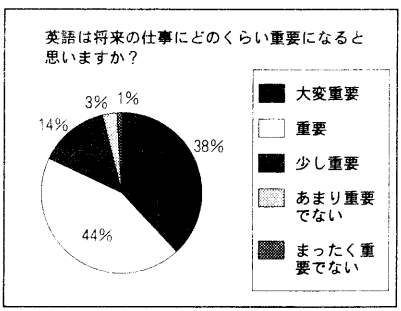
今年度末、1年次の学生を対象に英語授業に関連するアンケートを行った。自由参加であるので回答者総数は170人と、全体の5分の1以下であるが、英語授業やテストに対するある程度の傾向を見ることはできるであろう。教科書等に関する質問項目も含まれていたが、ここでは、本研究に関連する項目だけに限定して考察を加えることとする。

英語の重要性に関しての質問に対する結果は以下の表の通りである。これを見ると、学生が英語が自分にとって重要であると考えていることが解る。特に将来の仕事に関連して、英語の重要性が増してくると考えているようである。大変重要・重要と考える学生は、現在のこととして67%、将来のこととしては82%に達する。(表中の単位は人、未解答はカウントしていない。)

英語の重要性に関して	大変重要	重要	少し重要	あまり重要でない	まったく重要でない
英語はあなたにとってどのくらい重要ですか？	44	67	40	9	3
英語は将来の仕事にどのくらい重要になると思いますか？	63	71	23	5	2

円グラフに表示すると以下のようになる。



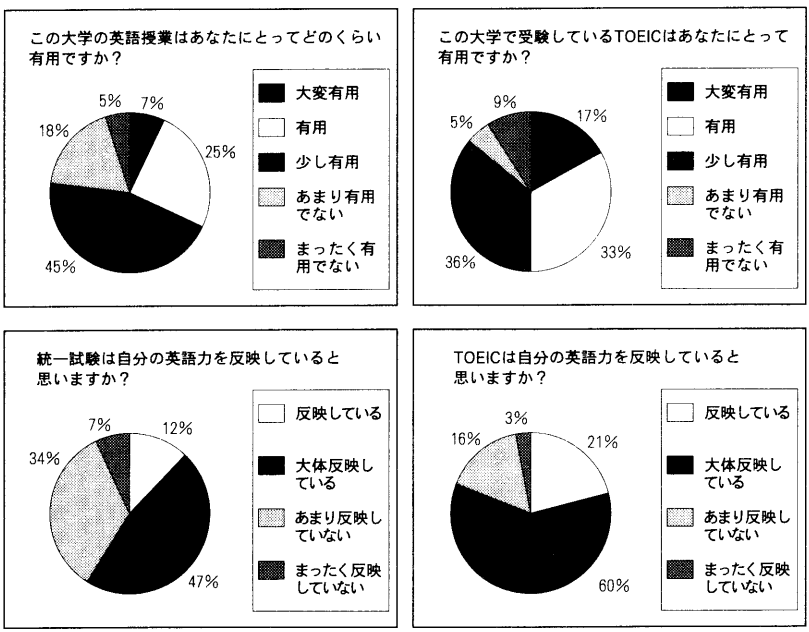


次に、英語の授業とTOEICの有用性に関するアンケート結果を以下の表に示す。この回答で注目すべきことは、授業も有用であるが、TOEIC受験はさらに自分にとって役立つ、つまり意味があると考えている点である。持続的に受けている授業よりも、年2回受験するだけのTOEICの持つ意味の方が大きいと感じていることからも、TOEICの社会的インパクトの強さが如何に大きいかが解る。この傾向は英語力の反映に関する質問によってさらに明確になる。

有　用　性	大変重要	重要	少し重要	あまり重要でない	まったく重要でない
この大学の英語授業はあなたにとってどのくらい有用ですか？	11	41	74	29	9
この大学で受験しているTOEICはあなたにとって有用ですか？	29	57	63	9	16

英語力の反映	反映している	大体反映している	あまり反映していない	まったく反映していない
統一テストは自分の英語力を反映していると思いますか？	20	77	55	12
TOEICは自分の英語力を反映していると思いますか？	35	97	26	5

円グラフでパーセンテージを表すと以下のようになる。



英語の授業の有用性に関してはTOEICに比べて、やや厳しい結果が出た。大変有用と有用を合計すると32%、しかし有用でない・まったく有用でないの合計はやはり23%あった。TOEICに関しては大変有用・有用を合計すると50%に上る。これは上記のように、TOEICに対する現在の日本の社会的ニーズの高さを見事に反映しているからに他ならない。学生は、さらに一年間受けてきた授業の期末試験である統一テストよりもTOEICの結果が自分の英語力を反映していると感じている。TOEICに関しては「反映している・大体反映している」の合計は81%という非常に高い率に上る。しかし、統一

テストに関してもこの合計が59%あることを考えると、学生は統一テストも自分の英語力を反映していると判断しているようだ。それにしても、入学時はTOEICに関してほとんど知識の無かった学生が、一年後にはTOEICが自分の英語力を統一テストより正確に測定し、しかもその結果は自分にとって大変有用であると考えるようになったわけである。この考え方の変化は著しいものがあるが、これはTOEICの表面的妥当性の高さによるものであろう。

9．結　論

これまで本学で実施している4テストの現状とテスト結果、またこれらのテストや英語の授業に関する学生のアンケート調査結果について述べてきたが、以上で明らかになった点をここにまとめ、さらに今後のカリキュラムの検討課題の材料を提示することにしたい。

まず信頼性という観点から検証した結果、統一テストは、項目弁別力の悪いものを除去するなどの改良が必要であるが、信頼性係数は十分高く、信頼性の高いテストと言える。

次に、4回のテストスコアの相関であるが、トータルスコアで見た場合はいずれも比較的高い値を示した。これらを総合的に判断すると、統一テストも学生の英語力を適切に反映していると言えるだろう。リスニングに比べると、リーディングセクションの相関が高いということも明らかになった。またレベル別に見ると、中位の相関が低く、これについては更なる考察が必要である。

しかし、今回は正答数に基づくいわゆる「古典的テスト理論」に則って諸分析を行ったため、中村（2002）が指摘するように分析が被験者グループに依拠するという限界がある。今後はIRT（項目応答理論）を含めた新しい理論に基づいた分析を行い、新しいテスト作成を試みる必要があるだろう。しかし、コースの目標を明確にして、授業内で教える項目をリストアップし細目を規定する、またテスト目的を記述し測定すべき構成概念を定義する等、まずはより良いテストを作成する基本的取り組みを、教員全体が自覚的に始めるべきであろう。アチーブメントテスト作成としては、集団基準準拠（NRT）ではなく、コースの目標に応じて設定された基準をゴールとする目

標基準準拠テスト（CRT）が適切であり、小山（1996）が提案したようにCRLTD（Criterion-Referenced Language Test Development:目標基準準拠テスト開発）はそれを実現するための一つの具体的方策と考えられる。

また、アンケートの結果からも明らかなように、学生からのTOEICテストに対する要請は無視できないものがある。今後、TOEICの得点をより積極的に成績に反映させる何らかの手立てを考えるとともに、将来的にはTOEICを継続的にカリキュラムに取り込むことも検討していかなければならない。

世の中には様々な語学学校が林立し、英語学習への関心は高まる一方であるが、社会的ニーズは大学のアカデミックな英語の必要性とは必ずしも一致していない。このような時にこそ、大学の英語教育の果たす役割とその中身に関して英語教員の間で活発に議論し、それぞれの教育現場でこれまで以上に明確な目的を設定することが求められていると言えよう。

参考文献

池田央（1992）.『テストの科学』東京：研究社

小山由紀江（1996）. Criterion-referenced language development の可能性.『LLA関西支部研究集録』6, 55-70

靜哲人等編著（2001）.『外国語教育リサーチとテスティングの基礎概念』大阪：関西大学出版部

靜哲人（2002）.『英語テスト作成の達人マニュアル』東京：大修館書店

清水裕子（2005）. 測定における妥当性の理解のために－言語テストの基本概念として－.『国際言語文化研究所紀要』16(1), 241-254

中村洋一（2002）.『テストで言語能力測れるか』東京：桐原書店

AERA, APA, & NCME.(1985). *Standards for educational and psychological testing*, Washington, DC: American Psychological Association, Inc

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*, Cambridge: Cambridge University Press

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

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge

University Press.
McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press
Messick, S. (1988). The once and future uses of validity: Assessing the meaning and consequences of validity. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp.33-45). Hillsdale, NJ: Lawrence Erlbaum.
Quinn, K. (2006). Creating an English Curriculum for Engineers. 『中部地区英語教育学会紀要』36, 327-332