An Analysis of the Use of External English Language Tests from a Consequential Validity Perspective

日本の高等教育機関における英語能力評価についての一考察：結果的妥当性の視点から

小林美代子

An Analysis of the Use of External English Language Tests from a Consequential Validity Perspective

Miyoko Kobayashi

テストの妥当性とは、測定しようとしているものを正確に測定できているかどうかを示すための指標であり、テストを開発・使用する際、考慮すべき最も重要な特質である。Messick（1989）は、結果的妥当性という概念を提唱し、テストの妥当性を確立するためには、得点の解釈と使い道、さらには社会への影響を考慮する必要があることを提言した。現在、日本の多くの高等教育機関で、外部テストを選抜や単位認定等の目的で教育カリキュラムに取り入れる傾向が見られるが、教育機関の評価の目的と外部テストの本来の目的が必ずしも合致しているとは限らない。本稿は、Messick の提唱する結果的妥当性という観点から、この問題を考察し、神田外語大学におけるテスト開発・妥当性検証の試みを紹介しながら、教育プログラムと連動した評価の重要性を論じる。

＊言語評価　＊テストの妥当性　＊神田英語能力テスト　＊テストの社会的責任

1. はじめに

テスト結果は、入学選抜、昇進・配属の決定など、我々の人生の岐路を決定するような重要な判断材料として利用されることが多い。そのため、テスト結果を利用する際には、テストの得点が持つ意味を正しく理解し、適切な判断をすることが必要となる。本稿は、Messick（1989）の結果的妥当性の枠組みを基に、現在日本の多くの教育機関で見られる外部テスト利用の状況について

考察した後、神田外語大学におけるテスト開発・研究の取り組みを紹介し、教育機関における望ましい英語能力評価のあり方について提言する。

テストの妥当性とは、測定しようとしているものが何であれ、それを正確に測定できているかどうかを表す指標であり、テストの開発及び使用の際に、考慮しなければならない重要な特質である。妥当性を検証するために、「このテストは、測定すべき能力を測定しているか。またそれ以外のものは測定していないか。」という質問を問いかけることができる。つまり、使用するテストが、測定したいと思う能力や知識を正しく測定する道具であることを確認する必要があるということである。ものの重さを計るためには、ものの大きさ・重さに応じた適切な「はかり」が必要なように、外国語の能力を測定するためには、それを正確に測定する道具が必要となる。

テストの妥当性の分類として、伝統的に以下の４分類が広く使われてきた。

・構成概念的妥当性
・内容的妥当性
・並存的妥当性
・表面的妥当性

これらについては、類書（Heaton 1988; Hughes 1989など）に詳しいので、ここでは詳細な説明は割愛するが、テストが測定しようとしている能力や知識の概念的理解が明確にできていて、確かにその能力・知識を測定しているか（構成概念的妥当性）、測定しようとしている能力・知識を満遍なく測定しているか（内容的妥当性）、受験者や教師の目に、測定すべきものを確かに測定しているように見えるか（表面的妥当性）といった側面から、妥当性を検証することができる。また、既に妥当性が検証されている他の測定法と比較し、類似の結果を得ることで妥当性を確証することもできる（並存的妥当性）。さらに、測定に揺れや誤差が少ないこと（信頼性）、人的・経済的資源に見合った測定法であること（実用性）、教育へのよい波及効果があることなども、望ましい測定法の特質として欠かせないものである。

これに加え、近年、テストが持つ社会的影響力に着目し、テストの有益性、

倫理、公正さ、社会的責任といったものも、妥当性検証の重要な要素として考えられるようになってきた（Alderson & Banerjee 2001, 2002; Bachman 1990; Bachman & Palmer 1996; Hamp-Lyons 1997; Messick 1989, 1994, 1996; Shohamy 1997）。特に、テスト得点の解釈、使い道、及びその結果の大切さを訴え、適切な根拠をもって示すことの重要性を強調したMessick（1989）の提唱する妥当性の枠組みは評価に関わる人たちに大きな示唆を与えた。この枠組みの中でMessickは、妥当性を単一概念としてとらえ、構成概念的妥当性がその中核を成すものとしている。妥当性検証に大切なことは「このテストの得点の解釈と使い道が妥当であることを支持する証拠は何か」と問いかけ、適切な根拠を集めることである。下の表1はMessickの妥当性の枠組みを示すが、Messickはこれを Progressive Matrix と呼び、左上から右下に進むに従って、一つ一つの根拠が加算されていく過程を示した。

表1：Messickの妥当性の枠組み（Messick　1989）

	テスト結果の解釈	テストの使用
証拠としての根拠	構成概念的妥当性	構成概念的妥当性 ＋使用の適切さ
結果的根拠	構成概念的妥当性 ＋価値判断	構成概念的妥当性 ＋使用の適切さ ＋価値判断 ＋社会的結果

この枠組みでは、テストそのものの質を問う従来の妥当性は、主に左上の単独の構成概念的妥当性に該当すると考えられるが、それに加えて、結果の解釈及び使用、そしてテストがもたらす社会的結果に配慮することの重要性が強調されている。使用の適切さとは、テスト作成の意図に合った使い方がされているかどうか、価値判断とは、テスト結果がどのような価値判断の基準になるのか、テスト結果を解釈する際の判断が適正であるかどうか、ということを意味する。さらに、教育や社会に及ぼす影響がテストの本来の目的に沿ったものであるかどうかまで検証することを妥当性検証の枠組みの中に取り入れたことは

非常に意義深い。これは、これまで波及効果という概念で考えられてきたものをさらに発展させ、それを妥当性検証の中核に据えた画期的な枠組みである。この考えによれば、テストそのものがどんなに精度の高いものであっても、本来の目的に合致しない解釈や使用をした場合には、テストの妥当性は低いものとなる。これは、取りも直さず、テストの妥当性というものが、テストそのものに内在する絶対的なものではなく、テストの使用目的や状況によって変わり得る、相対的なものであることを意味する。「妥当性追求は、終わることのない活動である」とされる所以である[1]。

では、実際にどのように妥当性を検証したらよいのか、その方策について見てみよう。Alderson, Clapham & Wall（1995）は、妥当性検証のためのチェックリストを挙げているが、その中に「受験者の反応を見る」という項目を含めている点が注目できる。これは、従来の表面的妥当性を超え、受験者にどのように回答したかを内省させることによって、回答のプロセスを解明しようとする試みで、受験者が回答の際に、テスト作成者が意図した能力を実際に用いたかどうかを検証する手法を提示している。このような情報を得ることで、テスト使用及びテスト得点の解釈の適切さを確認することができる。つまり、テストが作成時の目的及び作成者の意図に適った役割を果たしているかどうかを検証するという点で、Messick が提唱する「使用の適切さ」「価値判断」の正当性を確認すると考えてよいであろう。同様に、Luoma（2001）も、テストの妥当性検証の方策の枠組みを提唱しているが、その中で、従来の妥当性検証に加え、「影響」「結果」に関する項目も含めており、Messick の考えを反映させたものと考えられる。たとえば、よい影響を与え、望ましい結果を得るための手段として、「起こり得る波及効果について、学習者・教師・教材開発者・カリキュラム作成者・研究者など、広範囲にわたる関係者から情報を集めること」、「テスト開発のプロセスに関する情報を綿密に記録し、測定手段の特質を周知のものとし、得点解釈のための指標を示すこと」などを挙げている。

妥当性検証の実践例としては、Chapelle, Jamieson, & Hegelheimer（2003）が、実際のテスト開発の際に、結果的妥当性を含めた意味での妥当性を一つ一つ検証していく過程を示しており、テスト開発・妥当性検証に関わる者に大きな示唆を与えてくれる。また、Cheng & Watanabe（2004）は、結果的妥当性の

中でも最も重要な側面である「波及効果」の理論的枠組みを示した上で、世界のさまざまな状況における波及効果に関連する研究を紹介している。アンケート、授業観察、インタビューなど、波及効果検証のために必要なデータ収集の具体的な手法を豊富に例示しており、妥当性検証研究の新しい方向を示している。また、波及効果を焦点に当てた研究を詳細に紹介しているChen（2005）及びWall（2005）、Messickの妥当性の枠組みに従って妥当性検証の研究論文を分類・掲載したCumming & Berwick（1996）も示唆に富む。

以上、ここまでで、妥当性という概念が、テストの内在的な正確さや信頼性のみならず、テストの使用や得点の解釈、そして、社会的影響までも考慮に入れるべきものであることを確認し、妥当性検証には、適切な「根拠」に拠ることが必要であることを認識した。次節では、この観点から、日本の高等教育機関における外部テスト利用の状況について考察する。

2. 外部テストの利用状況

近年、日本国内で急速にその普及度を高めているTOEIC®（Test of English for International Communication国際コミュニケーション英語能力テスト）について見てみよう。TOEIC®は、1970年代に、日本の勤労者を主たる対象として、「英語コミュニケーション能力を測定する」ことを目的に謳ったテストである。アメリカのテスト開発機関Educational Testing Service（ETS）により開発・制作され、（財）国際ビジネスコミュニケーション協会TOEIC®運営委員会が運営する。「世界共通のテストであり、世界約60ヶ国で実施」と謳ってはいるが、1997年の受験者の92%は日本と韓国であり、62%が日本人受験者である（The Chauncey Group 2000, cited in Chapman 2005b）。受験者は、大多数が企業雇用者で、製造業に従事する人の割合が高いが、近年は、教育機関でも幅広く利用されるようになってきた。

表3はTOEIC®運営委員会が2004年9月より10月に行った独自調査に基づく資料である[2]。「全国の大学院・四年制大学・短期大学・高等専門学校における入学試験（推薦入試、AO入試を含む）および単位認定でのTOEIC®スコア所持者への優遇措置をまとめた」（TOEIC®運営委員会 2004）ものであるが、これによると、非常に多くの教育機関で入学選抜の判断材料や単位認定のため

にTOEIC®の得点を利用していることがわかる。

表３　TOEIC®テスト入学試験・単位認定における活用状況

	大学院	大学	短期大学	高等専門学校	計
調査実施校数	539	696	425	62	1722
入学試験活用校数	81	185	51	5	322
単位認定活用校数	※	230	50	31	311

※大学院における単位認定活用状況については調査対象外

同資料によると、入学試験での活用には、総合判定の際の参考資料として参照する程度とするものから、出願要件、学力試験への点数加算、学力試験免除とするものまで多岐にわたる。また、単位認定のための活用に関しても、認定のための水準及び互換単位数は教育機関によって様々であるが、スコアに応じて認定する単位が加算する方式を取る大学が多い（たとえば、400点で４単位、600点で８単位、730点で10単位など）。認定する単位数は、２単位から20単位程度までと幅があるが、４単位から８単位程度の認定が多い。これは、場合によっては卒業要件に必要な外国語の単位すべてが単位認定で取得可能ということになるのだろうか。公式のTOEIC®テストの他に、IPテスト（各教育機関で実施する非公式のテスト）のみを単位認定に利用している教育機関も見受けられるが、この場合は、大学独自の評価を実施する代わりに、学内でTOEIC®のIPテストを実施し、それを単位認定のための評価の道具として使っているのではないかと推測される。

このように、TOEIC®は、日本の高等教育機関で広く利用されていることが判明したが、このテストは、果たして、教育機関における英語能力評価をするための適切な道具と言えるのだろうか。

テスト内容のトピックを見てみると、勤労者を主な受験対象としているため、製品の開発、金融、ビジネス上の契約や商談、製造ライン、昇進・雇用・年金等人事関係の事柄、ビジネスミーティングや懇親会等での会話など、学生が身近に触れる話題や状況とはかなりずれがあることが伺える。

また、表4で示す通り、TOEIC® はリーディングとリスニングのみのテストで、コミュニケーションに不可欠のスピーキング及びライティングといった表出能力は測定していない。

表4　TOEIC® テストの概要

リスニング　(45分)	多肢選択問題100問
	Ⅰ．写真描写問題　(20問)
	Ⅱ．応答問題　(30問)
	Ⅲ．会話問題　(30問)
	Ⅳ．説明文問題　(20問)
リーディング　(75分)	多肢選択問題100問
	Ⅰ．文法語彙問題　(40問)
	Ⅱ．誤文訂正問題　(20問)
	Ⅲ．読解問題　(40問)
合計120分	合計200問

「英語によるコミュニケーション能力を幅広く評価する世界共通のテスト」[3]と謳うテストとしては、疑問が残る。ETS は独自の研究結果に基づき、スピーキング及びライティングテストとの相関（0.77）が十分に高かったとして、TOEIC® は表出能力も間接的に測定するテストであると主張しているが、Chapman（2005b）は、0.77という相関係数は十分に高いとは言えないと反論している。さらに、日本人受験者を対象とした Hirai（2002）は、典型的な受験者層である450－650点の範囲の得点者に絞った場合、スピーキングテストとの相関係数は0.49であったことを指摘している。このことから、TOEIC® の本来の受験対象者である企業雇用者を考えた場合でも、TOEIC® の点数だけでは、企業が本当に必要としている英語のコミュニケーション能力が測定できていないのではないか、という疑問を投げかけ、日本の多くの企業が TOEIC® の得点を偏重する傾向に警鐘を鳴らしている（Chapman 2005a; Hirai 2002）。もし教育機関が英語の教育目標に、「話す」「書く」といった表出能力も含めているとしたら、単位を認定するためには、その能力もきちんと測定する必要があるのではないだろうか。

また、Bonk（2004）によれば、リスニングとリーディング各部門間の相関

は0.81と非常に高い値となっており、リスニングとリーディングを２つの異なる能力として測定しているかどうかが危惧される。これは、上で述べた、ETS がスピーキング及びライティングも間接的に測定していると主張する根拠となった「十分に高い」という相関係数0.77よりも高いことに着目したい。

Bonk（2004）はさらに、TOEIC® の公式の信頼係数は0.95となっているが、これは、非常に幅の広い層の受験者を基準に算出したものであり、一大学内の学生の場合のように、能力のばらつきの幅が狭い受験者を基準にした場合には、もっと低くなることが想定されると述べている。同様に、Chapman（2005a）は Childs（1995）の研究結果を引用し、TOEIC® は、標準誤差が非常に大きいため、個々の学習者の短期的・中期的な学習効果を測定する道具としては不適切であることを指摘している。

また、TOEIC® の結果は、リスニングとリーディング各々と合計点の3種類のスコアが示されるだけで、受験者には、なぜそのような得点になったのかに関する詳しいフィードバックは与えられない。教育機関における評価は、学習効果や教育効果があったかどうかを確認し、それを次の指導に役立てていくことが大きな役割となっていることを考えると、得点の意味合いが不明確な TOEIC® の得点を利用することは、教育的効果が低いと言わざるを得ない。Messick の言う「使用の適切さ」「結果の解釈」及び「社会的結果」という点において、テストの妥当性に大きな疑問が生じる。

TOEIC® を大学生の英語能力を測定する道具として用いることについて、Bonk（2004）は次のように結論付けている。

- 対象としている受験者像は、勤労者であり、学生ではない。
- テストの内容が学生の身近な体験、教室での学習活動と合致していない。
- 学生が、授業の中で学習・習熟したことを測定する尺度としては不適切である可能性が高い。
- 個々の学生の得点を見ると、受験した時期・テストによって得点にばらつきがある。標準誤差を超える範囲で得点が上下している例が多く見られる。
- 一つの大学内の学生のように受験者層にばらつきが少ない場合、テスト信頼性は、公表されているものほど高くない可能性がある。

以上見てきたように、現在、日本の多くの高等教育機関でTOEIC®の得点が利用されているが、上記のような特徴をもつテストを学生に受験させ、教育機関内の成績とみなすことは、結果的妥当性の観点から、大いに問題がありそうである。入学選抜における利用に関しては、受験者の英語能力を、総括的な指標として客観的に他の受験者と比較することが可能である点を考えれば、「TOEIC®が測定している能力に限界があることを認識した上で」という条件付きで、参考資料として利用することは許容できるかもしれないが、単位認定のための利用となると、大きな疑問が残る。特に、多くの教育機関で、単位認定の基準を満たす得点を得た学生はその単位分の授業が免除になるという方式を採用しているようであるが、これでは、教育目標及び教育そのものの放棄にも等しく、教育の空洞化が懸念される。

3. 神田外語大学の紹介

次に、神田外語大学における英語能力評価の試みについて紹介しよう。神田外語大学は、外国語学部の単科大学で、全学で学生数3000人弱の小規模な大学であり、開学以来の教育目標のひとつとして、実践的な外国語運用能力を学生に身につけさせることを掲げている。その目標達成のために、外国語の授業では、学生主体の「コミュニケーションのやりとり」を重視する学生参加型の教授法を取り入れており、、グループワークや発表の機会が多くなっている。1－2年次のFreshman English courseの大部分を応用言語学・英語教育の修士号をもつ英語のネイティブ・スピーカーが担当しており、2005年度現在、これらのネイティブの教員は総計41名となっている。このような教育目標及び学生のニーズに見合った評価方法の必要性を認識し、開学後まもなく、独自のテスト開発が始まった。

表5は、KEPT（Kanda English Proficiency Test）開発の過程を示す。開発開始より16年が経過した2005年現在、妥当性検証等の研究もかなり進み、5つの平行テストが完成し、これにより、入学前及び入学後の4年間の各学年末の実験が可能となっている。現在、さらに過去の版を見直し、より質の高い平行テストの整備に当たっている。テストの問題作成は、ネイティブの教員のチームが担当しており、妥当性検証の研究にも多くの教員が従事している。

表５　KEPT 開発の過程

1987	開学
1989	KEPT 開発開始
1990	KEPT 第一版施行
1992	平行テストの開発
2000	KEPT 得点、評価に参入
2002	４つの平行テスト　完成；　妥当性検証研究活発化
2003	５つ目の平行テスト完成

４．KEPT の概要

本節では、KEPT についてさらに詳しく紹介する。通例、一回のテストの受験者数は、2300人程度である。受験者は、新入生と在学中の学生で、新入生は大学入学直前の３月末、在学生は一年間の授業終了後の１月末に受験する。新入生は、テスト結果によって進度別のクラスに分けられる。また、在学生は一年間の英語学習の効果を評価されることになる。このため、KEPT はプレースメントテストと進度テストの二つの側面を持つ。

表６は KEPT の構成内容を示す。

表６　KEPT の概要

リーディング	35分	５つのパッセージ；　多肢選択問題30項目
文法	25分	２つのパッセージ；　30項目の空所補充問題
リスニング	30分	５つのリスニング題材；多肢選択問題30項目
エッセイライティング	35分	エッセイ１題
オーラルテスト	15分	グループディスカッション

カリキュラムとの連携を強く意識しており、内容・形式ともに、教室での活動を反映している。たとえば、オーラルテストは、グループディスカッションの形を取っているが、これは通常の英語の授業で多く用いられる形態を再現するもので、１グループ４人で、与えられたテーマについて自由に討議する。試験官は、最初の指示以外、討議には加わらない。

リーディング、リスニング、文法の部門では、多肢選択問題形式を採用しているが、TOEFL® や TOEIC® などの外部テストでは、一つ一つのテスト項目が独立した問題であることが多いのに対し、KEPT ではテスト全体を貫く一つのストーリー・テーマによって、一定の文脈が与えられている。導入として、まずビデオで登場人物の紹介があり、テスト全体がこれらの登場人物が織り成すストーリーを中心に展開する。エッセイ及びディスカッションの題材も、このストーリーに関連したものとなっている。リスニングはビデオ映像を使用し、発話の状況について視覚情報を与え、より現実のコミュニケーションに近いものとなっている。

オーラルとライティングの評価には、ネイティブ・スピーカーの教員全員が参加する。評価基準は学内のウェブに公表されていると同時に、普段から学生も教師も持っており、授業中の活動の評価にも使用している。それによって、学生も教師も日頃から学習の達成目標を明確に認識し、学年末の KEPT 実施の際には、評価基準を熟知した状態で受験することが可能となる[4]。これは、教室での指導と連携させて開発した学内テストの大きな利点と言えるであろう。

テスト結果は、テスト実施後約一週間後には、大学のウェブに載せられ、インターネットを通して、学生は自分の成績を知ることができる。前述の通り、問題作成及びオーラル、ライティングの採点に教師たちが参加しているため、教師は、KEPT の得点の意味を十分に把握しており、各学生の成績を参考に適切な指導をすることが可能となる。ちなみに、学生の Freshman English course の評価には、KEPT の成績が20％の割合を占める。テストの内容がカリキュラムに連動していることを考えれば、進度テストとしての性格を持つことも当然と言えよう。

また、教師自身の評価者トレーニングとして、毎回テストの前に評価訓練を実施する。ここでは実際の解答例を用いて採点した上で、討議する。たとえば、オーラルテストでは、実際のテスト状況を収録したビデオを見ながら、各自採点した後に、討議し、評価基準の徹底を図る。また、試験実施後には、Multi-facet Rasch 分析という統計手法を用いた採点の結果分析データに基き、ブレの大きい採点官は、個別に特別の訓練を受けることになっている。

5. KEPT 妥当性検証のための研究

次に、これまで KEPT の妥当性検証のために行っている活動について述べたい。以下は、これまで行ってきた活動の一例である[5]。

- テストそのものの信頼性、一貫性（Bonk 2000; Bonk & Ockey 2003他）
- TOEFL® との相関（Bonk 2001）
- Multi-facet 分析でスピーキングとライティングテストの妥当性・信頼性検証
- スピーキング及びライティング評価基準中の語彙力項目と語彙力テストの相関
- グループの構成員の要素がオーラルテストの得点に及ぼす影響について（Bonk & Van Moere, 2004; Van Moere & Kobayashi, 2003; Kobayashi, Johnson & Van Moere, in press）
- 年次報告書の作成（テストの結果全般に加え、学生の進歩、新しいテスト開発の進展状況なども含めたレポート）
- Multi-facet の結果を基に、評価者の訓練を実施
- カリキュラムとの連携の確認（Van Moere & Johnson 2002他）
- 学生の英語力の追跡調査（Leaper & Lee 2004他）
- 学生へのフィードバック
- 授業への還元

たとえば、Multi-facet 分析の結果を見ることにより、タスク間の互換性、評価基準の順当性、評価者間の信頼性などを検証することができる。これにより、評価者間にずれがある場合には、特に「甘い」もしくは「厳しい」評価者を特定することができ、そういった評価者の再訓練が可能となる。また、グループディスカッションという形式を取るオーラルテストにおいて、グループの構成員の英語能力差や性格といった個々の学習者の要因が成績に及ぼす影響を調べることで、グループの構成を考える際の示唆が得られる。KEPT の妥当性検証において特筆すべきことは、学生へのフィードバックと授業への還元という点であろう。これは、学内で開発された、カリキュラムに連動したテストであ

るからこその特徴と言えるだろう。テスト開発に教員が加わっていること及び英語の指導目標と評価の目標が一致しているために、テストの得点を指導に生かしていくことが可能となる。このような環境の中でこそ、Leaper & Lee（2004）のような、学生の英語能力の変化を2年間にわたって追跡し、その要因を探る研究も可能となり、その研究結果が大きな意味を持つこととなる。

ここで、これらの妥当性追及のさまざまな活動を、先に挙げたMessickの枠組みに照らして考えてみよう。多くの活動が、まずは基本となる構成概念的妥当性を追及しようとしていることは容易に見てとれよう。テストの信頼性・一貫性の確認、TOEFL®との相関、Multi-facet分析による評価者及び評価基準の信頼性・妥当性の確認、Multi-facet分析結果による評価者の再訓練、年次報告書の作成等、いずれも、テストそのものが妥当であることを確認するために欠かせない活動である。

しかし、ここで注目に値するのは、上に挙げた妥当性追及の活動の多くが、単に、テストそのものの妥当性・信頼性を追及するのみに留まらず、使用の適切さ、得点の解釈の適切さ、波及効果の確認も行っていることである。学生の英語力の追跡調査や学生へのフィードバック、授業への還元等、いずれも、KEPT本来の目的に照らして、テスト結果を解釈し、使用していることを確認するための活動である。神田外語大学で取り組んでいるKEPTというテストの妥当性追及が、Messickの提唱する結果的妥当性を奇しくも実現しようとしていることが伺えるのではないだろうか。

6. 考察とまとめ

指導と評価は切り離すことのできないものである。教育における評価の大きな目的は、学習者が教育目標に到達できたかどうか、学習の効果はあったのか、次の段階に進めてよいのか、教室での指導がカリキュラムの目標通り進んだかどうか、教育プログラム全体が目標通りに運営されているかどうか、といったことを確認するために、必要な情報を集めることである。そのためには、必要な情報とは何かを見定め、適切な判断の根拠を収集する必要がある。Tierney, Carter & Desai（1991: 35-37）は、「評価は、教師と学習者が自分自身及び相手を理解する助けとなる」はずで「教師と学習者は、集めた情報を、洞察力を

もって分析できる力をつけていくべきである」と述べている。さらに、Rea-Dickins（2000：397）は、「教育における評価とは、教室での指導を導き、学習と学習者を育むための豊富な情報を教師に与えてくれる道具である」と評価の意義を唱えている。もちろん、評価の目的の中には、特定の教育機関の特定のプログラムの教育目標を越え、言語能力の習熟度全体のスケールに照らして学習者がどの位置にいるのかを知ることを目的とするような総括的評価もあり、教室における指導への直接の還元を必要としない場合もある。しかし、総括的評価のみでは、教師が必要とする的確な情報を得ることはできない。学習者の能力、進歩、指導、教育プログラムといった広範囲にわたる現象に関する情報を、ひとつの手段で収集することは難しく、すべてに対する答えを、KEPTのようなテスト一つで、解決できるわけではない。ただ、カリキュラムや学生に合ったテストを作成することで、教育目標を再確認する機会ともなる上、KEPTの妥当性検証の試みで見たように、教育効果を検証するための情報を収集する様々な手段を模索することも可能となる。また、教師自身が、自分の指導の効果を確認するための研究をする機会が得られ、職業的な向上のきっかけにもなるであろう。測定したい能力を正確に測るテストを開発するのみに終わらず、そのテストの適切な使用を求め、結果の解釈を適切に行い、指導に還元していく、それこそがMessickの説く結果的妥当性を追求することになるのではなかろうか。

以上、本稿では、テストが及ぼし得る波及効果や社会的影響を鑑み、テストの使用や得点の解釈が妥当であることを示すことを妥当性検証の要件としたMessickの結果的妥当性の概念に基づき、現在日本の高等教育機関で広く行われている外部テストの利用について考察し、神田外語大学で取り組んでいるテスト開発・妥当性検証の試みを紹介した。テストの得点は、我々が社会生活を営むにあたり、さまざまな形・用途で利用される。教育機関における学習効果の判断だけではなく、入学選抜、資格や学位の授与、昇進・配属の決定など、人生の岐路を決定する大きな判断材料ともなりうる。そのため、テストを作成する者もその得点を利用する者も、テストの及ぼす影響力をしっかりと掌握し、その社会的責任を認識する必要がある。そして、より望ましい、妥当性のあるテストを作成し、その得点を当初の目的に沿って、正しく利用する責務がある。

テストの得点を解釈する際には、そのテストがどのような目的で開発されたのか、どのような能力を測定するために、どのような尺度を用いて測定しているのか、得点の解釈及びそれに基づく価値判断がどのような社会的影響を及ぼすのか、といったことに十分留意しなければならない。そして、テストを本来の目的とは異なる目的に使用することは、慎むべきであることを力説したい。神田外語大学における取り組みが、日本の多くの大学や教育機関への参考となることを願ってやまない。

[謝辞]

本稿をまとめるに当たり、Bonk （2004）が非常に大きな示唆を与えてくれた。ここに感謝の意を表する。

[注]

[1] 妥当性の概念の時代的変遷及び Messick の概念については、Cumming（1996）、Chapelle（1999）に詳しい。
[2] http://www.toeic.or.jp/toeic/data/pdf/toeic_2004-2.pdf より。
[3] http://www.toeic.or.jp/toeic/about/index.html
[4] 評価基準については、KEPT の HP（http://kandaeli.brinkster.net/kept/SignIn.aspx）を参照のこと。
[5] 詳細は上記 KEPT の HP を参照のこと。

[参考文献]

Alderson, J.C. and Banerjee, J. (2001). Impact and washback research in language testing. In C. Elder, A. Brown, E.Grove, K.Hill, N.Iwashita, T.Lumley, K.McLoughlin, & T. McNamara (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies.* (pp.150-161). Cambridge: Cambridge University Press.

Alderson, J.C. & Banerjee, J. (2002). Language testing and assessment (Parts 1 & 2). *Language Teaching, 34,* 213-236 & *35,* 79-113.

Alderson, J.C., Clapham, C., and Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bonk, W. J. (2000). KEPT Argentina-2000 administration: report and analysis. *Studies in*

linguistics and language education (Research Institute of Language Studies and Language Education, Kanda University of International Studies), *11,* 163-224.

Bonk, W. J. (2001). Predicting paper-and-pencil TOEFL scores from KEPT data. *Studies in linguistics and language education* (Research Institute of Language Studies and Language Education, Kanda University of International Studies), *12,* 65-86.

Bonk, W.J. (2004). TOEIC and KEPT: A critical comparison. Paper presented at a lunch-time seminar. Kanda University of International Studies. June, 2004.

Bonk, W. J. & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20,* 89-110.

Bonk, W. J. & Van Moere, A. (2004). L2 group oral testing: The influence of shyness/ outgoingness, match of interlocutors' proficiency level, and gender on individual scores. Paper presented at the 26th Annual LTRC. Temecula, California.

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

Chapelle, C.A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20,* 409-439.

Chapman, M. (2005a). A case study of the need for change in the language testing policies of a Japanese corporation. *JLTA Journal* (The Japan Language Testing Association), *8,* 51-67.

Chapman, M. (2005b). TOEIC: Claim and counter-claim. *The Language Teacher, 29,* 11-15.

The Chauncey Group International Ltd. (2000). *TOEIC Report on Test-Takers Worldwide 1997-98.* Princeton, NJ: The Chauncey Group International Ltd.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study* Cambridge: Cambridge University Press.

Cheng, L. & Watanabe, Y. with Curtis, A. (2004). *Washback in Language Testing: Research Contexts and Methods.* New Jersey: Lawrence Erlbaum.

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In Brown and Yamashita (Eds.), *Language Testing in Japan.* (pp. 66-75). Tokyo: The Japan Association for Language Teaching.

Cumming, A. (1996). Introduction: The Concept of Validation in Language Testing. In Cumming, A. & Berwick, R. (eds.) (1996), *Validation in Language Testing.* (pp.1-14). Clevedon, Avon: Multilingual Matters.

Cumming, A. & Berwick, R. (eds.) (1996), *Validation in Language Testing.* Clevedon, Avon: Multilingual Matters.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14,* 295-303.

Heaton, J.B. (1988). *Writing English Language Tests. Second Edition.* Harlow: Longman.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation Newsletter, 6,* 2-8.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Kobayashi, M., Johnson, K. & Van Moere, A. (in press). Effects of quantity and quality of students' output in group oral tests. *Studies in Linguistics and Language Teaching* (Research Institute of Language Studies and Language Education, Kanda University of International Studies), *16.* 275-295.

Leaper, D.A. & Lee, P.T. (2004). The KEPT score tracking project. *Studies in Linguistics and Language Teaching.* (Research Institute of Language Studies and Language Education, Kanda University of International Studies), *15,* 131-148.

Luoma, S. (2001). What does your test measure?: Construct definition in language test development and validation. Unpublished Ph.D. dissertation, University of Jyvaskyla.
Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.
Messick , S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256.
Rea-Dickens, P. (2000). Classroom Assessment. In Hedge (2000). *Teaching and Learning in the Language Classroom.* (pp.375-401). Oxford: Oxford University Press.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing,* 14, 340-349.
Tierney, R.J., Carter, M.A., & Desai, C.E. (1991). *Portfolio Assessment in the Reading-Writing Classroom.* Norwood, Mass.: Christopher-Gordon Publishers.
TOEIC® 運営委員会 (2004).『TOEIC® テスト 入学試験・単位認定における活用状況 —大学院・大学・短期大学・高等専門学校』東京：(財) 国際ビジネスコミュニケーション協会
Van Moere, A. & Johnson, F. C. (2002). Communicative Assessment in a Personal Curriculum at Kanda University of International Studies. *JALT 2002 Pan-SIG Conference Proceedings.*
Van Moere, A. & Kobayashi, M. (2003). Who speaks most in this group? Does that matter? Paper presented at the 25th LTRC, Reading, UK.
Wall, D. (2005). *The Impact of High-stakes Testing on Classroom Teaching: A case study using insights from testing and innovation theory.* Cambridge: Cambridge University Press.