コミュニケーション学科論集
茨城大学人文学部紀要　No.13　2003.3

英語教育改革プロジェクトにおける プレイスメントテストに関する考察

土　平　泰　子，熊　澤　孝　昭

1．はじめに

1.1　習熟度別クラス導入の過程と問題点

各学生のレベルが様々に異なることは効率よく学習を進めるためには望ましくない。教師が，指導内容，教材，指導法を決める際にも，また学生にフィードバックを与える際にも，同クラス内の学生のレベルがある程度設定されていれば，よりそのレベルに焦点を絞った指導が可能である。

その利点も含めて現在様々な大学で習熟度別クラス編成を導入，または導入への検討を行っている。茨城大学も例外ではなく，2002年度入学の人文学部社会学科の学生を対象に教養教育プログラム内の英語科目でプレイスメントテストを実施し，4月より試験的に習熟度別クラス編成を行った。本稿では，まず始めに様々なテスト理論の中から本稿に関わるものを概観し，次にプレイスメントテストをどのように選び，どのように実施，分析するかを検討する。

1.2　様々な測定と理論

1.2.1　集団基準準拠テストと目標基準準拠テスト

テストの種類は相対評価と絶対評価に分かれ，言語テストの用語では，それぞれ集団基準準拠テスト（norm-referenced test）と目標基準準拠テスト（criterion-referenced test）と呼ばれている（Bachman，1990，1996；Brown，1990，1995，1996；Brown & Hudson，2002；Henning，1987；大友，1996）。集団基準準拠テストとは受験者の得点結果を他の受験者の結果と比較し，相対的に計算する手法である。統計学的には，受験者全体の試験結果を正規分布図で表し，その母集団の平均値からの距離，標準偏差をｚ-得点（素点－平均／標準偏差），またはＴ得点（10×ｚ-得点＋50）に変換し，得点を導きだす手法等がある。集団基準準拠テストにはTOEFLやTOEICなどの資格試験があてはまる（土平，1996）。集団基準準拠テストの特徴としては以下のような内容が挙げられる（Brown，1996）。

1）言語運用能力を測定する。

2）受験者の得点を広範囲に弁別する。

3）テスト内容が事前に公表されていない。

4）項目困難度が広範囲にわたる項目。

5）出題テスト項目が比較的多い。

目標基準準拠テストとは特定の学習目的を受験者がどれだけ習熟したかを測定するものであり，集団基準準拠テストと比較して，より特定の目的を測定する項目から構成されている。また，相対

的に得点を解釈するのではなく，絶対的に個々の受験者がどれだけ特定の学習目的，または評価基準を習熟できたかをパーセンテージで計算し得点を算出する。目標基準準拠テストには分割点(cut-off score）が設置される場合が多く，受験者の得点がその分割点以上か，否かで合否を決定するようなテスト解釈にも用いられる。また，目標基準準拠テストには英語検定試験（STEP Test）や国際英語検定試験（G-TELP）などの検定試験があてはまる（土平，1996)。目標基準準拠テストの特徴としては以下のような内容が挙げられる（Brown，1996)。

1）受験者の学習目的習熟度を測定する。

2）受験者の得点を弁別するのではなく，受験者の習熟度を測定する。

3）テスト内容が事前に公表され，受験者がすでに学習した内容である。

4）テスト項目の難易度が比較的一定で，評価基準を基にテストが構成されている。

5）出題テスト項目が比較的少ない。

1.2.2 項目応答理論

項目応答理論または項目反応理論（item response theory）とは，受験者がある項目困難度を示す項目に応答できるかどうかをもとに，受験者の能力などの特性を推測する理論である（Henning，1987;大友，1996)。項目応答理論を用いる利点としては以下の3つが挙げられる（Henning，1987;大友，1996；土平，1996)。

1．受験者集団に依存せずにテスト項目統計量を求められる（sample-free item calibration)。
受験者が異なる場合でも項目の特性を分析した結果は変わらない。よって，受験者の能力にかかわらず，項目困難度の値は不変である。

2．テスト項目の特性に依存しない受験者能力の測定ができる（test-free person measurement)。
テストの特性が異なる場合でも受験者の能力を分析した結果は不変の値が求められる。よって，異なる項目困難度から構成されているテストを実施しても，受験者の能力値は不変である。

3．測定の精度を受験者ごとに求められる（multiple reliability estimation)。
受験者ごとに生じたテストの測定誤差を値として求めることができる。

さらに，豊田（2002）は項目応答理論の利点を5つ指摘している。

1．複数のテスト間の結果を比較することが容易である。これは先に述べた，テスト項目の特性に依存しない受験者能力の測定を示唆していること。

2．測定精度をきめ細かく確認できること。テストの特性は受験者の特性に適しているかを示すテスト情報関数（test information function）からテストの精度を明示することができる。

3．平均点をテスト実施前に制御できること。すでに試験的に実施し，分析済みの項目を仕立て式項目（tailored items）といい，その項目に関しては項目困難度などが分析結果として明示されるので，その項目困難度をもとにテストを構成すれば，およその平均点などを実施前に制御することができる。また,仕立て式項目の集まりを項目プール(item pool)または項目銀行(item

bank）と呼ぶ。

4．テスト得点の対応表が作成できること。これは2つの難易度の異なったテストを同じ尺度上で得点を比較することができる。また，その計算過程をテストの等化（test equating）と呼ぶ。

5．受験者ごとに最適な問題を瞬時に選び，その場で出題できること。技術の発達とともに，コンピューター上でテストを実施するコンピューター式テスト（computer-based test）が可能になった。さらに，項目応答理論とコンピューター技術を駆使して受験者の特性に的確な項目を出題する方式で受験者の能力を測定する適応型テスト（computer adaptive test）の開発が近年進んでいる（大友，1996；Sands，Waters，＆McBride，1997）。TOEFLの例を挙げると，リーディング問題は同じ問題と問題数を全受験者にコンピューター上で出題するコンピューター式テストになっており，聴解問題は受験者の特性に合わせて項目と項目数を出題する適応型テストになっている。Brown（1997）では，コンピューター式テストと適応型テストの利点と難点を挙げるなど文献研究を行い，今後の研究課題について言及している。

項目応答理論にはいくつかのモデルがある。項目応答理論の考案者George Rasch開発の二値型データ（dichotomous data）を分析する1パラメータロジスティクモデル（one-parameter logistic model）またの名称をラッシュモデル（Rasch model），2パラメータロジスティクモデル（two-parameter logistic model），3パラメータロジスティクモデル（three-parameter logistic model）のほかに1パラメータロジスティクモデルを応用し多値型データ（polytomous data）を分析する部分採点モデル(partial credit model）とメニーファセットラッシュモデル（many-faceted Rasch model）などが代表的なモデルである（Brown，2002；Hambleton，Swaminathan，＆Rogers，1991；Henning，1987；McNamara，1996，大友，1996）。ラッシュモデルは，多岐選択肢問題などの正解，誤答で断定できる二値型データを分析し，項目困難度と受験者の能力値を推測することができる。2パラメータロジスティクモデルでは項目困難度，受験者の能力値とある項目がどれだけ受験者の能力を弁別するかを数値で表す項目弁別力パラメータ（item discrimination parameter）を測定する。さらに，3パラメータロジスティクモデルでは項目困難度，受験者の能力値，項目弁別力の他に，推量パラメータ(guessing parameter）を測定することができる。ラッシュモデルの開発が進み，二値型データだけではなく，1，2，3，4，5のような部分点を分析する部分採点モデル（partial credit model）がある。このモデルを応用し，リカートスケール（Likert scale）をラッシュモデルで分析するようになり，心理学の分野でも一般的に使用されるようになった（Embretson & Reise，2000）。さらに，言語テストの分野では話す能力や書く能力を測定するパフォーマンテスト（performance test）が盛んになったが，受験者の能力値，項目困難度と審査官が厳しい得点をつける特性（severity）があるかを分析するメニーファセットラッシュモデルの使用が一般的になった（McNamara，1996）。

1.3 カリキュラムとテスト選択

もう一つ言及しておきたいのは，テスト選択とカリキュラムの関係である。一連のカリキュラムの過程では，テストは主に２つの目的で実施される。１つは，プログラムレベルでの使用目的（administrative purpose）で，もう１つはクラスレベルでの使用目的（classroom purpose）とがあり，それぞれ結果の使用法も異なる。集団基準準拠テストを使用した場合，プログラムレベルでは２種類の判断が行われる。１つは能力テスト（proficiency test）結果を用いて，その受験者の結果を母集団から相対的に比較し，どの程度の運用能力を保持しているかを知った上で，受験者がそのカリキュラムのレベルに適合するかの判断に用いるもので，入学試験にあたる。もう一つはプレイスメントテスト（placement test）であり，能力テストと同じく，テスト結果を基に受験者がどのレベルの講座に適合するかの判断をくだすのに用いる。但し，一度入学試験時に能力テストで受験者のレベルがスクリーニングされているので，プレイスメントテストはプログラム内の受験者のレベルに合った難易度から構成された，弁別力の高い項目で構成されていなければならない（Brown，1995，p.111）。

目標基準準拠テストは，クラスレベルでは２種類の判断に役立てることができる。１つは到達度テスト（achievement test）で，どれだけ学習者が学習目的を習熟したかを測定する。また学習者の得点が分割点に到達したかで合否を判断する。もう１つ診断テスト（diagnostic test）では，どれだけ学習者が学習目的を熟知しているかを測定することにより，学習者の苦手な学習内容と，得意としている学習内容を診断する。また，目標基準準拠テストの結果は教師にとっても非常に重要な情報となる。学習者が習熟していない内容や，学習者が不得意としている内容はより授業時間をかけ効果的に教える必要がある。但し，到達度テストと診断テストを実施するには，授業の目標と目的が明確になっていることと，その目的を効果的に測定するテスト項目を出題することが必須である（Brown，1995）。

1.4 プレイスメントテスト

1.4.1 先行研究

先に述べたように，プレイスメントテストに適しているテストは集団基準準拠テストであり，カリキュラム内の学生のレベルを弁別し，各クラスが同等のレベルの履修者で構成されるようにクラスを編成するのに用いる。日本の大学は一般的にTOEFLやTOEICなどの外部作成の能力テストが使われている。Bachman（1990）は，以下の３条件を満たす場合のみ能力テストをプレイスメントテストとして使用することを指示している。

１．学生のレベルや背景が多様な場合

２．カリキュラム内でクラス目標と目的が設定されていない場合

３．学生のレベルが毎年異なる場合

しかし，Culligan and Gorsuch（1999）によれば，この３点は日本の大学のカリキュラムには当て

はまらない場合が多い。彼らはこの研究で，Educational Testing Service（ETS）開発の能力テストSecond Level English Proficiency Test（SLEP）を，プレイスメントテストとして日本の大学生539名に実施した結果から，その適合性を検証した。

まず，彼らはSLEPの各項目の正答者数を受験者数で割った数値である項目困難度（item facility）と，正解率の高い上位27%の集団の項目困難度から正解度の低い下位27%の集団の項目困難度を引いた数値である項目弁別力（item discrimination）の結果を分析した。項目困難度とは集団基準準拠テストの分析法で，古典的テスト分析の一つだが，項目の正解率を0から1に数値化したものである。大友（1996）によれば，最適項目困難度は，0.5+0.5（1／選択肢数）で算出することが出来，その項目困難度はその周辺であることが望ましい（p.31）。選択肢が4つの場合は0.65，3つの場合は0.667となる。項目弁別力とは項目の弁別力を-1～1の範囲で数値化したものであり，.2～.29は弁別力が低く項目の修正が必要とされている（Brown，1989，1996）。SLEPを受験者に実施し，分析した結果，全150項目の内，66項目が項目弁別力.2以上だと報告している。つまりこのプレイスメントテストでは，最も重要視されるべき項目弁別度の数値が低い項目から構成されていて，その受験者のレベルを広範囲に弁別しない傾向があったことを示唆している。

また，彼らはSLEPを実施した時のテスト信頼性を全項目と不良項目を除いた場合とで検証した。その結果，信頼性係数は全150項目での内部一貫性信頼性キューダー・リチャードソン20の公式（internal-consistency reliability Kuder-Richardson formula 20（KR20））では.81で，項目弁別度値.2以上の66項目では比較的項目数が少ないが，KR20は.84であり，後者のほうがより信頼性が高いことが分かった。

最後にこの結果を，先に述べたBachman（1990）の3つの条件について彼らは以下のように結論付けた。

1．当てはまらない。日本の大学生は，高校での学習内容や教材が類比しており，大学内の学生レベルも類似している場合が多いので，能力テストでは学生のレベルを広範囲に弁別する項目は少なくなるのではないか。能力試験をプレイスメントテストとして使用する場合は，受験者に実施後，項目困難度と項目弁別力を分析し，より弁別力の高い項目のみクラス編成時の判断材料として使用するべきである。

2．当てはまらない。既存のカリキュラムには明確な目標がある。しかもクラスは全体的にスピーキングを重視しているが，SLEPはその技能を直接測定しておらず，テストの妥当性が疑問である。クラスの目的をより効果的に測定するテストを実施する必要があり，そのためには複数のテストの実施が必要である。

3．当てはまらない。毎年ほぼ同等レベルの学生が入学して来る。

このようにBachman（1990）で提示された3点はどれも当てはまらず，彼らのカリキュラムに能力試験SLEPの結果をクラス編成に使用するのは不適切だという結論となった。よって，より精度の高いクラス編成を行うためには，時間と労力を費やすが，信頼性があり，そのカリキュラムに妥当

な独自テストの開発が必要であると述べている。また，Culligan and Gorsuch（2000）では，彼らは同じデータを用いて項目応答理論のモデルの一つであるラッシュモデルで分析を試み，ラッシュモデルによって分析した能力推定値の方が，SLEPテストの素点の結果よりも精密なクラス編成が可能となることを示した。

ジョンソン，玉井，加須屋（1999）はSLEPを140名の大学1年生に実施し，その信頼性，項目困難度と項目弁別力の3点において検証した。信頼性については，SLEP全体のKR20は.82であったが，聴解力理解問題は全項目数75問中（k＝25，18，12，14，6），一部の信頼性が際立って低いことを示した（KR20＝.39，.77，.25，.18，.35）。また，項目困難度の値の.30～.70，項目弁別度と点双列相関係数(point biserial correlation coefficient）の値が.30以上を基準として古典的項目分析を試み，リスニングでは「よい項目」は75問中44問，リーディングでは75問中39問が弁別力の高い項目であったこと示した。

Kondo-Brown and Brown（2000）はハワイ大学日本語学部に入学する学生を対象にプレイスメントを目的として，リスニングテストを888名に，文法テストを1284名に，漢字と仮名の認識テストを1120名に，ライティングテストを234名に実施した。そして，その結果をリスニングテスト，文法テスト，認識テストについて，項目応答理論を用いて項目分析を行った。分析手法は3パラメータロジスティクモデルで，項目応答理論のソフトウェアであるXCalibre version 1.10e（Assessment Systems，1997）を使用し，3つのテストの内部一貫性信頼性キューダー・リチャードソン21の公式(internal-consistency reliability Kuder-Richardson formula 21（KR21)），残差（residual），項目困難度パラメータ，項目弁別力パラメータ，当て推量パラメータを算出した。結果，リスニングテストは.797（k＝14），文法テストは.955（k＝70），認識テストは.951（k＝50）という信頼性のあるテストであることが明らかになった。

さらに彼らは，分析結果を基に項目の良否を検証し，不良項目をテストから除外していった。データから判断基準に該当する項目を削除し，再度項目分析を行った結果，リスニングテストは残差が.2以上の数値を示す項目が1個あり，その項目を削除した結果，信頼性が.801（k＝13）となった。文法テストは残差が.2以上の数値を示す項目が6個あり，分析の対象外になった。さらに，テストの項目数を減らし効率性を高めるため，弁別力パラメータの数値が1.00以下の項目と，項目困難度パラメータが1.00以上の項目を削除し，再度分析を行った結果，信頼性が.948（k＝50）となった。認識テスト（recognition test）は50項目中20項目が2.0以上の残差を示し，削除の対象となり，再度信頼性を算出した結果.922（k＝30）となった。このように彼らは項目応答理論を用いて，項目数を減らしながらも高い信頼性を保つことができることを示し，より効率的なテスト開発を提案するに至った。以下の条件が彼らの用いた基準である。

1）残差（residual）の数値が2.00以上の数値を示す項目（Flgの欄にRが表示される）

2）項目困難度パラメータ（bパラメータ）が－2.95～＋2.95の範囲を超えた数値を示す項目（Flgの欄にPが表示される）

3）項目弁別力パラメータ（aパラメータ）が.3以下の数値を示す項目，及び当て推量パラメータ（cパラメータ）が.4以上の数値を示す項目

以上，古典的テスト理論，項目応答理論共にプレイスメントテストのためのテスト作成を中心に見てきた。先行研究の結果をまとめると，クラス編成，プレイスメントテスト実施，分析において注意すべきことは以下の2点である。

1）受験者の能力を広範囲に弁別する項目の多いテストを選び，以下のような項目はクラス編成に用いない。

a．古典的項目分析で項目弁別度の低い項目。

b．項目応答理論で項目分析を行い，残差（residual）の数値が2.00以上の項目，項目困難度パラメータ（bパラメータ）が−2.95〜+2.95の範囲を超えた数値を示す項目，項目弁別力パラメータの数値が.3以下の項目，当て推量パラメータ（cパラメータ）が.4以上の数値を示す項目。

2）受験者の得点を素点ではなく，z-得点かT得点に変換した数値をクラス編成の判断材料とする。さらに，より正確なクラス編成を試みるのであれば，Culligan and Gorsuch（2000）が提唱するように，項目応答理論のモデルを分析に使い，受験者の能力値をクラス編成の判断材料とする。

これ以外にも,項目応答理論のラッシュモデル，2パラメータロジスティクモデル，3パラメータロジスティクモデルのどれかを用いる場合は，項目適合（item fit）を分析する方法もある。.8〜1.2の範囲外の数値を示す項目はモデルとの不適合（misfit）と解釈し，テストから除外するのが一般的で，クラス編成の判断材料としては扱わない（Bond & Fox，2001，p.179）。

1.5　本稿のねらい

本稿の目的は先の2点を踏まえ，以下の3つの問いについて議論し，今回のプレイスメントを検証することである。以下，下記の3つの疑問に基づいて議論をすすめたい。

a．実施したテストは信頼性，弁別性，妥当性のあるものであったか？

b．今後，どのようなクラス編成を行えは良いか？

c．今後，どのようなプレイスメントテストを実施すれば良いか？

今回の分析においては，コンピューター適応型テスト，センター試験の正解不正解数のデータが得られないという制限もあるが，以下の点に絞ってできる限りで行ってみたい。

1）今回のプレイスメントに用いたテスト得点の分析

2）古典的テスト理論を用いた項目弁別力の分析

3）項目応答理論を用いた項目困難度，項目弁別力，残差の分析

4）z-得点を用いた再クラス編成のシミュレーション

2．方　法

2.1　参加者

平成14年度から茨城大学では，4技能育成とした英語教育の実験プロジェクトを開始した。参加者はこのプロジェクト授業を履修している人文学部社会学科（人社）1年生，230名である。各試験で欠席者，欠損データ等生じたところは，それぞれに整理し分析を行った。英語力については，参加者の殆どが日本の大学1年生の平均的なところであるといってよいだろう。

2.2　手順

前期開始第2回目の授業内で人社の学生230名にリスニングテストを実施した。コンピューター適応型テストであるQuick Placement Test（QPT）は，コンピューター30台ほどを用いて，3日間に分けて実施した。さらに，希望者のみ，翌週第2回目のQPT受験を促した。

集計結果後，平成14年度大学入試センター試験英語の得点とリスニングテストの素点を2倍にした得点を合計したものをクラス編成の判断材料とした。そして，得点順に中級（Intermediate）1クラス，準中級（Pre-intermediate）4クラス，初級（Beginner）2クラス，基礎（Basic）1クラスの4レベル，8クラスに分けた。

2.3　使用したテスト

今回のクラス編成には，平成14年度実施英語センター試験の結果，平成14年度4月に人文学部社会学科1年生対象に行ったQuick Placement Test（Oxford，2001）と全50問からなる多肢選択式のリスニングテストの結果を考慮した。以下に3つの試験の特徴を述べたい。

2.3.1　Quick Placement Test

University of Cambridge Local Examinations Syndicate（2001）が開発したQuick Placement Test (QPT) は集団基準準拠テストでプレイスメントテストである。筆記版と適応型テスト版があるが，今回は適応型テスト版を使用した。

QPTは全問多岐選択型で，各項目4つの選択肢中，1つが正解で，残りの3つは錯乱肢からなる。リスニング問題，リーディング問題と，文法能力と語彙知識を測定する語法問題の3部構成である。リスニング問題ではイギリス英語で発音された刺激（stimulus）を2回まで聞くことができ，4つの選択肢の中から正解を選ぶ内容になっている。リーディング問題では，刺激を読み質問の内容にあてはまる正解を選ぶ形式になっている。語法問題は文章中の空欄箇所にあてはまる単語を選択肢から選ぶクローズテスト形式である。

QPTは適応型テストなので出題項目と出題項目数は受験者の特性により異なり，テスト所要時間も異なるが，20から25項目で受験者の応答から能力値を測定する仕組みになっており，所要時間はおよそ20分程度である。最高得点は100点に設定され，受験者の得点は6つのレベルに分割される。

QPT得点の0から39がレベル0 (Beginner)，40から49がレベル1 (Elementary)，50から59がレベル2 (Lower Intermediate)，60から69がレベル3 (Upper Intermediate)，70から79がレベル4 (Advanced)，80から100がレベル5 (Very Advanced) にそれぞれ相当する。また，QPTでの得点レベルはAssociation of Language Testers in Europe (ALTE) 開発のALTE Level Descriptionと比較できることからテストの妥当性があることがマニュアルに示されている。

今回QPTを採用した理由としては，実用性の観点からが大きい。コストが低い，コンピューターの自動採点で受験者の得点がその場で分かること,そして一人20分という実施時間の短さ等である。また,大学入試センター試験では実施されていないリスニング能力を測定できることも挙げられる。しかし，その一方でコンピューターへのインストールやシステムの管理にかなりの労力と技術が必要になることが，今回の実施を通じて分かった。

2.3.2　リスニングテスト

今回実施されたリスニングテストは50問から成る多肢選択型のテストで，学習内容の習熟度を測定する目的で作られた到達度テストである。選択肢は3つで，正解は1つで，残りの2つが錯乱肢である。試験所要時間はおよそ30分で，1問ごとに北米英語の発音で録音された刺激が2回続けて放送される。

試験内容は5つのパートに分かれており，第1部は5～10語から成る文を聞き，最も絵の内容を表していると思われる選択肢を選ぶ問題で，8問出題される。第2部は10語から成る質問文を聞き，的確な応対を選択肢から選ぶ問題で，10問出題される。第3部は約10文から構成されている2人の人物による会話を聞き，内容に相応した選択肢を選ぶ問題で，10問出題される。第4部は1文から5文で構成された文を聞き，内容に相応した選択肢を選ぶ問題で，10問出題される。第5部はやや長めの文章を聞き，合計12問の質問に答える形式である。内容は10文から構成されている学術的な談義になっている。このリスニングテストは当初予定されていたものではなく，QPTのシステム作動に問題があったため，急遽受験者のリスニング能力を測定する手段として行われた。

2.3.3　大学入試センター試験

大学入試センター (2001) が開発した大学入試センター試験英語 (以下センター試験) は項目総数51問で，6つのパートから成る。全て4肢選択肢項目で，所要時間は80分である。第一問は8項目から成り，語彙中や文中のどの箇所にストレスを置くかを答える，発音の強勢に関する問題である。これについては，発音を“文字の試験で測るという点で基本的な問題を抱えている”(大学入試センター，2001，p.582) という分析もあり，改善が望まれている。第2問は17項目あり，会話表現や語彙問題など文法的能力や社会言語的能力を問う問題である。第3問は5項目から成り，接続詞などの文脈理解の力を問う問題である。第4問は5項目から成り，読解能力よりスキャニングやスキミングなどのストラテジーを測っているように思われる。第5問は5項目出題され，会話文

を理解し，適切な表現を選ぶリーディング問題である。第6問は8項目出題され，物語性のある文章を読み，理解力を問うリーディング問題である。

センター試験は，以下の理由から基本的に目標基準準拠テストの特性を持っていると考えられる。

1）実施目的が高等学校学習指導要領の基礎的学習の習熟度の判別である。

2）試験範囲が事前に公表され，試験内容が受験者にとって学習済みである。

3）実施前に項目ごとの配点が決まっており，200点満点中何点正解したかで得点が算出される。

4）全試験項目が51問と比較的問題数が少ない。

しかし，その一方でセンター試験には集団基準準拠テストの要素も持つと思われる。高等教育3年間の学習内容の習熟度という広い試験範囲を持ち，石塚，中敏，内田，前（2001）では，2001年1月249,256人に実施したセンター試験の結果から,項目難易度も点双列相関係数も広範囲な数値を示している（IF=.17～.97，r_{pbi}=.023～.495)。

そして，コミュニケーション能力測定テストの要素も指摘されている。近年は学校の英語教育でもコミュニケーション能力の育成が重視され，センター試験にもコミュニケーション能力を測定する能力試験的な要素があるのではないかという意見もある（大学入試センター，2001，p.571)。Brown and Yamashita（1995）はセンター試験や他大学の入試試験の問題を，readabilityと項目形式の観点からテスト分析を調査している。さらに，Watanabe（1996）は，センター試験が教授法や授業内容などに与える影響を,波及効果(washback effect)の観点から言及している。大学審議会(2000)では，今後の更なるセンター試験の改善点が検討されており，リスニングテストの導入，項目銀行の開発，テストの等化などが挙げられている。

3．結果と分析

3.1 使用したテストの得点分析

3.1.1 リスニングテスト

先ず，以下にリスニングテストの記述統計と分布を示す。本稿での分析はすべてSPSS ver.11.0 (SPSS，2002）を用いている。

Table 1. Descriptive statistics for the listening test

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Listening test	228	29.00	19.00	48	37.75	4.44

Skewness	Std. Error	Kurtosis	Std. Error
−.52	.16	1.11	.32

Figure 1. Distribution of the listening test scores

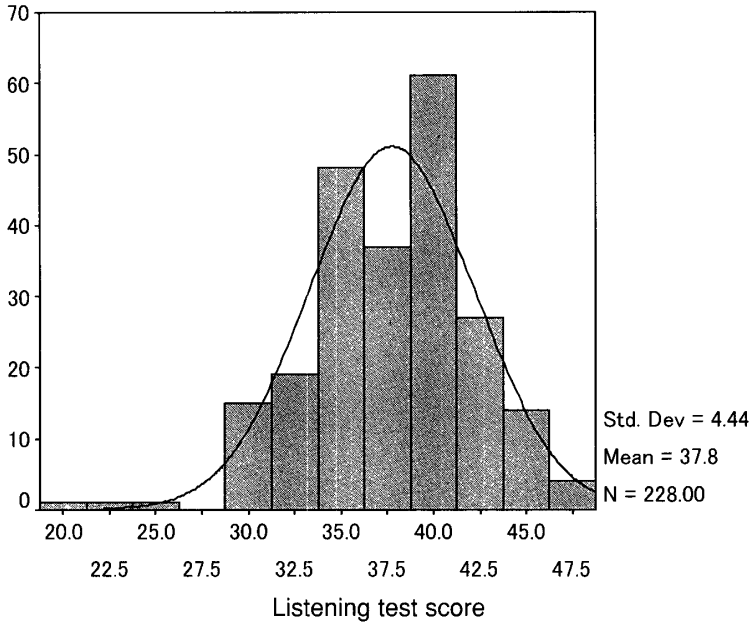


Table 1を見てみると，歪度の値が−2.04（Std. error=0.17），尖度の値が12.46（Std. error = 0.34）となっている。歪度の値が負であることは分布のピークがやや得点の高い方にあることを示し，また大きな尖度の値は分布が一部分に非常に集中していることを表わす。Brown（1997）によれば，歪度，尖度の値が標準誤差の値の2倍を越えるかどうかで標準性の判断ができる。この場合，共に大きく2倍を上回っているので標準分布とはかけ離れていると判断される。さらに尖度については100，歪度については200を越える大きなサンプルについてはこの方法が当てはまらないこともある。しかしそのような場合でも，グラフを用いて視覚的に分布を見，補う必要がある(Tabachnick & Fidell，2001，p.74）。そこでFigure 1を見ると，やはり分布が偏り，そして40点付近に集中しており，標準分布とは異なっていることが結論付けられる。

この分布は，受験者である学生が皆似たような方法と環境で英語を学んできたこと，およそ同じ入試方法で選抜されてきたことなどを考えると当然のことであるが，プレイスメントテストとして活用するためには，より分布を広げる項目を増やし，弁別力を高めることが必要である。

次に用いられたテストの項目分析を行った。Table 2にその結果を示す。

この分析によって，不良項目として削除すべきものが数項目存在することが分かった。先ず，古典的テスト理論に関する分析であるが，項目困難度については大友（1996）の基準を用いると，選択肢が3つのテストの最適項目困難度は.667である。先ずこれを基準とすると全く数値が及ばないものが4項目見られる。次に項目弁別力については大友（1996）やBrown（1996）に述べられている.19以下を不良項目とする基準を用いた。その結果，18項目が不良項目として取り除かれるべきであることが分かった。この50項目中18項目というのは非常に大きな割合であるが，その一つの理由として先ず，このリスニングテストが到達度測定のために作成されたものだということが挙げら

Table 2. Item statistics for the listening test

Item	IF	ID	r_{pbi}	Item Dis.	Item Diff.	Residual
1	.77	.54	.62	.45	−2.00	1.59
2	.77	.54	.62	.50	−1.83	.71
3	.100	.01*	.12*	.68	−3.00*	2.22*
4	.97	.08*	.46	.60	−3.00*	1.29
5	.99	.02*	.34	.68	−3.00*	1.94
6	.99	.03*	.30	.65	−3.00*	1.91
7	.98	.05*	.34	.65	−3.00*	1.91
8	.98	.05*	.45	.68	−3.00*	1.75
9	.87	.28	.41	.52	−2.59	.65
10	.91	.21	.40	.56	−2.89	.52
11	.73	.53	.44	.54	−1.75	1.42
12	.97	.06*	.23*	.58	−3.00*	1.73
13	.88	.27	.45	.53	−2.61	.36
14	.84	.31	.41	.45	−2.51	.79
15	.97	.08*	.28	.64	−3.00*	1.40
16	.93	.17*	.34	.57	−3.00*	.43
17	.96	.10*	.39	.62	−3.00*	1.31
18	.66	.58	.46	.46	−1.14	.92
19	.18*	.26	.38	.47	2.10	1.52
20	.35*	.33	.32	.42	.82	.80
21	.50	.43	.44	.44	−.12	1.32
22	.75	.38	.35	.42	−1.95	1.53
23	.75	.32	.35	.41	−1.91	2.17*
24	.71	.36	.38	.43	−1.51	1.39
25	.91	.14*	.28	.52	−2.98*	.52
26	.57	.43	.44	.41	−.60	1.61
27	.75	.24	.28	.38	−2.04	1.94
28	.40	.40	.39	.46	−.54	.79
29	.85	.14*	.33	.48	−2.55	1.33
30	.91	.13*	.40	.57	−3.00*	.23
31	.89	.23	.47	.59	−2.57	.80
32	.94	.10*	.38	.62	−3.00*	.75
33	.81	.27	.43	.46	−2.28	1.23
34	.42	.45	.48	.38	.49	1.81
35	.82	.18*	.35	.42	−2.50	1.84
36	.21*	.18*	.31	.44	1.90	1.22
37	.73	.35	.35	.41	−1.76	1.54
38	.80	.30	.42	.48	−2.11	.64
39	.73	.44	.44	.44	−1.65	.71
40	.71	.44	.49	.47	−1.44	.82
41	.47	.58	.53	.43	.06	1.05
42	.80	.28	.37	.45	−2.23	1.58
43	.54	.40	.42	.46	−.37	1.08
44	.42	.26	.25	.37	.44	2.28*
45	.81	.29	.38	.48	−2.14	.72
46	.95	.08*	.30	.63	−3.00*	1.02
47	.95	.13*	.42	.65	−3.00*	1.07
48	.78	.24	.34	.49	−1.93	.74
49	.55	.21	.26	.39	−.49	1.85
50	.37*	.26	.35	.39	.83	2.70*

IF：item facility（項目困難度）

ID：item discrimination（項目弁別力）

r_{pbi}：point-biserial correlation（点双列相関係数）

Item Dis.：item discrimination（IRT）（項目応答理論による項目弁別力）

Item Diff.：item difficulty（IRT）（項目応答理論による項目困難度）

Residual（残差）

＊：削除対象項目

れる。これでは使用用途を誤っているため，次回は使用用途に合った項目構成が望まれる。また，もう一つ挙げられるのは，大学入試による選抜で，受験者が元から非常にあるレベルに集中した分布を成しているということである。これは大学のプレイスメントテストの抱える本質的な問題であり，解決のためには受験者のレベルにより焦点を当てた項目作成をする必要があると思われる。

次に，点双列相関係数を用いた弁別力の分析であるが，これはHenning（1987）の，.25以下を不良項目とする基準を採用した（p.53）。その結果，不良項目となったのは2項目となり，先の分析とは非常に異なる結果となった。この点においても，受験者の分布がある範囲に集中していることが理由として挙げられる。Henning（1987）によれば，点双列相関係数の値はピアソンの相関係数と同様，母集団の数と広がりに左右される（p.53）。後に述べる項目応答理論による分析も考え合わせると，この結果は双点列相関係数による分析がこの分布の特殊性に影響をうけたことによるものではないかと思われる。この2項目はクラス分けの判断材料からは削除することが望ましい。

次に，項目応答理論を用いた分析ではKondo-Brown and Brown（2000）の基準を用いた。先ず項目困難度による分析では，彼らの項目困難度を−2.95〜+2.95を許容範囲する基準を用い，結果15項目が該当した。また古典的テスト理論の項目弁別力の分析に似た結果となったのも興味深い。弁別力の分析ではパラメータの数値が.3以下の項目は存在しなかった。当て推量パラメータの値は，今回は2パラメータモデルで分析しているので存在しない。残差では，基準の2.00を上回るものは4項目見つかった。

以上，全ての結果を総合すると50項目中23項目が削除の対象となった。これらの項目は除いて，クラス編成を行うことが適切だといえるだろう。

3.1.2　Quick Placement Test（QPT）の結果分析

先にも述べたようにQPTの得点はネットワーク等の障害があったため，2回にわたって実施され，その得点を統合するなどしてデータとして活用した。記述統計の結果，分布は次頁のようであった。

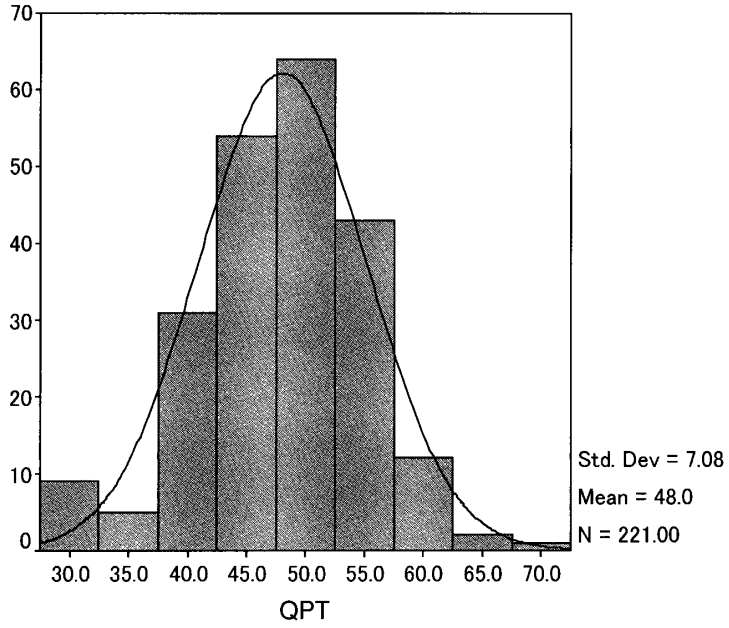
統計，グラフを見ても分かるように，歪度が−0.24（Std. error=0.16），尖度が0.47（Std. error 0.33）と，共に標準誤差の2倍の範囲内に収まっており，この分布があまり偏ったものでないことを示している。

Table 3. Descriptive Statistics for QPT scores

	N	Range	Minimum	Maximum	Mean	Std. Deviation
QPT	221	42	28	70	48.03	7.08

Skewness	Std. Error	Kurtosis	Std. Error
−0.24	0.16	0.47	0.33

Figure 2. Distribution of QPT scores



先述の通り，このテストは2度にわたって実施されたため，その得点の相関を見ることでtest-retest methodまたはparallel forms method（Henning，1987）による信頼性係数を求めた。QPTは適応型テストであるから，各受験者の全問の出来に応じて次に提示する問題が変えられていく。そのため，受験者は1度目と2度目の受験で異なる問題に接している可能性もあり，完全なtest-retest methodではなく，同じ難易度であることから，parallel forms methodと呼ぶほうが適切かもしれない。しかしながら，ここでは，コンピューターによってより適切な項目を選択して提示し，より正確な得点算出がされている。つまり2度のスコアは非常に近くあるべきであるという前提のもとに，test-retest methodの方法で，1度目と2度目に共通した受験者の得点の相関を求めた。相関分析を用いるが，まず前提として1度目のQPT（QPT1）と2度目のQPT（QPT2）のデータがおよそ標準分布をなしている事，極端なoutlierを除いておくことが必須であるため（Tabachnick & Fidell，2001）それを調べた。以下に記述統計を示す。

Table 4. Descriptive Statistics for QPT1 and QPT2

	N	Range	Minimum	Maximum	Mean	Std. Error	Std. Deviation
QPT1	79	37	28	65	46.99	0.76	6.77
QPT2	79	44	19	63	45.85	0.92	8.16

Skewness	Std. Error	Kurtosis	Std. Error
−0.36	0.27	0.92	0.53
−0.38	0.27	0.55	0.53

Table 4 にあるように，QPT1では歪度が−0.36（Std. error=0.27），尖度が0.97（Std. error=0.53），QPT2では歪度が−0.38（Std. error 0.27），尖度が0.55（Std. error 0.53）と，共に標準誤差の2倍の範囲内に収まっており，この分布があまり偏ったものでないことを示している。また散布図も検討したが，特に極端なoutlierもないことから相関分析をこのまま用いて良いことを示している。

Table 5. Correlational analysis of QPT1 and QPT2

		QPT2	QPT1
Pearson Correlation	QPT2	1.00	0.47**
(two-tailed)	QPT1		1.00
		**=p<.01	n=79

QPTを2度受験した受験者の得点の相関分析は，2つのスコアが r =0.47（p <0.00）と高い確率で相関関係を示すということを示した。しかしながら，2つが同じテストである以上これは至極当然であると考えられる。Brown（1996）によれば，統計的な有意（significance）は必ずしも意義（meaningfulness）に繋がるものではない。このmeaningfulnessの判断のために，Brown（1996）は相関係数を2乗した決定係数（coefficient of determination）を提案している。今回の分析では，$r^2=(0.472)^2=0.22$となり，これが2つのスコアの重なり合った部分の分散を示すということになる（p.167）。つまり，ここでは2度のテストが共に測る部分はおよそ22パーセント，残りは測定誤差であるという解釈が成り立つ。

この測定誤差が生じた理由についてはいくつか可能性のあるものが挙げられる。

1）新入生を対象に行ったため，コンピューター操作に不慣れな学生が多くいたこと。

2）コンピューターを通じてリスニングテストを行う際，音声がやや聞き取りにくかったこと。

3）中学高校とアメリカ英語を学んできた学生が大多数であるのに対し，QPTではイギリス英語を採用していること。

4）テスト自身の測定誤差

これらの要因が複雑に絡んで，学習者のパフォーマンスの妨げとなり，測定誤差を生じたと考えられる。特に1），2）についてはコンピューター不安，テスト不安と関連し，それぞれがどのくらいの割合を占めるか興味深いところである。

1度目のQPTスコアと2度目のスコアとの回帰分析も行った。結果，QPT1とQPT2は以下のような回帰式で表わされた。

QPT2＝19.094＋.569×QPT1

ここにQPT1の平均点にあたる46.798を入れると，45.722となり，これが統計的に予測される2度目のQPTスコアであるといえる。さらに，QPT1が60の時，QPT2は53.054となりこのように計算していくと多くの受験者は1度目より2度目の方がスコアをやや落としていることが分かる。勿論測定誤差が大きかったわけであるから仕方がないが，2度目の受験で適応型テストにも慣れているはずであるので，何故なのか興味深い。受験者に話をしてみたところ，「1度目の方が緊張感があった。」「慣れたから。」などの意見が挙がった。コンピューター不安やテスト不安の捉え方も，見直していく必要があるかもしれないと考えさせられた。いずれにしてもこれだけの測定誤差が含まれる試験をプレイスメントに使用しつづけることは困難であろう。

3.1.3 各試験得点の関係

さらに今回実施した各試験の得点の関係をセンター試験の得点も交えて分析した。これによって各試験の特性も明らかになればと思う。

3.1.1で示したように，リスニングテストのスコア分布は，尖度が高く，標準分布から非常に離れているために，そのまま相関分析を行うことは出来ない。これは，相関分析がデータの標準分布と前提としているからであり，データ変換が必要である。Tabachnick and Fidell（2001）に示されたいくつかの方法から，今回はz-得点を用いる変換が行われた。その結果，歪度の値は－.322(Std. error＝.173)，尖度は.247（Std. error＝.344）となり，共に標準誤差の2倍の範囲内に収めることができた。よってセンター試験，QPTのスコアも同方法でデータ変換し，相関分析を行った。詳しく記すことはできないが，センター試験のスコアの尖度も同方法で解決され，相関分析が行われた。結果はTable 6の通りである。

Table 6. Correlational analysis of three test scores

		ZCENTER	ZAL	ZQPT12
Pearson	ZCENTER	1.000	0.299**	0.285**
Correration	ZAL		1.000	0.207**
(two-tailed)	ZQPT12			1.000
		**p<.01		n＝198

Table 6に見るように，全てのスコアが高い相関を示した。センター試験のスコアはリスニングテストのスコアとr=.299（p<.01），QPTのスコアとはr=.285（p<.01），リスニングテストのスコアとQPTのスコアはr=.207（p<.005）の相関を示した。ここで一つ興味深いのは，リスニングテストのスコアがQPTのスコアよりも，センター試験のスコアとより高い相関を見せているということである。QPTにはリスニングセクションがあるため，本来はリスニング問題の無いセンター試験よりも，リスニングテストテストとより高い相関を示すだろうと思われていたのだが，逆となってしまった。この理由は様々に推測できると思われるが，一つ考えられることは，リスニングテストで使用された問題が，リスニング能力よりも文法力や言語知識をより大きく測っていたのではないかということである。この結果は,用いられたリスニングテストが本来到達度テストの目的で作られたこととも関連していると思われる。リスニング能力のみならず，授業などで習った語法，文化に関する知識を試す問題が多かったのではないかと推測される。

3.2　z-得点を用いたシミュレーション

今回のクラス編成では単にセンター試験とリスニングテストの素点の合計が用いられた。また，リスニングテストは50点満点となっているため，クラス内での音声英語の占める役割の大きさも考え，素点を2倍してセンター試験と合計した。

しかしながら，先にも述べたようにクラス編成は集団に基づいて行われるので，集団に基づいた標準化を行い，それぞれに重み付けをした得点，つまりz-得点やT得点などの標準得点を用いた方がより適切なクラス編成が可能であったように思われる。ここではz-得点を用い，クラス編成を再シュミレーションした。今回のクラス編成でセンター試験の満点が200点，リスニングテストが50×2＝100点であったので，同様の重み付けで計算し，シュミレーションを行った。結果は次頁のようになった。

Table 7のパーセンテージは，シミュレーションによるクラスと現在所属するクラスの一致している割合である。以上2つの表を見て気づくことは，初級1から準中級1にかけて素点によるプレイスメントとz-得点によるプレイスメントの違いが大きいことである。Table 7では，基礎クラスから初級1，2と一致の確立が小さくなり，また中級へ向けて高い確率で結果が一致していく。Table 8では準中級1の部分で最も違いが出てきており，平均値のみを見ると初級2のクラスよりも下回っている。しかしながら，Table 7を見ると，初級クラス2の学生はレベル2から5（つまり初級1から準中級2）の範囲に広く分布しており，必ずしもこちらの学生の能力が上回るとはいえないことがわかる。このようなことが起こる大きな原因の一つは,やはり素点だけでプレイスメントの処理をしてしまうことである。プレイスメントは同グループ内で行われるものであるから，グループ内での比較，つまり相対的尺度を用いることが望ましい。そのため，標準得点であるz-得点又はT-得点を総計に用いることが望ましい。これらのクラス編成がどの程度影響を与えるかを測ることは難しいが，このような標準得点を用いることでより適切なクラス編成が可能になると考えられる。

Table 7. z-得点による再シュミレーション

		シミュレーション結果								
		基礎	初級1	初級2	準中級1	準中級2	準中級3	準中級4	準中級5	
現在のクラス		1	2	3	4	5	6	7	8	
基　礎	1	20	5							80.0%
初級1	2	5	10	4	3	1				43.5%
初級2	3		5	6	7	4				25.0%
準中級1	4		3	12	14					73.7%
準中級2	5				5	24				82.8%
準中級3	6						29			100.0%
準中級4	7							27	1	96.4%
中級5	8							1	22	95.7%

Table 8. z一得点シミュレーション後のクラスレベルの平均*

	現在	z-得点
基　礎	1	1.2
初級1	2	2.35
初級2	3	3.45
準中級1	4	3.38
準中級2	5	4.83
準中級3	6	6.00
準中級4	7	7.00
中級5	8	7.96

*クラスレベルは，基礎＝1，初級1＝2，初級2＝3，準中級1＝4，準中級2＝5，準中級3＝6，準中級4＝7，中級＝8で換算した。

4．考察

4.1　実施したテストは信頼性，弁別性，妥当性のあるものであったか。

先ず，リスニングテストであるが，このテストを今後クラス編成の材料に使用することは以下の点から疑問視される。1）このテストは学習内容の習熟度を測定す到達度テストなのでクラス編成の目的として使用するのは妥当な使用法ではない。2）プレイスメントテストではいかに受験者の能力を弁別するかが重要な要素であるが，リスニングテストの記述統計の結果をみると40点付近に

分布が集中していて，分散していない。また，古典的項目分析と項目応答理論の分析結果からいえることは，人社の学生の能力を測定するに適切な難易度と弁別力を示す項目が比較的少ないのでクラス編成の目的に合っていない。3）リスニングテストの信頼性係数の数値（Cronbach α ＝.68）は望ましい数値ではなく，32%はなんらかの理由で誤差が生じていて標準誤差（standard error of measurement）は2.61で，受験者の真の得点は96%の可能性で実際の得点から－5.22～＋5.22点の範囲に収まることを示している。誤差が大きいため，信頼性の観点から項目の改訂や別のテストの検討が望まれる。

QPTについても，本稿の分析結果などから今後継続してプレイスメントテストとして使用することは以下の理由から再検討が必要であると思われる。1）QPTは適応型テストを基に開発され，信頼性と妥当性が実証されたテストであるが，おそらくコンピューター慣れ，コンピューター不安などの問題から，同じ学習者，同じテストであっても2度の受験得点の相関係数はr＝.47（p＜.01）であり，測定誤差が大きかったこと，2）履修登録等の都合上，迅速にクラス編成する必要があり，現状を考えると大勢の受験者に対応するにはコンピューターの台数などに制限があるため，多数の受験者を一斉に実施できる筆記試験の方が効率的であること，等である。

センター試験については，1）受験者の結果をみると得点の分散が十分であること，2）目標基準準拠テストではあるが，先述の通り項目難易度や点双列相関係数の数値等を考慮すると集団基準準拠テストの特性も持っていること，等からセンター試験結果をクラス編成の材料として使用することは可能ではないかと思われる。但し，センター試験には英語の音声を用いたセクションが無いため，4技能の養成を目標としたプログラムで採用する際にはリスニングやスピーキング等の音声に関するテストを併用する必要があると思われる。

4.2　今後，どのようにクラス編成を行えば良いか。

プレイスメントテスト等の得点を用いて，クラス編成を行う際は以下の点に注意すべきである。

1）受験者の能力を広範囲に弁別する項目を多く含む集団基準準拠テストを選ぶべきである。

2）古典的項目分析を行い，以下3の条件を満たす項目のみクラス編成の判断材料に使用する。

(ア)　4岐選択肢項目は難易度が.65程度の数値を示す項目（大友，1996，p.31）

(イ)　項目弁別度が.19以上の数値を示す項目

(ウ)　点双列相関係数が.25以上の数値を示す項目

3）クラス編成では，受験者を相対的に比較するため，素点ではなく，z-得点かT得点に変換した数値をクラス編成の判断材料とする。さらに,より正確なクラス編成を試みるのであれば，Culligan and Gorsuch（2000）が提唱するように，項目応答理論のモデルを用い，受験者の能力値をクラス編成の判断材料とする。

4.3 今後，どのようなプレイスメントテストを実施すれば良いか。

検討の結果，更なる信頼性，妥当性，実用性が高いプレイスメントテストを調査し，試用することとなった。ここでは，簡略化した調査表をAppendixに載せるが，テスト内容，実施時間，結果返送までの所要時間，実証研究データの有無，受験料，支払方法等の様々な要素ついて調査，検討した。

その結果，TOEFL ITP，TOEIC IP，G-TELPが有力候補として挙がった。しかし，TOEFL ITPは得点結果返却にかかる期間が比較的長く，クラス編成には充分な猶予ではないと判断した。TOEIC IPは実施時間が120分であるため，授業時間内での実施が難しく，また，結果返却のための期間も1週間程と長い。この2つのテストは知名度も高く，ニーズもあるであろうが，コストも割高であるため，今回は断念した。

そして，以下の理由からG-TELPを採用することに決定した。

1）実用性の観点から…G-TELPは結果報告が3日で，また実施1週間後には受験者全員に認定書が配布される。また，古典的項目分析結果と生データが入手できる。実施時間もレベルごとに異なるが90分程で授業内に実施することもできる。

2）テストの特性の観点から…一般的にクラス編成試験は集団基準準拠テストを使用するが，Culligan and Gorsuch（1999）同様，1．4．1で触れたBachman（1990）が提唱する条件にも当てはまらないことから，目標基準準拠テストでも項目困難度がある程度制御されているG-TELPを採用し，クラス編成は標準得点を用いて行うのが適切と考えられる。

平成14年10月にはパイロットとして学生数を半分に分けレベル2とレベル3を実施し，分析結果から採用レベルを決定する試みを行う。また，この新カリキュラムの品質保証の一環として，プレテストとポストテストを実施し，期間中に学生の得点がどの程度上昇したかを測定する。このようなプログラム評価を行う目的には目標基準準拠テストであるG-TELPは適していると思われるが，難易度等の受験者特性との相性は未知である。今後も分析し更なる信頼性や妥当性の検証が必要である。

5．最後に

今回のプレイスメントテストの実施，分析，そして新しいプレイスメントテストの選択を通じて，カリキュラム，授業内容，受験者の特性を考慮することの重要性を痛感した。カリキュラムや指導内容，受験者のレベルに合わないテストを選択すれば，どんなテストを用いても信頼性，妥当性，弁別力に欠けてしまう可能性がある。しかしながら，日本の大学生に合ったテストは現在よく実施されている外部テストでも非常に少なく，理論的にはCulligan and Gorsuch（1999）に述べられている通り，受験者に合う外部テストが無ければ，作成することが最も望ましいであろう。しかし，よい自作テストはそう簡単に作れるものではない。そしてプレイスメントを履修登録に問い合わせるには，採点や振り分けにもスピードと入力が要求される。そこで，外部テストの導入を検討するわけだが，外部テストの採用にあたっては，実施の際の費用，時間，手続き，クラス編成のための時間等を考えると，理論的な面では相当な妥協を強いられることも覚悟しなければならない。本プロ

ジェクトにおいても，今回は最終的に目標基準準拠テストであるG-TELPを採用した訳であるが，これはこのテスト結果を実験プロジェクトの効果測定にも用いることを前提としたもので，クラス編成に用いる際には不適切な項目を除き，また得点を標準化して用いることが望ましい。自作テストが難しく，外部テストを用いる際には，各テストの特性の理解と適切な使用法，信頼性，弁別性等の分析が必須であるといえるだろう。

参考文献

Assessment Systems. (1998). ITEMAN (version 3.6). [Computer Software] St. Paul, MN: Assessment Systems.

Assessment Systems. (1997). XCalibre (version 1.10e). [Computer Software] St. Paul, MN: Assessment Systems.

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

Bachman, L. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bond, T. & Fox, C. (2001). *Applying Rasch model: Fundamental measurement in the human science*. New Jersey: Lawrence Erlbaum Associates.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23(1), 65-83.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal*, 12, 121-140.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall.

Brown, J. D. (1997, April). Skewness and Kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 1(1), 16 - 18. Retrieve from http://www.jalt.org/test/bro_1.htm.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, 1(1), 44-59.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D. & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal*, 17(1), 7-30.

Culligan, B. & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal*, 21(1), 7-25.

Culligan, B. & Gorsuch, G. (2000). Using item response theory to refine placement decisions. *JALT Journal*, 22(2), 315-325.

大学入試センター（2001）．*平成13年度大学入試センター試験：実施結果と試験問題に関する意見・評価*．大学入試センター．

大学審議会（2000，November）． *大学入試の改善について*（答申）．http://www.mext.go.jp/b_menu/shingi/12/daigaku/toushin/001102.htm.

Embretson, S. & Reise, S. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

石塚智一，中敏菜緒子，内田照久，前川真一（2001）．テストレットモデルにより英語試験問題の分析．*大学入試センター研究紀要*，30，21-37．

ジョンソン，J.，玉井光江アレン，加須屋裕子（1999）．SLEPテストによる英語能力測定：文京女子大学1年生の分析．*文京女子大学研究紀要*，1（1），141－162．

Kondo-Brown, K. & Brown, J. D. (2000). The Japanese placement tests at the University of Hawai'i: Applying item response theory. *NFLRC Net Work* #20 [HTML document]. Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center. http://www.nflrc.hawaii.edu/NetWorks/NW18/[2002/10/01].

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

大友賢二（1996）．*項目応答理論入門*．大修館書店．

Oxford University Press. (2001). Quick Placement Test [Computer Software]. Oxford: Oxford University Press.

Sands, W., Waters, B., & McBride, J.(Ed.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington D.C.: American Pychological Association.

SPSS. (2002). SPSS for Windows (Version 11.02) [Computer Software]. Chicago: SPSS.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

豊田秀樹（2002）．*項目反応理論［入門編］：テストと測定の科学*．朝倉書店．

土平泰子（1996）．言語テストの歴史と現在：項目応答理論のもたらした新時代．*日本語英語教育史研究*，11，93－110．

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary finding from classroom-based research. *Language Testing*, 13, 318-333.

Appendix

テスト検索結果

テスト名 （開発会社＆公式サイト）	測定技能（出題数）	測定目的，特徴	値段 （購入法）
ACT's ESL Placement Test (ACT;http://www.act.org/esl/sample.html)	-Grammar/Usage (adaptive) -Reading (adaptive) -Listening (adaptive) -can be used separately or in combination -proficiency test	-placing ESL students in university programs -20-30mins to administer -on-line test -computer adaptive -scores delivered immediately	Low
ACCUPLACER (The College Board; http://cpts.accuplacer.com/)	-Reading (adaptive) -Grammar (adaptive) -Listening (adaptive) -Writing	-ESL placement for incoming college students in North America -on-line test -computer adaptive -scores delivered immediately	Low
英語コミュニケーション能力測定テスト （日本英語検定協会） http://casec.evidus.com/	-Vocablary(adaptive) -Grammar(adaptive) -Listening(adaptive) -Dictation(adaptive)	-proficiency test -about 40mins to administer -on-line test -computer adaptive -scores delivered immediately with estimate TOEIC score	初年度に限り使用数無制限で60万
JACET Intermediate Listening Comprehension Test(JACET)	-Listening (30Qs)	-proficiency test -intended for intermediate students -about 40mins to administer -paper & pencil multiple-choice -scores delivered within 10 days	1名1000円 (銀行振込)
国際英検 G-TELP(International Testing Services Center; http://www.eigotown.com/eigocollege/exam/exam1_5.shtml)	-Grammar (L2=26 Qs; L3=22Qs) -Listening (L2=26 Qs; L3=24Qs) -Reading/Vocabulary (L2=28Qs; L3=24Qs)	-criterion-referenced test -90mins for Level 2 -75mins for Level 3 -paper & pencil multiple-choice -scores delivered within 2-3 days ーレベル2：ビジネスなどの日常生活の実際の場面で，ネイティブと支障のないレベルでコミュニケーションができる。 ーレベル3：日常生活の限られた範囲の表現方法を用いて，ネイティブとコミュニケーションができる。	Level2 団体受験では200名で2300円 Level3 団体受験では200名で1840円 (銀行振込)
Michigan English Placement Test (Michigan English Language Institute; http://www.lsa.umich.edu/eli/testpub.htm#EPT)	-Listening (20Qs) -Grammar (30Qs) -Vocabulary (30Qs) -Reading (20Qs)	-proficiency test -placement -intended for beginning to advanced intermediate students -75mins to administer -paper & pencil multiple-choice -scored on same day local hand scoring	200名で $650 (クレジットカード)

テスト名 (開発会社＆公式サイト)	測定技能（出題数）	測定目的，特徴	値段 (購入法)
SLEP(ETS; http://www.toefl.org/educator/edslep.html)	-Listening (75Qs) -Reading (75Qs)	-proficiency test -placement -intended for secondary schools and community colleges -85mins to administer -paper & pencil multiple-choice -scored on same day local hand scoring	200名で $1550 (クレジットカード)
TOEFL-ITP (ETS; http://www.cieej.or.jp/toefl/index.html)	-Listening (50Qs) -Structure/Written expressions (40Qs) -Reading (50Qs)	-proficiency test -intended for intermediate to advanced students -115mis to administer -paper & pencil multiple-choice -scores delivered within 2 weeks	200名で一人あたり ￥2560 (口座振込)
Pre-TOEFL ITP (ETS; http://www.cieej.or.jp/toefl/index.html)	-Listening(30Qs) -Structure/Written expressions(25Qs) -Reading(40Qs)	-proficiency test -70mins to administer -targeting students whose scores are below 500 TOEFL -paper & pencil multiple-choice -scores delivered within 2 weeks	200名で一人当たり ￥2430 (口座振込)
TOEIC(ETS; http://www.toeic.or.jp/)	-Listening (100Qs) -Reading (100Qs)	-proficiency test -120mins to administer -intended for mainly intermediate but all levels of students -paper &-pencil multiple-choice -scores delivered within 2 weeks	公開テストは￥6615 非公開テストは3850円 (口座振込)