名桜大学紀要，(19)：65-76（2014）

【研究ノート】

TOEFL iBTにおけるスピーキング測定とライティング測定の妥当性，信頼性，実現性の検証

Examination of the Validity, Reliability and Practicality of the Speaking and Writing Test of TOEFL iBT

渡慶次　正　則

要旨

本稿は，TOEFL iBTのスピーキング力測定とライティング力測定の妥当性，信頼性，実現性を検証することを目的とする。ワークショップでの筆者の受験者と採点者としての体験と資料を基に検証を行った。結果は，第１にスピーキングもライティングも採点の信頼性を高めるかなりの配慮がなされている事が明らかになった。第２にスピーキングの妥当性については，表面妥当性と内容妥当性は高いが，構成概念妥当性については弱い。第３にライティングの妥当性は複数の言語技能を用いて統合タスクを課しているが，構成概念は伝統的な論述的形式で，講義と文献のポイントを網羅的にまとめる基本的な書く能力に留まっている。最後に，TOEFL iBTの実現性については，表面的なアピール度は高いが，施設や費用，採点者などの訓練など課題は多い。

キーワード：外国語，評価，妥当性，信頼性，TOEFL

１．はじめに

経済や，政治，文化分野等での世界的なグローバリゼーションの進展とともに，英語の世界共通語（Lingua Franca）としての役割やニーズは増大している。

日本人学生のTOEFL（Test of English as a Foreign Language，以下TOEFL）平均スコアがアジア諸国25か国中23位である事実（ETS 1996～1997）[1] を受け，文部科学省は日本人の英語能力を伸長させることを喫緊の課題とし，グローバリゼーションに対応しようと，「英語の使える日本人を育成するアクションプラン」（文部科学省，2002年）を実施した。具体的な目標としては，例えば，英語教員の英語能力を英検準１級レベル以上，またはTOEFL PBT（紙媒体試験）550点以上，TOEIC（Test of English as International Communication，以下TOEIC）730点以上を獲得させること，英語によるイマージョンプログラムを行うSELhi（Super English Language high school）の高等学校を国内で増やすことなどを政策として打ち出した。しかし，そのアクションプラン実施後もTOEFL iBT（インターネット・ベースト）における日本人の英語平均点数は相変わらずアジア諸国では最下位に近い（ETS，2011）[2]。

この様なTOEFLテストの日本人受験者の危機的な英語能力に対して，突然提案された改善策がTOEFL iBTをAO入試等として大学入学選抜方法の一部として導入する計画である（文部科学省，2013）。この提案は高等英語教育への波及効果（backwash）だけではなく，大学における英語能力の測定にも今後大きな影響を与える。さらに，留学の推進が叫ばれる昨今，米国のほとんどの大学はTOEFL iBTスコアを具体的な入学条件としており[3]，日本人学生には留学への大きな壁となっている。

この様に日本の社会的かつ大学の状況において，TOEFL iBTの存在は無視できない状況となりつつある。従って，本稿では妥当性，信頼性，実現性の３要素（Hughes, 2003）を中心に，筆者が参加したPropell Workshop for the TOEFL iBT Test（CIEE主催）での受験者と採点者の双方の役割の体験と，講師や参加者からのフィードバックを基に，実証的な検証を試みる。TOEFL iBTを分析することにより，テストの内容と測定基準を精査し，同テストを日本の大学へ導入する可能性を探りたい。TOEICやこれまでのTOEFL試験とは異なり，TOEFL iBTに初めて導入されたスピーキングとライティングを焦点化して検証する。

２．先行研究

2.1　英語能力テスト発達の概観

歴史的に英語能力テストは単語や文法，発音などの部分的な言語能力を図る弁別的な（discrete-point）試験（例，Rivers, 1968）から始まり，ヨーロッパにおける1970年代の各国間の労働者の移動の活発化に伴い，共通された言語テストのニーズが高まり，言語のコミュニケーション能力を測定する統合的なテストに移行してきた歴史的経緯がある（Weir, 1990）。従来の文法や語彙，発音などの弁別的な測定ではなく，コミュニケーション能力を測定するために４技能の言語知識を統合的に測定しようとCloze Testや面接などを用いた測定がなされた（Ollers, 1983）。その概念はTOEFLテストでも引き継がれ，英語４技能について個別に弁別的な測定を行うのではなく，日常の生活に近い状況で４技能が統合して用いられる統合的な測定（例えば，リスニングとライティングを組み合わせて測定）を試みているのがTOEFL iBTで，特にスピーキングとライティングでは複数の技能を統合した英語能力測定が行われている。

一方，日本でもヨーロッパの言語測定の影響を受けて，言語使用の目的や概念に応じて英語を測定する動きが1980年代中盤から始まった（青木・田中，1985）[(4)]。その測定方法は，現在ではCEFR（Common European Framework Reference）（Council of Europe, 2001）として英語のみならず日本語を含めた諸言語の測定方法として用いられている[(5)]。その言語観は文部科学省にも引き継がれ小学校，中学校，高等学校の各々の学習指導要領で「言語の働き」を中心に指導することが明記されているが，実際には使用されている英語教科書のほとんどが文型中心であり，理想と現実の乖離が生じている。

2.2　TOEFLテストの発達

TOEFLはこれまでPBT，CBT，iBTと３世代のプロセスを経て発達してきた。TOEFLは米国留学の判定試験に用いられ，一方，イギリスやオーストラリアへの留学判定試験にはIELTS（International English Language Testing System）[(6)] が頻繁に用いられる傾向がある。

まず，PBT（ペーパーベーストテスト）は1999年まで公式に実施され，最高点は677点であり，その構成はListening（聴解）50問（30分～40分間），Structure and Written Expressions（文法）40問（25分間），Reading（読解）55問（50分間）の３つのセクションで測定される。つまり語彙･文法，聴解，読解は測定するが「書く能力」と「話す能力」を測定しない。所要時間は約２時間程度である。難点は，試験中にメモを取ることが一切許されなかった事である。解答方法としてはマークシート用紙を用いる。その点では，TOEICや実用英語検定試験と相似している。「書く能力」については，他にTest of Written English（約30分）で１つのトピックについてエッセイを書かせる試験が用いられた。

PBTテストの次に2000年から登場したのがCBT（コンピュータ・ベースト）テストで300点が最高点である。CBTテストはコンピュータを用いて解答し，PBTのテストに加えて「書く能力」を測定することができる。特徴としては受験者に応じて問題を変えることができる，いわゆる項目応答理論（IRT）[(7)] に基づいたコンピュータ・アダプティブ・テスティング（computer-adaptive testing）を用いた方式があり当時は脚光を浴びた。聴解力と読解についてはヘッドホンとコンピュータスクリーンを用いて解答する。「書く能力」（writing）については与えられたトピック１題について直接，コンピュータに入力する方法を採る。「書く能力」の試験のみメモを取ることが許された。しかし，CBTの登場は2006年から採用された「話す能力」（speaking）を追加して，４技能をコンピュータで測定するiBT（日本では2006年より利用）の出現によりほとんどの地域で姿を消してしまった。

TOEFL iBTについては次のセクションで詳しく述べる。

他にTOEFL IPT（インスティチュショナル・プログラム・テスト）[(8)] があり学校などの団体で英語の能力別クラス編成や達成度能力測定に用いられるが，公式なTOEFLスコアとしては米国の大学が認めるケースは少ない。

2.3　TOEFL iBTについて

TOEFL Teacher Manual（2012年改訂，以下，TOEFL Manual と呼ぶ）によるとiBTは次の特徴を持つ。

⑴　リーディング（30点），リスニング（30点），スピーキング（30点），ライティング（30点）の配点で合計120点満点である。

⑵　テストの目的は米国大学での講義や大学生活で必要な英語能力が備わっているかを測定する。使用目的としては米国大学の学部や大学院の入学基準の一つとして用いられる。必要とされる英語能力は学究的英語（academic English）と大学生活に必要な日常会話が多い。

⑶　インターネットを用いて，リスニングはヘッドホンを通してスクリーン上で解答，リーディングはスクリーン上の英文を読んで解答，ライティングはキーボードを用いて書き込む，スピーキングはヘッドホンを用いて聞き取り（あるいはスクリーン上の

表１．TOEFL iBT テストの形式（出典：TOEFL Test Prep Planner（ETS ホームページ PDF pp.11-12, p.27, p22.）

セクション	質　問　の　数	所　要　時　間
リーディング	大問３から４　各14小問*	60分から80分*
リスニング	大問４から６の講義，各６小問* 大問２から３の会話で各５小問*	60分から90分*
休憩		10分
スピーキング	６タスク：Independent question ２問 Integrated question ４問	20分
ライティング	Integratedタスク １問 Independent タスク１問	20分 30分

補足説明
・リーディング 問題（60分から80分*），リスニング問題（60分から90分*）と時間に幅があるのはいずれかで２問程度，ETSの調査サンプルのために利用されるためである。受験者はどの問題が調査に用いられるはわからない。スピーキングとライティングが調査サンプルとして利用されることはない。
・リスニング問題*は，主に大学生活の２，３名の会話（約３分間）について各５問，講義（３分から５分間）について各６問が出題される。
・リーディング問題の内容は大学の講義で読まれる様々な学問分野（例，心理学，生物学，歴史学）の基礎的な文献である。基本的に３つの形式があり，(1) 多肢選択から１つの解答選択肢を選ぶ問題（例，文中の用語を尋ねる，語彙の意味，パラグラフの大意，選択された部分の大意，全文の大意など），(2) 全文の中で適当な箇所に一文を挿入する問題（例，“Some historians believe they can be established.”の文を全文中にある■の部分に適当な挿入箇所をクリックする），(3) 複数の解答選択肢に１つ以上の解答を選ぶ（例，サマリー文に対して６つの選択肢から３つを選ぶ），この問題は配点が高い。（参考: Barron's TOEFL iBT 13th edition）

英文を読み），タスクについて付属のマイクロホンに向かって話し，電子データが録音される。

(4) リスニングとリーディングは従来通りだが，スピーキングとライティングは他の４技能と組み合わせた統合（integrated）テストである。

(5) 試験は１週間に１回程度受験できる。

(6) 試験時間はTutorialを含めないで４時間以上を要する。

(7) 受験料は225ドルで約21,825円（１ドル：97円の為替レートで計算）と安価ではない（受験７日前までの金額）。

TOFL iBTの試験形式の概要は表１で示す通りである。本稿では主にTOEFL iBTテストのスピーキング能力とライティング能力の測定について後述の部分で詳しく論じる。

基本的に，リーディング問題とリスニング問題については，TOEFL PBTやTOEFL CBTの問題形式と大きな差異はない。但し，講義の聞き取り問題は，各々３分から５分間所要する講義を６題ほど聞き取るために，集中力の持続と要点を理解するメモの取り方の工夫が求められる。リーディング問題については，出題傾向は大きな変化はないが，コンピュータの特性を生かし，文章中に一文が挿入される部分をクリックする点が真新しい。リーディング問題とリスニング問題については詳しく本稿で述べることは避ける。

2.4　妥当性，信頼性，実現性について

本稿では，Huges（2003）により示された妥当性（validity），信頼性（reliability），実現性（practicality）の観点からTOEFL iBTを検証する。以下に妥当性，信頼性，実現性について説明する。

妥当性とは，意図するものを的確に測定しているかどうかであり，的確に測定していれば「そのテストは妥当性がある」と言える。妥当性には様々な妥当性があるが，本稿ではテストの内容が測定すべき事項を含んでいるかを示す内容妥当性（content validity）を検証する。例えば，文法の測定には文法項目が含まれると妥当性が高いと言えるが，スピーキングの測定を発音記号の知識で測定するのは内容妥当性が高いとは言えない。さらに，言語能力に対する理論的な概念の妥当性を測る構成概念妥当性（construct validity）は，言語能力の理論的構成要素に基づき検証することで，妥当性では最も重要である（Hughes，2003）。例えば，英語コミュニケーション能力を測定する場合に，「コミュニケーション能力が高い」とは具体的にどのような構成要素を測定するか，理論的に重要な問いである。他に，受験者や管理者，保護者にとって試験が表面的に説得力があるかどうかを判断する表面妥当性（face validity）を検証する。

信頼性とは，測定する言語能力が同じ条件下で測定した場合に，同様な測定結果を示すかという事である。本稿では，受験者のデータは入手できないために，受験者の点数についての信頼性の検証はできないが，信頼性を高めるための問題作成者の観点について検証する。さらに，採点者によって採点結果が偏りがないかを判断する

採点者信頼性（inter-rater reliability）の観点から論じる。

実現性とは，テストが妥当性や信頼性が高くても実際に実施できるかどうかを判断することである。例えば，英語コミュニケーション能力を測るために，大学センター試験で受験者全員に対して英語で面接試験をする事は，妥当性も信頼性も高いと思えるが，実際には費用の面や，時間的な制約，面接官の養成などで実施は困難である。

３．調査方法

筆者はETSから公式認定を受けているCIEE（国際教育交換協議会日本支部）主催によるTOEFL iBTの指導者対象ワークショップを受講した。認定トレーナーから採点の具体的な留意点の指導を受け，受験者としてテスト（主にスピーキングとライティング）を受講する機会を得た。ワークショップでは講師と参加者でテストについて具体的な討論があり，テストに関する多くのフィードバックを得ることができたので，文献資料とワークショップのフィードバックを用いて検証したい。ワークショップについて以下の様に実施された。

参加ワークショップ：
Propell Workshop for the TOEFL iBT Test
日　　時：平成25年11月17日（日）
午前９時30分～午後４時
主 催 者：CIEE
参 加 者：講師と受講者18名（ほとんどが大学英語教員）
配布資料：TOEFL Teacher Workshop Manual，TOEFL More Skills & Activities，TOEFL Reading, Listening, Speaking & Writing，TOEFL Prep Planner

４．調査結果と分析

スピーキングとライティングについて受験者の体験に基づいたフィードバック，さらに採点者としての体験での懸念や疑問を主に妥当性，信頼性，実現性の観点から論じる。

4.1　スピーキングの検証

4.1.1　スピーキング問題の体験と検証

筆者はTOEFL Reading, Listening, Speaking & Writing Activities（ETS, 2012，以下 TOEFL Activities）を用いて（pp.30-31），独立タスクの個人的嗜好タスクを体験した。

個人の馴染みのある話題について話すので，話題の展開については難しくないだろうが，個人の体験や具体的な事例を挙げながら45秒間，よどみなく話せない受講生もいた。

4.1.2　スピーキング応答の測定

TOEFL Teacher Manual（前出）によると，採点は総合的（holistic）に測定され，最低０から最高４の範囲で，判定される。６つのタスクに対して各々違う採点者が採点を行う。重要な３つの基準項目として，発表の仕方（Delivery）つまり，明瞭さ，流暢さ，躊躇などに注目し，言語使用（Language Use），つまり文型や語彙の選択などに注目し，トピックの展開（Topic Development），つまり文章の整合性と構成，論拠の提示などに注目する。

ワークショップでは，筆者は実際にスピーキングの採点を体験し，受講者で採点基準説明（rubrics）[9] や方法について討論をした。以下に筆者のコメントを中心に受講者の判定をレベル別に記載する。確認のために，ワークショップで使用されたCDを再度聞きながら記載した。TOEFL Teacher Manual (pp.36-37) の注釈も参照した。次の独立タスクのスピーキングの採点をした。

独立タスク:「キャンパス内とキャンパス外ではどちらに住む方が良いか述べなさい。」

サンプル話者Ａ：［ほとんどの受講生がレベル１（最も低いレベル）と判定］
（音声データ49秒）

Umm, soなどを頻繁に用いていた。話す速度が遅く，語の固まりが不自然に途切れる。語彙は的確な語を用いているが，完全な文で話すことがほとんどない。理解はそれほど難しくない（筆者の評価コメント）。聞き手が理解するのに大きな努力を必要とされる。個々の単語を用いて意味を伝えようとしている。例えば，“...students / relationships / communication so/very familiar /human / care...”（Manual注釈を追加）。

サンプル話者Ｂ：［レベル２（実際には受講生はレベル１と判定した場合とレベル２と判定した場合に分かれた）］
（音声データ46秒）

話す速度はサンプル話者Ａより速い。流暢(fluidity)さはある。話者Ａよりは，一度に話す語の長さは長い。しかし，発音は話者の母語の影響があり，話者Ａより聞きづらい。文はほぼ完全な文で話してお

り，文法的な問題はないが，レベル１との違いは流暢さであるようである（筆者評価コメント）。自分の考えを充分に支持できる具体的な例を示せていない。例えば，“it is better to live in the dorm because it is easy for them to live,...don't raise much problems”（Manual注釈を追加）。

サンプル話者Ｃ：［レベル３（受講生はレベル３と判断した人が多かった）］
（音声データ45秒）

全体的に明瞭で理解しやすい。ほぼ自然な流れで話している。話者のくせなのかbecauseを頻繁に用いる。やや理解し辛い発音もあるが，寮に住む利点を幾つか指摘した（筆者評価コメント）。小さな文法的な誤りがある。例えば，“to make student live in the dormitory”とか“I think it safe”などと言っていた（Manual注釈を追加）。

サンプル話者Ｄ：［レベル４（最高レベル）（受講生のほとんどがレベル４と判定）］
（音声データ45秒）

話す速度は自然で，流暢さがあり，躊躇（えーと，あのー等）がほとんどない。語彙や表現が豊富で，いつくかのポイントを指摘していた。母語の影響を少し受けた発音であるが，充分に聞き取れる。４名のサンプルでは最もレベルが高い（筆者評価コメント）。説明に対する理由を述べ，考えがスムーズに展開している（Manual注釈の追加）。

採点者の体験と採点基準説明資料（rubrics）から明らかになった事は以下である。評価で重要視している事は，流暢さや豊富な語彙や文の使いこなし（control），全文で話すこと（choppyにならない），ポイントを的確に述べること，話すことを制限時間内で持続する（sustain）ことなどである。逆に，評価が低いのは，文が細切れである，躊躇の語を頻繁に用いる（Ummなど），全文で話さないこと，使用する語彙や文法が貧弱であるなどである。レベル４でもわずかな誤りは許容され，発音の違いやイントネーションは理解を妨げなければ大きなマイナス印象にはならない。

筆者は実用英語検定試験の面接試験官として約９年間の経験を持つ（2005年から2013年まで準２級と２級受験者約110名の面接試験をした）。実際にTOEFL iBTスピーキングの採点を行い，レベルごとの明確な違いを理解するのはかなりのトレーニングが必要だと感じた。ワークショップ講師に，「実際には受験者を相互比較しながら，採点を修正することがないのか」との問いを発したが，原則としてはないとの回答であった。今回のトレーニングはレベルの違いが顕著なサンプルが選択されていたので違いは明白であったが，実際のテストではRubricsに厳密に従い，レベルを判断するのは容易ではない。

4.1.3　スピーキング測定の妥当性

このセクションでは，スピーキングの内容妥当性（content validity），構成概念妥当性（construct validity），表面妥当性（face validity）について述べる。

表２で示されるように，スピーキング問題は独立タスクと統合タスクに大別される。独立タスクは個人的経験や考え，対比する事例を用いながら論拠を述べたりする，日本人学生には馴染みのある問題である。統合タスクは大学の講義を想定して，読む，聞く，話すを組み合わせて，話す能力を測定する。

内容妥当性はスピーキングの内容に相応しい内容であるかということである。TOEFL iBTは大学の講義で想定されるスピーキングまたは，キャンパス内外で起こる大学生活についてのスピーキングを想定している。TOEFL iBTスピーキング問題（表２）によると，独立タスクのタスク１「個人的な嗜好」は，大学生活で頻繁に起こり得る話題について自分の嗜好を表し，理由を述べる事で，基本的なレベルで想定されるスピーキング活動である。タスク２「選択」は，対立する行動や行動過程について個人の立場を選択し，理由を述べる事は，論拠に基づいて個人の価値判断をする必要なスピーキングだと考える。

タスク３，４，５，６はスピーキングとリスニング，リーディングを組み合わせた統合タスクである。大学の講義ではスピーキングとライティングの組み合わせも考えられるが，限定されたテスト時間ではスピーキングと他技能のすべての組み合わせは不可能であり，最も起こりうる状況での組み合わせのタスクであると推測する。Barron's TOEFL（2010）によると，例えばタスク３は「留学生の健康保険の加入の必要性について」，タスク４は「南極での天然資源についての講義」，タスク５は「大学生の生活費の問題と解決策についての会話」，タスク６は「文献資料の活用についての講義」である。いずれのタスクも留学生が大学生活や講義で一般的に経験する内容や場面に必要なスピーキングと一致していることは間違いないが，全体的に必要な言語能力の測定を網羅しているかという問題は次の構成概念妥当性で論じる。

構成概念妥当性は，測定内容や方法が言語を構成する概念や理論と一致しているかを問う。TOEFL iBTのスピーキング測定では，大学講義で必要なスピーキング能力とコミュニケーション能力を測定する事を目的としている。まず，大学の講義で必要な言語能力は基礎的な「読

表２：TOEFL iBTスピーキング問題（出典：TOEFL Test Prep Plannerを要約, 2013, p.15）

タスクの種類	タスクの説明	所要時間
独立タスク（Independent Tasks）		
タスク１：個人的嗜好	あるカテゴリ（例:重要な場所や人々，場所，出来事，活動など）から個人的な嗜好を表現し，正当性を主張する。	準備時間：15秒 応答時間：45秒
タスク２：選択	２つの対比する行動や行動過程について個人的な選択をし，正当性を主張する。	準備時間：15秒 応答時間：45秒
統合タスク（Integrated Tasks）		
リーディング／リスニング／スピーキング		
タスク３：キャンパスの場面 話　　題：適合と説明 タスク４：アカデミックコース 話　　題：一般／特定	・キャンパス関連の読解文（75語から100語）を読む ・読解文の問題にコメントする聴解文（60秒から80秒，150語から180語）を聞く ・読解文章の範囲内で話者の意見をまとめる ・用語やプロセス，コンセプトなどを定義する読解文章（75語から100語）を読む ・用語やプロセス，コンセプトを説明する例を示す講義から抜粋（60秒から90秒，150語から220語）を聞く ・読解文章と聞いた講義から理解した重要な情報を組み合わせて，伝える	準備時間：30秒 応答時間：60秒 準備時間：30秒 応答時間：60秒
リスニング／スピーキング		
タスク５：キャンパスの場面 話　　題：問題／解決	・学生に関連した問題とその解決に関する会話である聴解文章（60秒から90秒，180語から220語）を聞く ・問題の理解と解決についての考えを示す	準備時間：20秒 応答時間：60秒
タスク６：アカデミックコース 話　　題：サマリー	・用語やコンセプトを説明する講義の抜粋で，説明のための例を示す聴解文章（90秒から120秒，230語から280語）を聞く ・講義をまとめ，例と全体とトピックとの関係について理解した事を示す	準備時間：20秒 応答時間：60秒
総　　計		20分

む，書く，話す，聞く」言語能力と講義で必要な学究的な言語能力が必要となる。TOEFLテストでは，これまで4技能が個別に測定されてきたが[10]，日常生活に近い形で，４技能を組み合わせたiBTテストは高く評価できる。しかし，通常の講義ではグループ・ディスカッションや講師との質問と応答において，双方向性である事が通常であるが，iBTが一方向性のスピーキングのタスクであることは，自然さに欠け，次のコミュニケーションの議論と関連する。

コミュニケーション能力については様々な定義があり（例，Canale & Swain, 1980; Backman, 1990），一致した学問的な見解は確立されていないが，共通したコミュニケーションの特質についてiBT スピーキングを論じたい。まず，「意味のやりとり」（negotiation of meaning），つまり双方向性である事が必要されるが，iBTスピーキングは受験者がタスクに対してコンピュータに一方的に話すのみで，対面式面接の様なやりとりは発生しない。次にコミュニケーションでは場面に応じた（context），つまり対話者や状況によって適切な表現や言い回しが求められるが，iBTのタスクでは一般的な聞き手を想定したスピーキングになっており，弱点となっている。一方，スピーキングの流暢さ（fluency）と正確さ（accuracy）の両方を測定する事を評価基準としており，コミュニケーションの特質と合致している。全体的にはスピーキングのコミュニケーション能力を測定するテストとしては弱いと言わざるを得ない。しかし，スピーキングとリスニング，リーディングを組み合わせた統合タスクを利用している点は評価されるべきで，次の表面妥当性の論議と関連する。

表面妥当性とはテストが受験者や教員，学校管理者などが求めている試験内容や結果になっているかを問う。iBTスピーキングでは，これまで教育機関で個別に4技能が学習・指導されかつ測定されてきた事と，実際の生活では4技能が組み合わされて使われていることの矛盾の解決を試みており，見かけとしてはアピール度が高い。

4.1.4　スピーキング測定の信頼性

iBTスピーキングテストの正確な受験生のデータを入手していないので，このセクションではスピーキング測定で懸念される，採点者の測定の信頼性に焦点化して論じる。

TOEFL iBTは採点者の採点の偏りや一貫性の欠落，

体調不良等による不適格などの問題を克服するために，試験当日に，採点者の検査（calibration）を行い，適格と認められた採点者のみが採点を行う方式を採用しており，信頼性は高まっていると考える。

スピーキングの採点では，６タスクに６名の異なる採点者が測定を行い，採点の偏りや主観性に影響されないように採点者間信頼性（inter-rater reliability）を高める方式を採用している。６名の採点者が独立して採点を行い，極端な採点については統計処理により点数の調整がなされるために，信頼性はかなり高いと考える。

4.1.5　スピーキング測定の実現性

実現性とはテストが費用や施設，採点者などの点で実際に実現可能かどうかを問う。このセクションでは日本の大学というコンテクストで論じる。TOEFL iBTは日本の都市地区では約１週間ごとに受験が可能で，受験の回数は頻繁にあり，大学での到達テストやプレイスメントテストとしての活用は不可能でない。しかし，受験料が高額である点や多数のコンピュータ台数が必要であることが困難点である。さらに防音の施設を持たない試験会場でのスピーキングは，隣の受験生の声に妨害される事が頻繁に起こり得るため，信頼性にも係る大きな問題である。

4.2　ライティング測定の検証

このセクションではライティングの採点体験に基づき検証，妥当性，信頼性，実現性について検証する。

4.2.1　ライティング採点の体験と検証

筆者はスピーキングと同様にワークショップに参加してライティング採点を体験した。ライティングは実際には行わなかったが，TOEFL Teacher Workshop Manual（ETS, 2012，pp.42-53）を用いてライティングの採点を体験した。以下に体験とフィードバックを述べる。ライティングの測定結果は，０から５までの６段階のスケールで評価される。採点は，細部を採点するのではなく，全体的（holistically）に文章の一貫性と正確性を測定する。

タスク実例（統合ライティング）タスク（表３を参照）
題　目：「グループプロジェクトの長所と短所」
活動１：題目について300語程度の文章を読む（３分間）
活動２：題目に関連した教授の講義を聞く（２分間，300語程度）
活動３：聞き取った文のポイントをまとめ，読んだ文章に関連付けて説明する文を書く

［ライティング・サンプル］

ライティング・サンプルを用いて採点を行った。受講者の間では，採点に大きなばらつきが生じ，ライティング測定と指導の難しさを痛感した。紙幅の関係で，２サンプル（サンプルＡとサンプルＢ）のライティングを比較しつつ，本稿巻末の資料の基準文例（サンプル基準文レベル１）と注釈も参照しながら述べる。

サンプルＡ：レベル３（筆者はレベル４と判定）

The lecture and the paragraph are completely contradictory. First of all, the lecture says that the work the group is going to be recognized in a hole, not individually, giving no space for this last. Second, the lecture says that the work is going to take more time once that the group is going to take more discussions and take time to decide how are they going to proceed within the group. Besides that the one who leads the group has the power, somehow, to see if the discussion is going to proceed or if it's going to be hidden, and, in this way, make his/her personal desire kept in front of the others. We can also percept that a group there will be someones who's going to carry the group on their back and there'll be others who'll be static without doing even a single thing. By these points of view we can see that there advantages and disadvantages in group working.

文法的な誤りが，散在するが意味の理解を大きく妨げる程度ではない。講義と資料文のポイントを列挙している。単語の綴りミスは１か所である（以上，筆者のコメント）。この応答文はよく構成されておりポイントを正確に伝えているが，書き手は資料文と講義の内容をうまくカバーしていない。認識についてのポイントと影響者の役割はあいまいである。応答文には不正確で不明瞭な表現がある（以上，Manualの注釈）。

サンプルＢ：レベル５（著者はレベル４と判定）

In the lecture, doubt was expressed concerning the advantages of the recent trend of forming teams to tackle projects, which was mentioned in the reading.

To begin with, the lecturer argues that although a group tends to have a greater resource of skills and expertises, these resources may not

necessarily used. according to a recent company project, it was found that one or two members dominated over the whole group, when the dominant members asserted or banned an idea, most of the other group members would follow their ideas and 'suppress' the other ideas that were suggested, even if the other ideas were more creative and innovative.

Secondly, it was proved that, or the contrary of the reading, progress in the project was very slow. this was the result of long debates over reaching a compromise as ideas were diverted and concenus took a lengthy period of time.

Thirdly, as a group would be credited collectively, quite a number of unfair situations appeared. in the group. It was found that some members did not work hard at all and got a "free ride". However, those worked harder were not rewarded for their extra efforts as their individual efforts would not be recognized.

Concluding, via the results of a recent company that adopted the "group method" of tacking projects, the lecturer projected doubts that contradicted with the central standpoint of the reading. The lecturer believes that skills and expertise cannot be maximized in a group, progress is slow and the overall results of the team is not a fair assessment of the individual members of the group-which contradicts with the central standpoint of the reading.

文法的な誤りが２か所ある（experitises［第２パラグラフ］，concenus［第３パラグラフ］）。文の始まりが小文字で始まっている文が２か所ある。講義と説明文の内容のポイントをもらさず，説明している（以上，筆者コメント）。この応答文はよく構成され，書き手は講義と説明文の対比する点を説明している。唯一識別できる誤りは，文の最初の語を大文字にすること，綴りミスなどの機械的な事と小さい文法ミスである（以上，Manualの注釈）。

採点の体験と採点基準説明資料（rubrics），Manualの注釈から明らかになったのは以下の点である。文法や綴りミスは大きな誤りでなく，講義や説明文の内容のポイントをもらさず説明すれば高得点が得られる（サンプルＢの場合）。綴りミスが少なくても，講義と説明文の内容のポイントをすべて含まなければ高得点は得られない（サンプルＡの場合）。文章の量は制限時数を超えても減点はない。本稿の巻末の資料はレベル１の基準文であるが，資料からコピーした文が多く，難解な語を用いているにも係らず，書き手自らの言い換えや要約がほとんどないために非常に低い採点になったと考える。

4.2.2　ライティング測定の妥当性

このセクションではライティングの妥当性について述べる。

表３によると，統合ライティング・タスクは，一般的に講義などで観察される文献を読んだり，講義を聞いて対比しながらポイントを書いてまとめたりするタスクである。独立タスクは，個人的な意見や立脚点について論拠を示しながら論述する従来の形式のライティングである。

ライティングの内容妥当性については，独立タスクは対立する考えに対して，書き手の立脚点を選択し，論拠を示しながら論述的なエッセイを書く伝統的なスタイルである。英語圏の大学では，問題点を持つ題目について論述形式のエッセイ課題を与えられるのは，一般的な学習形態で講義の内容と合致していると言える。統合タスクは，講義で良く観察される光景で，事前に与えられた文献資料を読み，講義では講師が文献資料に異なる視点から批評や新しい利点や欠点の情報を与える。受講者（書き手）は各ポイントを自分の表現で書き直し（paraphrase），情報を分類化したり，主要なポイントと末梢的ポイントを区別してまとめたりし（summarize），全体の趣旨の一貫性やバランスを整えて仕上げる（synthesize）。iBTテストでは，６つ程度の講義を聞くリスニング問題で，講義のノートテイキングを行い，リーディング問題でアカデミックな内容の資料を読むので，やや重複する部分がある。

構成概念妥当性については，前述のparaphrase, summarize，synthesizeはアカデミックな最も基本的なスキルである。アカデミック・ライティングは学習の段階，学問分野，学習課題により多様な形態を持つが，TOEFL iBTでは大学の初年次を想定していると考える。初年次の学習スキルについては，例えば，Cottrell（2003）によるとアカデミック・ライティングの共通する特徴は，根拠資料を利用する事，比較・対比する事，評価する基準を用いる事，問題解決の複雑さへの認識を持つ事，論旨に従う事，問題に対する立脚点を持つ事，決められたスタイルに従う事，論旨が一貫している事，感情的に中立である事などとしている。iBTの独立タスクと統合タ

表３．TOEFL iBT ライティング問題（出典：TOEFL Test Prep Planner要約, 2013 p.32）

タスクの種類	タスクの説明	所要時間
タスク１		
統合ライティングタスク リーディング／リスニング／ライティング	・アカデミックな題目について短い文章（230語から300語）の文章を読み，メモを取っても良い（講義を聞いている間は文章がスクリーンから消え，書くときに現れる） ・同じ題目について違う観点からの講義（230語から300語）を聞く（読解文章に関連した付加的な情報を聞き，メモを取っても良い） ・聞き取った文章での重要なポイントを英語の散文でまとめ，読んだ文章と関連付けて説明する文を書く（150語から225語程度で書く。それ以上を書く事には減点はない）	・３分間で読む ・２分間聞く 合計所要時間 20分
タスク２		
独立タスク 経験と知識からのライティング	・ある問題について自分の意見を説明し，支持するエッセイを書く（最低300語を含むこと。それ以上書いても良い） ・単に個人的な嗜好や選択を上げるのではなく，自分の意見や選択に論拠を示さなければならない ・一般的なエッセイの質問は次の様な文章になる —次の文章に賛成ですか，反対ですか あなたの答えを支持する理由と具体的論拠を用いなさい —［X］と信じている人もいるし，［Y］と信じている人もいる。この２つの立場のどちらを好み／賛成ですか。理由と具体的な論拠を用いなさい。	所要時間30分

スクがすべてのアカデミック・ライティングを包括しているとは言い難いが，大学の講義において最も基本的な共通するライティング能力を測定しているとは言える。

表面妥当性についてもiBTライティング・タスクがアカデミック・ライティングを全体的に網羅しているようには決して見えないが，米国の大学教育を経験した者には大学の講義での典型的なライティングである事は直感的には理解できる。

4.2.3　ライティングの信頼性

iBTテストではライティング測定結果の信頼性に大きな配慮をしていることが伺える。

試験当日に行う採点者の検査（calibration）に加えて，TOEFL Test Prep Planner（2013）によると４人（種類）の採点者によって採点される。統合タスクは２名の採点者によって採点され，独立タスクは１名の採点者と機械による採点プログラム（e-raterと呼ばれる）によって採点される。スピーキングテストと同様でライティングテストも採点者の主観や採点の偏りなどが大きな課題である。人間の採点だけではなく，文法や綴りなどの言語的な誤りを機械で採点するのはより主観性を排除した採点方法である。さらに，複数の採点者を用いる事により，採点の偏りや主観性を最小化し，レベル別基準文（Benchmark）と照合しながら採点をする方式を導入しているために，信頼性は大きく進歩している。

4.2.4　ライティング測定の実現性

ライティングテストは，２つのタスクに対して応答文をコンピュータに打ち込むために，技術的な困難点は少ない。スピーキングテストの様に，隣の受験者に集中力を乱されることは少ないだろう。但し，タッチタイピングなどやキーボードへの入力の技術は必要となるが，現代社会においては必須な一般的学習スキルであり，問題点は少ないだろう。

５．結論

これまでTOEFL iBTの採点資料やManual等を通読し，またTOEFL iBTのワークショップに参加した経験から主にスピーキングとライティングの妥当性，信頼性，実現性について論じてきた。本稿のまとめを行い，大学の教育現場への示唆を述べる。

最初に，信頼性については，スピーキングもライティングも採点者の検査（calibration）に合格した複数の採点者を用い，採点基準説明文等に沿って厳しい採点方式が用いられていることから信頼性が高いといえる。

第２にスピーキングの内容妥当性と表面妥当性については，タスクは大学生活での学生同士の会話等，キャンパス内での状況や講義の内容を題材として表面的に妥当性を持ち，受験生と関係者にアピール度は強い。概念妥当性については，コンピュータに受験生が一方向性で話し録音する形態であり，講義で予想されるグループ討論，講師や級友とのやりとりのコミュニケーションが基本的

には当てはまらない。

第３にライティングのタスクは，内容妥当性としてアカデミック・ライティングを全体的に網羅している訳ではないが，伝統的な論述的エッセイと講義でのノートテイキングを測定する基礎的かつあらゆる講義に共通する書く能力を測定しており充分ではないが，必修な内容であると言える。構成概念としては，アカデミック・ライティングの一部を形成する能力を測定しており限定的であると言える。

最後に，実現性の問題とTOEFL iBTの大学環境への導入の可能性を述べる。これまでTOEICや実用英語検定試験ではスピーキング力とライティング力を測定できなかったために，４技能を測定するTOEFL iBTの可能性について期待は非常に大きい。大学でのプレイスメントテストや卒業時の英語能力到達テスト，留学派遣の英語能力試験などについても活用の可能性は高い。

スピーキングとライティングに共通する課題は，多数のコンピュータ機器を要する事，採点者のトレーニングに専門的ノウハウと時間や経費を要する事，受験費用が高額である事，受験時間が長時間である事，試験センターが不足している事などが挙げられる。さらに，具体的な問題点としてスピーキングテストの時に隣の受験者のスピーキングに邪魔され，集中できない試験場環境の問題がある。

TOEFL iBTは大きな可能性を秘めた英語能力試験である。本稿で検証した妥当性，信頼性，実現性が大学の状況下で活用できる可能性を探る端緒となれば幸いである。

注

（１）日本人受験者のTOEFL PBT（677点満点）（1996-1997）の平均点は496点でアジア25か国中，23位。ちなみにタイ国受験者平均点は491点，モンゴル国受験者平均点は，490点，韓国平均点は518点，中国平均点は555点である。

（２）米国の学部への入学が許可されるTOEFL iBTのスコアは79点（PBT550点）が通常である。その点数を満たさない場合には語学学校で英語科目を受講することが入学への条件となる場合が多い。

（３）2011年現在のETS実施のTOEFL iBT（120点満点）によると日本人受験者の平点は69点で31か国中で30位である。ちなみにラオス国受験者平均点が68点で最下位，隣国の韓国受験者平均点は82点，中国受験者平均点は77点である。

（４）ヨーロッパでは，van EkやWilkinsなどにより1970年中盤からEC（欧州共同体）を中心として，広がる国籍を超えた労働者の英語能力を測定するために，Thresholdを作成し，ヨーロッパ諸言語に共通な測定方法を開発した。

（５）CEFRは低い英語レベルからＡ１，Ａ２，Ｂ１，Ｂ２，Ｃ１，Ｃ２の６レベルでなり，言語で行動できることを各レベルで示している。ちなみにＢ１レベルはTOEFL 56点から86点の範囲に該当し，日本人大学生が目指すべきレベルである。

（６）IELTSは得点ではなく，バンドスケールの０（非受験者）から９（エキスパートユーザー）で評価される。大学学部への入学が許可されるのは，バンドスケール6.0から6.5程度である。TOEFLがETSによって運営されているのに対して，IELTSはケンブリッジESOLやブリティッシュ・カウンシルなどに共同運営されている。所要時間は約３時間を要する。日本では受験会場が全国で５地域ほどである。受験料は２万５千円程度。試験内容は，TOEFL iBTと同様に４技能を測定する。リスニング（60分，40問）は問題用紙にメモ取りが許され，正確な情報の理解も求められ，解答選択肢が５つ以上の問題もある。リーディング（60分，40問）はGeneral分野では日常生活で読解力，Academicでは学校を想定した問題が出題される。ライティング（60分，２題出題）では，General分野では日常の手紙などを書く力，Academic分野では大学でのエッセイを書く能力が測定される。スピーキング（15分から20分）では，対面式で自己紹介に続いて最初の課題について自己の体験や課題を試験官に伝え，２番目の与えられた課題ではディスカションをし，話した内容は録音され，試験センターに送付される。

（７）項目応答理論（IRT）は，評価項目に応じて被験者の能力や，項目の難易度というパラメータにより出題を変える方式であるが，受験生間の公平な測定になっているかどうか議論がある。

（８）TOEFL IPTは２レベルを持ち，レベル１はPBTとほぼ同レベルで，レベル2 pre-TOEFLはTOEFL500点以下の学生を対象に実施される。

（９）採点基準説明（rubrics）は４技能についてそれぞれETSが作成した公式基準がある。
スピーキングは，独立スピーキングと統合スピーキングで個別に詳細に記載されている。例えば，独立スピーキングの３項目の一つである「話し方」(Delivery）で「１」の判定記述は以下である。「常に発音や強勢，イントネーションの困難点は常に聞き手の［聞き取る］努力を要し，話し方は細かく，途切れ，電報での伝え方［２，３語レベル］で，

頻繁にポースと躊躇［えーと，あのーなど］がある」［ ］内は筆者が解説。

(10) TOEFL PBTではリスニングとリーディングを測定，スピーキングはTSE，ライティングはTWEで測定された。

参考文献

青木昭六・田中正道.(1985).『伝達重視の英語教育』東京：大修館書店.

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

Canale, M. & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language and Teaching and Testing. *Applied Linguistics*, 1, 1-47.

Cottrell, S. (2003). *The Study Skills Handbook*. (2nd ed.) London: Macmillan.

ETS. (1997). *TOEFL Test and Section Score Data Summary 1997-98*. ETS.

ETS. (2012). *TOEFL iBT Total and Section Score Means. Educational Testing Service.* ETSホームページ.

ETS. (2012). *TOEFL Teacher Workshop Manual* (Revised). New Jersey: ETS.

ETS. (2012). *TOEFL Reading, Listening, Speaking & Writing Activities* (Revised). New Jersey: ETS.

ETS. *TOEFL Test Prep PLANNER*. ETSホームページ. アクセス2013年11月24日. http://www.ets.org/s/toefl/pdf/toefl_student_test_prep_planner.pdf#search='TOEFL TEst+Prep+Planner' pp.11-13, 14-17, 21-22, 26-27.

文部科学省.(2013). 高大連携に関する意見. アクセス2013年11月23日. http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo12/shiryo/attach/1326461.htm

Hughes, A. (2003). *Testing for Language Teachers* 2nd ed. Cambridge: Cambridge University Press.

Oller, J.W. (1983). *Issues in Language Testing Research* (eds.). Rowley: Newbury House Publishing Co.

Rivers, M.W. (1968). *Teaching Foreign Language Skills*. Chicago: The University of Chicago Press.

Sharpe, J.P. (2010). *TOEFL iBT* (13th ed.). New York: Barron's.

Weir, C. J. (1990) *Communicative Language Testing* Cambridge: Cambridge University Press.

資料

サンプル基準文 レベル１（出典: TOEFL Teacher Workshop Manual, 2010 p.48)

In this lecture, the examples only one of the group succeed the project. Why the group will succeed on this project it is because of few factor.

First of all, a group of people has a wider range of knowledge, expertise, and skills than any single individual is like to process, and easier to gather the information and resources to make the work effectively, and the group will willingly to trey something is risky decision to make the project for interesting and successful, it is because all the member of group carries the different responsibility for a decision, so once the decision turn wrong, no a any individual one will be blame for the whole responsibility.

On the other way, the groups which are fail the project is because they lay on some more influence people in the group, so even the idea is come out. Once the influenced people say that is no good, then process of the idea will be drop down immediately instead taking some more further discussion! So the idea will not be easy to settle down for a group.

The form of the group is very important, and each of the members should be respect another and try out all the idea others had suggested, then it will develop a huge idea and the cooperate work environment for each other effectively work!

注釈（ETS）

この基準文１応答で使われたレベルの言語は非常に低く，第２パラグラフは最もレベルが低く，講義への唯一の参照である。読み手はパラグラフからの意味を掴み取るのは困難であり，応答は一貫した情報に貢献していないので，１と採点された。

Examination of the Validity, Reliability and Practicality of the Speaking and Writing Test of TOEFL iBT

TOKESHI Masanori

Abstract

This paper will examine the validity, reliability and practicality of the speaking test and writing test of TOEFL iBT on the basis of the writer's experience as both an examinee and a scorer in the workshop, including reference to training materials. The results revealed the following. Firstly, considerable efforts are made to enhance the reliability of scoring on both the speaking and writing tests. Secondly, regarding the validity of the speaking test, face validity and content validity are strong, but construct validity is weak. Thirdly, the writing test requires an integrative task using multiple language skills, but it ends up with traditional argumentative writing and fundamental writing skills, such as holistically summarizing the main points of lectures and written references. Lastly, the practicality of TOEFL iBT reveals appealing strengths, but it also contains several weaknesses, such as facilities, the testing fee, and scorer training.

Keywords: foreign language, assessment, validity, reliability, TOEFL