科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２９　日現在

機関番号：３２６１２
研究種目：挑戦的萌芽研究
研究期間：2012～2013
課題番号：２４６５２１３１

研究課題名（和文）英語スピーキング伝達能力客観テスト

研究課題名（英文）An Objective Test of Communicative English Proficiency

研究代表者
バティ　アーロン（Batty, Aaron Olaf）

慶應義塾大学・環境情報学部・講師

研究者番号：８０４０６６８６

交付決定額（研究期間全体）：（直接経費）　　1,800,000円、（間接経費）　　540,000円

研究成果の概要（和文）：研究者たちは新しく「スピーキング伝達能力客観テスト（OCST）」を開発した。OCSTはタブレットPCを用いた計時情報ギャップ型テストである。伝統的な口頭運用能力テストの構成要素がテスト所要時間の原因になる事を前提に、発話者が評定者に新しい情報を述べる時間を計る。英語が第一（L1）及び第二言語（Ｌ２）である86名を対象にテストを行い、L2タスク完了時間にL1基準のスコアを当て、データは多相ラッシュ・モデルで解析された。仮説通りテストの客観的デザインは評定者の影響を弱め、評定者をモデルから除外できた。受験者の信頼性係数として0.88が観測され、多くの主観的なスピーキング能力テストの数値を上回った。

研究成果の概要（英文）：The researchers developed a new test of communicative speaking proficiency, called the Objective Communicative Speaking Test (OCST). The OCST is a timed information-gap task-based test delivered via tablet computers. The OCST measures the time required for a speaker to relate a piece of information unknown to the rater, on the assumption that more traditional components of oral proficiency will contribute to time to completion. The test was administered to a sample of 86 first- (L1s) and second-language (L2s) speakers of English, and their task completion times were assigned an L1-referenced score. The data were analyzed via many-facet Rasch analysis. As hypothesized, the objective design of the test reduced rater effects, and raters could be excluded from the model. An examinee reliability coefficient of 0.88 was observed, surpassing that of most subjective tests of speaking proficiency.

研究分野：人文学

科研費の分科・細目：言語学・外国語教育

キーワード：学力検査　言語試験　英語教育　スピーキング伝達能力

１．研究開始当初の背景

一般的に第二言語のスピーキング能力テストは評定者によって評価されてきた。このようなスピーキング能力テストは、全体的に厳格性や寛容性の違い、ハロー効果、そして得点範囲の縮小または中心化傾向を含め、評定者の変動的な判断により不公平性に関する多くの問題を引き起こす可能性がある（具体的な概要については Linacre, 1989, and Saal, Downey, & Lahey, 1980 を参照）。これらの問題は根本的に主観的な性質を持つ評価方法に原因があると考えられる。多くの場合、採点基準の主観的な性質に加え評価者バイアスがかかる恐れがある。例えば、受験者の母語もしくは文化的背景に伴う評定者の熟知度がバイアスに影響する懸念がある (Carey, Mannell, & Dunn, 2011; Huang, 2013; Kim, 2009; Winke, Gass, & Myford, 2013)。上記に示した「静的」なバイアスの恐れ以外にも多くのスピーキング能力テストにおいて時間経過と共に評価が変動するという評定者ドリフトの問題があり、たいてい評定者の疲労感によって引き起こされる(Wilson & Case, 2000; Wolfe, Moulder, & Myford, 2001)。以上、これらの問題によって伝統的な外国語スピーキング能力テストの得点は信頼性に欠けるものがある。

主観的に評価する評定者を採用した能力テストの心理統計的な信頼性に関する懸念だけでなく、社会的公正に関する問題もある。十年以上にわたって不公平なテストの実施における結果と社会公正的な立場からテストの妥当性を解釈することに焦点が当てられており、言語テストにおける社会的な側面はますます注目されている(McNamara & Ryan, 2011; McNamara, 1998, 2001, 2006 を参照)。主観的に評価する評定者を採用したテストにおいて、無意識的あるいは意識的であれ、いずれにせよ評定者のバイアスが得点に影響する恐れがある。

そこで、本研究ではスピーキング伝達能力の評価方法に対する代替的なアプローチを提案する。このアプローチとはスコア（得点）が対話者への情報伝達における正当性と速力性によって評価される「客観的」な伝達能力テストであり、上記に示した多くの問題を軽減する可能性を持つ。

２．研究の目的

伝統的な外国語スピーキング能力テストでは受験者のスピーキング能力を主観的に評価する評定者を採用してきた。しかし、それは多くの場合テストの信頼性を低下させる。この課題に対処するため、研究者たちは Objective Communicative Speaking Test (スピーキング伝達能力客観テストまたは OCST)と呼ばれるスピーキング伝達能力における新しいテストを開発した。OCST とは、計時する情報ギャップタスク型テストである。対にされたタブレット PC を通して実施され、伝統的な口頭運用能力とは対照的に、スピーキング伝達能力に焦点を当てる。伝統的な口頭運用能力テストにおける構成要素がテスト終了時までに必要とする時間の原因になるという前提をもとにテストを行う。スピーキング能力をいくつかの評価尺度(例えば、発音、流暢さ、文法、語彙）によって主観的に評価するのではなく、寧ろ OCST は発話者が評定者にとって未知である新しい情報を評定者に伝えるまでに必要とする時間を測定する。

３．研究の方法

本研究ではテストの配布とデータ収集を容易にするために、アップル社のタブレット型 PC である iPad 向けのウェブアプリケーション開発を依頼した。このアプリケーションは PHP、JavaScript そして node.js を用いて、受験者のタブレットへ問題のコンテンツと評定者のタブレットへ解答のコンテンツを配布する。待ち時間を補正するために異なる 2 台のウェブクライアントが同期された。また、項目間のスムーズな移行を確実にするためコンテンツのバックグラウンドを読み込むプッシュサービスを用いた。各評定者は、評定者のクライアントを受験者のクライアントに繋ぎ、「通信路」を作成する。これは複数の評定者が他のテストセッションによるクロストークの影響を受けることなく、同時に試験を実施することを可能にする。

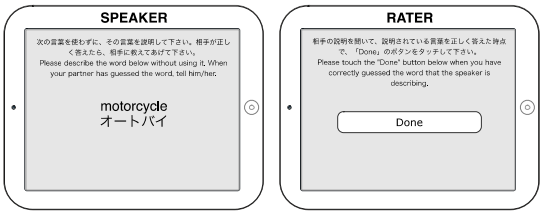
しかしながら、評定者は配布される項目の単純なパターンに気づき、受験者の発言に耳を傾けることなく解答を予測する可能性が懸念されたため、複雑な疑似ランダム項目配布系列を開発した。この項目配布系列により 54 回のテストセッションにわたって各項目がそれぞれ同回数ずつ配布されることが可能となり、各項目が必ず 4 回のテストセッションのうち 1 回配布されることを確実にする。この項目配布系列を使用することによって受験者は同じ項目を繰り返すことなく、連続的に異なる評定者と 4 回テストを受けることが可能となる。一方で、評定者は次に配布される項目を予測することができない。

テストセッションが始まる際に、評定者は通信路の名前、疑似ランダム系列から使用されるセッション開始番号、受験者の名前もしくは識別番号を入力する。後続するセッションでは、受験者の名前のみ入力する。通信路の名前はセッション間で持続され、セッション番号は自動的に進行する。項目が完了すると、その解答、その解答に対する二値型得点、そして解答完了時間がウェブサーバー上の MySQL のデータベースに記録され、次の項

目が開始する。テストセッションが終了すると評定者は入力インターフェースに戻され、項目配布系列から次の項目セットが待ち行列に加えられる。

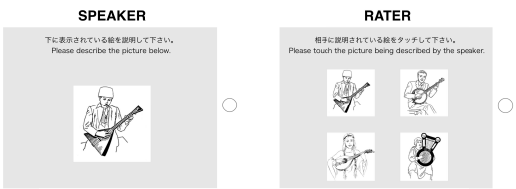
本研究では「言葉」・「絵」そして「アカデミック説明」の 3 種類のタスクがデザインされた。まず、「言葉」のタスクタイプはアメリカ合衆国のパーティーゲーム“ Taboo（タブー）”と同様である。発話者の iPad に言葉が表示され、発話者はその言葉を使用せずに表示された言葉を聴き手に説明する。発話者が説明している言葉を聴き手が正しく答えた時点でタスクは完了し、時間が記録される。このタスクの難易度は初級として作成されている（例：<図 1>）。

<図 1　例：「言葉」のタスク>



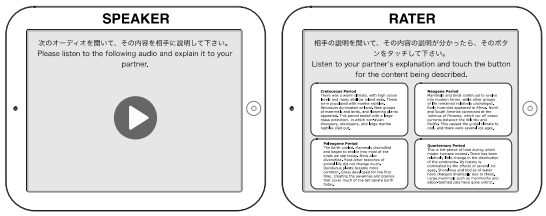
次に、「絵」のタスクでは発話者の iPad に絵が表示され発話者は聴き手に絵の説明を行う。聴き手は持ち手の iPad に表示された 4 つの絵の中から、説明と一致する絵を選択する。聴き手が絵を選択した時点でタスクは完了し、スコアが与えられ、時間が記録される。このタスクの難易度は中級として作成されている（例：<図 2>）。

<図 2　例：「絵」のタスク>



最後に、「アカデミック説明」のタスクではヘッドフォンを通して発話者が自身の母語で興味・関心のある一分間の短いレクチャーを聞く。その後、発話者は聴き手に向けて自身が聞き取った文の主なポイントを説明する。聴き手は発話者の説明を聞いた上で持ち手の iPad に表示された 4 つの概要文から一つ選択する。選択した時点でタスクは完了し、解答にスコアが与えられ、時間が記録される。このタスクはコンテクストを共有していない他者に向けて視聴した詳細な文章の内容を説明する能力が要求されるため、難易度が最も高いものとして作成されている（例: <図 3>）。

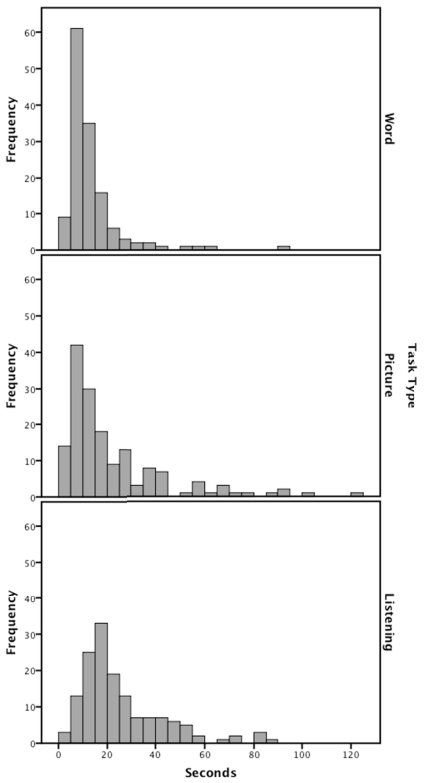
<図 3　例：「リスニング」のタスク>



以上 3 種類のタスクタイプには、発話者が自身で得た情報を自発語で正解を未だ知らない聴き手に伝える能力が必要とされている。発話者が重要な点を素早く、効果的に伝えることができなければ聴き手は正当することができない。これらのことから、本テストは面接試験と異なる。また、時間的要素を付け加えたことで評定者の評価方法に存在する可能性がある多くの問題を解決することができる。

次に、受験者のタスク完了時間のスコアを判定する参考基準を作成するために、英語を第一言語とする 43 人（L1）を対象にテストを実施した。各タスクタイプの完了時間の分布を測定するために、L１のデータを分析した。以下の図 4 に見られるように、各データの分布は左に大きく偏り、極端な外れ値も確認される。しかしながら、L1 のデータは比較的狭い範囲にまとまってあり、単峰型の分布であることが分かる。ここでは参考基準を設定することを目的とし、外れ値の影響を弱めるためデータを変換する必要性があった。

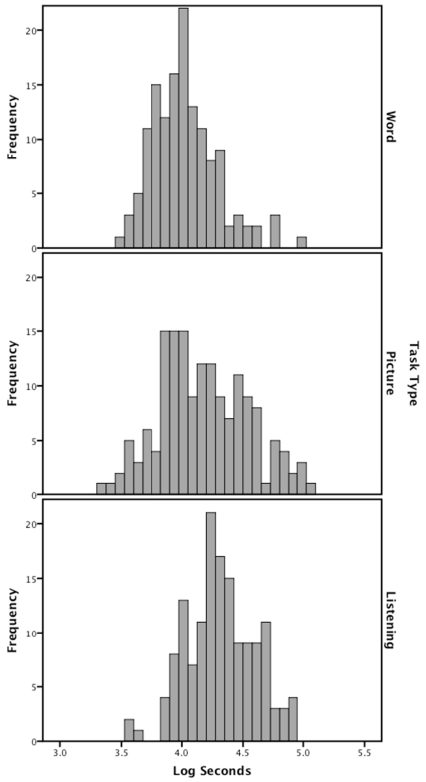
<図４　L１のタスク完了時間の分布>



L1 のタスク完了時間の歪度を調整するため、底を 10 とする対数変換を行ったところ各タスクタイプにとってより正規分布モデルを適用するのにふさわしい形となった（<図5>）。そして、尺度を設定するために L１対数時間は z スコアに標準化された。L1 の平均値を最高スコアとし値が平均値よりも遅くなるにつれて 4SD にわたってスコアが低くなる、分かりやすい 6 段階の評定尺度を作成した。

この評定尺度を TimeScore 尺度と呼び、表 1 に詳細を示す。

<図５　対数変換後の L1 のタスク完了時間の分布>



<表１　TimeScore 評定尺度>

スコア	範囲	説明
5	≤0 標準偏差	L1 の平均に等しい又は平均より速い
4	>0 - 1 標準偏差	L1 の３４パーセンタイルに等しい又はより遅い
3	>1 - 2 標準偏差	L1 の１４パーセンタイルに等しい又はより遅い
2	>2 - 3 標準偏差	L１の９８%より遅い
1	>3 - 4 標準偏差	L１の９９．９%より遅い
0	>4 標準偏差	L１の９９．９９７%より遅い

次に、第二言語として英語を話す（L２）43 名を対象にテストが実施され、L2 受験者のタスク完了時間に L１のタスク完了時間を参考基準としたスコアが割り当てられた。能力推定値の信頼性における様々な要因の影響を測定するため、Facets という多相ラッシュ・モデルソフトウェア（Linacre, 2012）を用いて解析を行った。

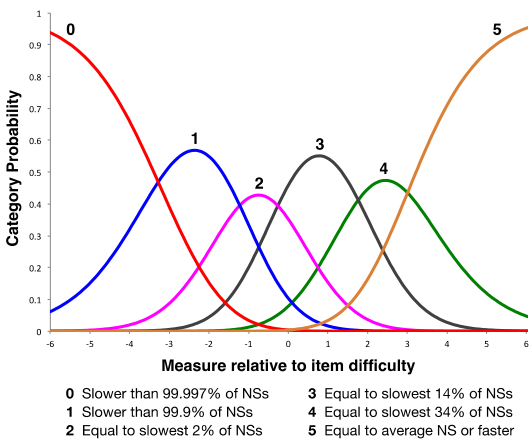
４．研究成果

期待通り、テストの客観的なデザインは評定者による影響を減少させ評定者をモデルから除外することができた。このために、評定者たちを多相ラッシュ・モデルから除去することができ主観的なスコア変動の最大要因を取り除くことができた。

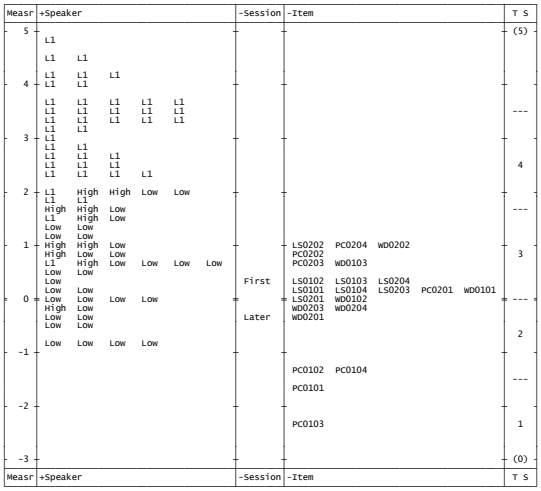
また、受験者がテストを受けた第一回目のセッション後に項目の困難度推定値が減少するという有意な練習効果を確認することができた。受験者、セッション（第一回目またその後、合計４回にわたって行ったもの）、項目を含む 3 相のラッシュ・モデルは信頼性の高い結果を出すことが分かった。

図 6 に示されるように TimeScore 尺度はタスク完了時間を分かりやすい 6 段階に分け、それぞれの段階は最頻値を持つ。また、図７に示されるように、OCST は確実に L２から L１を区別し、L１の平均値における 1 標準偏差範囲内で L２の能力推定値が下がることはなかった。受験者の信頼性係数として 0.88 が観測され、多くのスピーキング能力における主観的なテストの値を上回った。図 7 は L1 の参考基準と L2 発話者のサンプルを比較したものである。受験者は自身が持つ TOEFL スコアで表されている。TimeScore 尺度は“TS”、言葉のタスク項目は“WD”、絵のタスク項目は“PC”、そしてリスニングのタスク項目は“LS”として示されている。

<図６　TimeScore 尺度のカテゴリー確率曲線>



<図７　グラフ化した Facets 分析の結果>



本研究は現時点で OCST **がコミュニケーション伝達能力を客観的に測る信頼性を明らかにしているが、さらに研究を進めて行く必要性がある。テストは全体能力の面から解答者を確実に分類できたが、一般化可能性を高めるには多様な受験者を募ることができる、より大きな組織が必要である。さらに、**OCST **と受験者を様々な能力カテゴリー（例えば、発音、語彙、文法など）から評価する伝統的なスピーキング能力テストの比較研究は、**OCST **の伝達能力の構成概念の位置づけを促進させ、その他の** L2 **スピーキング能力測定方法との比較も容易にするであろう。**

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　１件）

① Batty, A. O., & Stewart, J. (in press). Theoretical underpinnings of a computerized objective test of communicative competence in Japan. In V. Aryadoust & J. Fox (Eds.), *Current Trends in Language Testing in the Pacific Rim and the Middle East: Policies, Analyses, and Diagnoses.* Cambridge Scholars Publishing.

〔学会発表〕（計　２件）

① Batty, A. O., & Stewart, J. (2014/06/06). A proposal for a socially-fair and objective method of scoring communicative ability. Presented at the Language Testing Research Colloquium (LTRC) 2014, Amsterdam, Netherlands.

② Batty, A. O., & Stewart, J. (2013/08/03). A more objective test of communicative competence in English: Establishing native speaker norms with MFRM. Presented at the Pacific Rim Objective Measurement Symposium (PROMS) 2013, Kaohsiung, Taiwan.

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
https://github.com/aaronpropst/cst
テストのソフトウェアのソースコードは GitHub **で配布している。オープンソースである。**

６．研究組織
(1)研究代表者

バティ　アーロン（Aaron Olaf Batty）
慶應義塾大学・環境情報学部・講師
研究者番号：80406686

(2)研究分担者

スチュワート　ジェフリー（Jeffrey Stewart）
九州産業大学・語学教育研究センター・講師
研究者番号：40536306