様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １７ 日現在

機関番号：１２１０２
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５８０１２２

研究課題名（和文）英文読解におけるタスクとエクササイズの違い：タスク型英文読解テストの開発

研究課題名（英文）Examining the Feature Differences Between Task and Exercise in Teaching EFL Reading: Toward the Development of Task-Based Reading Test

研究代表者
卯城　祐司（USHIRO, Yuji）

筑波大学・人文社会系・教授

研究者番号：６０２７１７２２

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究はタスク型英文読解テストの開発に向けて，テストの妥当化を行う前提条件であるテスト得点の信頼性を検証した。6種類のタスクを日本人大学生122名に与え，その解答をタスク依存・タスク独立尺度の2種類で採点した。一般化可能性研究の結果，両尺度は平行して用いることができるものの，採点手順の質には違いが見られた。決定研究の結果，タスク独立尺度を用いて，7種のタスクを3名で採点することが，評価の信頼性を確保する上で重要であることが分かった。

研究成果の概要（英文）：The present study described the first step in the development of a task-based reading test. We designed 6 tasks in which test-takers were required to transfer what they comprehended from passages. To validate extrapolations on the construct validity of the test, we examined the reliability of the test scores by a generalizability theory study. 122 Japanese university students completed the 6 tasks. Their outcomes were scored by 6 raters using either a task-dependent or task-independent rating scale. G-study showed that the 2 types of rating scale could be used alternatively, but qualitative analysis revealed that the 2 rating procedures differed in scoring of local errors associated with detailed information, appropriate reorganization of passage contents, and sociolinguistic appropriateness. D-study showed that 7 tasks by 3 raters are desirable using the task-independent rating scale in order to obtain a high-reliability coefficient (.80).

研究分野：英語教育学

キーワード：英語教育　リーディング　テスト　タスク中心教授法　妥当性　信頼性　測定　一般化可能性理論

１．研究開始当初の背景

教育現場で行われる英語力の測定は，学習指導要領に基づき，生徒が指導内容をどれだけ身につけ，授業の目標がどの程度達成されたのかを調べ，教育の改善へ活用することを目的としている。最新の学習指導要領 (文部科学省, 2008, 2009) では，外国語科の目標として「外国語で発信された情報や考えを的確に理解し，その内容に基づいた適切なコミュニケーションを行う姿勢や能力の育成」が掲げられており，旧来の英語学力テストから脱却した新しいテストの開発が求められている。

英語のコミュニケーション能力を測定するために，McNamara (1996) は実社会に即したタスクを行う過程で得られた言語データを利用することを提案している。しかし，具体的なコミュニケーション場面を想定しやすいスピーキングやリスニング，ライティングに比べ，リーディング能力の測定では，英文を読むことによって解決するタスクや場面の設定が難しく，従来型の問題にとどまることが多い。

従来の英文読解テストとしては，1 つの英文に様々な観点 (文法・語法・和訳など) を盛り込んだ総合形式や, 英検・TOEIC・TOEFL などが採用する多肢選択式がある。しかし，従来型のテストでは，「英文全体は読まずに解答に必要な部分にだけ目を通す」という，本来の「読む」という行為とはかけ離れた能力をも測定してしまう可能性がある (靜, 2002)。このようなテストでは「コミュニケーションのための言語使用」という観点が欠如しており，求めている能力を正しく測定できないという問題がある。さらに，従来型のテストは教室での読解指導の在り方をも歪めてしまう可能性がある。すなわち，テストのために文法・語法・和訳などを練習 (エクササイズ) させる指導が中心に行われ，英文から理解した内容を誰かに伝えるというコミュニケーション能力の育成が伴わない恐れがある。

以上の問題点を解決するために，本研究では「コミュニケーションのための言語使用」という観点を英文読解テストにも反映させるべくタスク中心教授法 (e.g., Ellis, 2003) に基づく英文読解テストを開発することとした。

２．研究の目的

本研究の第一の目的は，タスク型英文読解テストの妥当化の第一歩として，構成概念妥当性の推論に必要な統計分析を妥当なものとするため，評価の信頼性を検証することとした。二つ目の目的として，評価の信頼性を揺るがすタスクの特徴，評価者の特徴，評価尺度の特徴を明らかにし，タスクに基づく読解パフォーマンスを評価するための尺度を提案する。具体的には次のリサーチクエスチョン (RQs) を検討した。

RQ1: タスクに基づく言語パフォーマンスの評価方法として提案されている 2 種類の評価尺度（タスク依存型・タスク独立型）にはどの程度平行性があるか。

RQ2: 2 種類の評価尺度の間に，評価者間信頼性の違いはどの程度生じるか。

RQ3: 評価尺度の違いによってタスクや評価者による評価特徴はどの程度異なるか。

RQ4: 信頼性のある評価を行うのに評価者やタスクの数はどの程度必要か。

３．研究の方法

(1) マテリアル

本研究で作成したテストは「テキストの内容を誰かに伝えるために読み，かつ理解した内容をメモやメールといった形式で適切に伝えられるか」という情報転移課題である。このタスクが測定するのは産出された言語そのものであるため，英語で産出させた場合には，そのタスクが読解能力を反映しているのか，英語での産出能力を測定しているのかという議論がある (Horiba & Fukaya, 2012; Jenkins et al., 2004; Trites & McGroarty, 2005)。言語産出から内容理解を評価する際には，英語での産出能力が十分でないために理解内容が反映されないというケースを避ける必要がある。そのため，本研究では，英語でテキストの内容を読解し，日本語でそれを相手に伝える形式を採用した。

テスト作成の手順として，TOEIC (Bridge) の公式問題集から (a)「テキストを読んで内容を相手に伝える」という目的が達成できるもの，(b) 設定可能な読みの目的が実生活で起こり得るもの，(c) 英文の難易度が受験者にとって適切だと想定されるものを収集した。収集したマテリアルを改編し，情報転移の指示文 (e.g., あなたはロンドンの学校で日本語を教えている教員です。ある日，職員室で以下の掲示を目にしました。出張で職員室にいない日本人の同僚に，このことを伝えたいと思います。必要だと思われる情報を過不足なく含め，掲示の内容を同僚に伝えるためにパソコンでメールを作成して下さい) を加えた。

(2) 実施手順

日本人大学生 122 名を対象に，6 種類のタスクを 90 分の授業の中で 2 つずつ, 3 回のセッションに分けて行った。1 セッションはタスク A → 質問紙 A，タスク B → 質問紙 B という順で行われた。今回のタスクは通常授業の一環として行われたが，データの使用方法については十分に説明し informed consent を得た。協力者は本研究のタスクがテストであることは伝えられていない。

セッションでは，まずタスクと解答用紙が印刷された冊子を配布した。タスクに書かれ

ている問題内容を全員で確認し，タスクで求められていることを周知した (e.g., 友人に○○を伝えるためのメールを作らなければならない)。その後，タスクを完成させ，タスクに関するアンケートを行った。最後にペアでタスクの完成度の評価，全体で英文内容の理解度の確認を行った。

読解パフォーマンスの採点基準として，Ellis (2003)，Hudson (2005) や Norris et al. (2002) に従い 5 件法 (タスクの完成度に応じて 1. *inadequate* ~ 5. *adept*) のタスク依存尺度とタスク独立尺度を作成した。各尺度はタスクの完成度を評価するための基準とし，1~2 では，受験者はタスクで求められたコミュニケーションの目的を達成できておらず，3~5 では完成度に差はあるものの受験者はタスクで求められたコミュニケーションの目的を達成できたという基準になっている。

(3) 分析方法

RQ1 の検証のために各タスクの評定値の平均が 2 つの評価尺度で異なるかどうかを分散分析により確認した。さらに Shimizu (2009) に倣い，評定の分散が評価尺度間で等しいことを確認するために F test を行った。本研究はサンプルサイズが比較的大きいため，平均値の比較をする際には効果量を中心に結果を解釈した (水本・竹内, 2008)。

RQ2 の検証では，各タスク×評価尺度の違いにより，評価者間信頼性がどの程度変化するのかを検討するために Cronbach's α を算出した。また，評価者間信頼性が低い場合，どの評定者が入らなければどの程度 Cronbach's α が改善されるのかを検討した。その際，信頼性の低かった評定者がどのような採点を行っていたのかを評定者間の討論の記録から質的に検討した。

RQ3 および RQ4 について，本研究の G-study は，「評価者」と「タスク」という 2 つの facets を設定し，測定の対象を「協力者」とした 2 相完全クロス計画とした。また，評価尺度の違いによる評価特徴の差異に注目するため，タスク依存尺度とタスク独立尺度の比較を行いながら分散成分の推定値の検討を行った。具体的には，各評価尺度で採点された学生のタスク・パフォーマンスの評価得点を従属変数として，評定者とタスクの種類の組み合わせによる分散成分を特定した。続く D-study では，分散成分の推定値を用いた G 係数のシミュレーションを行った。

４．研究成果

本研究はタスク型英文読解テスト開発の最初のステップとして，作成したテストの評価方法がどの程度信頼性をもって行われるのかを検証した。タスクに基づく言語パフォーマンス研究で提案されている 2 種類の評価尺度について検討した結果，(a) 両尺度ともある程度平行性をもって使用できること，(b) 採点の評価者間信頼性は，詳細情報，情報の再構成，社会言語学的な要因への考慮により変化し，(c) タスク依存尺度の場合に信頼性が低下する傾向にあること，(d) 同じ種類のタスクを 7 つ以上使用し，かつ 3 名以上の評価者がタスク独立尺度を使って採点することが評価の上で望ましいことが明らかとなった。

RQ1 に対応する結果 (a) について具体的に見ると，タスク依存尺度とタスク独立尺度のそれぞれで学習者の読解パフォーマンスを採点しても，平均点およびその分散に統計的な差は現れないことが分かった。さらに両尺度間には強い相関関係 (r = .92, n = 122) があったため，それぞれの尺度は採点時に並行して使用できることが示された。

ただし RQ2 に対応する結果 (b) について，採点者の評価プロセスを検討すると，採点に使用する尺度間で質的な違いが見られた。特に，タスク依存尺度を使用すると，学習者の解答プロトコルに (a) 詳細情報をどのように提示するか，(b) 相手に伝えるべき情報をどのような順番で提示するか，および (c) コミュニケーション相手の社会的立場を考慮しているか，という 3 点が含まれている場合に，評価者の採点が一致しにくくなることが分かった。これらの結果から，タスクの特徴に応じた採点基準を個別に設ける際には，複数の評価者間で採点方法や基準の話し合いをしたり，あらかじめ評価トレーニングを行ったりすることの必要性が示唆された。

RQ3 および RQ4 に対応する結果 (c) について，シミュレーションの詳細な結果を図 1 と図 2 に示す。これらの図から分かる通り，評価者数を増やしたとしても信頼性係数が大幅に向上することは無く，タスク数を増やすことが重要となる。テストの実用性から考えても，評価者数の確保よりも，取り組ませるタスクの数を増やすほうが望ましいと言える。

具体的に見ると，十分に高い信頼性係数 (.80) を得るためには評価者数が 3 名の場合はタスク数が 7 つで済むのに対し，タスク数が 3 つの場合は 25 名もの評価者が必要になるという結果になった。

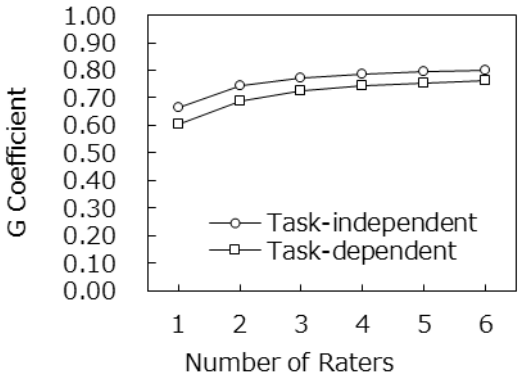

図 1. 3 つのタスクに対し評価者数を 1 名から 6 名に変えたときの G 係数の変化.

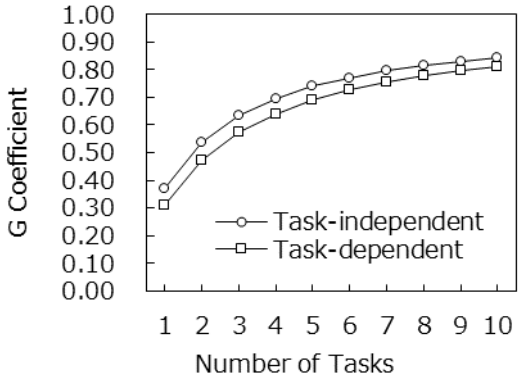

図 2. 評価者 3 名に対しタスク数を 1 項目から 10 項目に変えたときの G 係数の変化.

また，タスク依存尺度とタスク独立尺度の間で大きな差は無く，評価者間信頼性という観点では両尺度は並行して使用可能であると言える。

以上の結果を踏まえると，タスク型英文読解テストの採点においてはタスク独立尺度の方が，実行可能性が高いと言える。一方，十分な信頼性を得るためには，この尺度を用いたとしても，7 つ程度のタスクが必要であった。評価者の数を増やすよりは実効性が高いと考えられるものの，教室環境において，この数のタスクを同時に実施することは難しい。そのため，タスク型英文読解テストを行う際には，数回の実施を想定して，事前に授業計画にテスト実施のスケジュールを織り込み，学習者にあらかじめ周知しておくことが望ましいと考えられる。

同時に，採点者数の増減による影響は比較的小さかったことから，採点基準についてトレーニングを受けた，または基準に関する合意を得た教師を数人程度確保することが大切だと考えられる。具体的には，評価者間信頼性を脅かしうる各種変数（受験者の社会言語学的能力，方略能力，情報を要約する力など）に対して，各々の採点基準を事前に協議しておくことが望ましい。さらに，完成度が高いと判断される模範解答を事前に作成し，評価者間で合意を得ておくことも重要である。

外国語の習得においては，学習項目の体系的な暗記にばかり偏重せず，英語でできることを拡張していくことが必要であり，教師は生徒のどのような能力を育てたいかを明確に想定しておく必要がある。タスク型読解テストの開発が進むことで，以上のような点で外国語教師や学習者への正の波及効果も生じると考えられる。

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 Ushiro, Y., Hamada, A., Hasegawa, Y., Dowse, E., Tanaka, N., Suzuki, K., Hosoda, M., & Mori, Y. (2015). A generalizability theory study on the assessment of task-based reading performance. *JLTA Journal*, *18*, 92–114. Retrieved from http://ci.nii.ac.jp/naid/110010000679/en

〔学会発表〕（計２件）

①卯城祐司. (2016, 3 月 12 日).「状況モデル構築を目指した中高の英語リーディング指導」. 日本英語教育学会第 47 回年次研究集会招待講演. 早稲田大学（東京都新宿区）

② Ushiro, Y. (2015, 7 月 2 日). Understanding how students learn English in Japan from elementary school to university: Looking ahead at how to address globalized needs and environments. 岡山大学言語教育センター特別講演会. 岡山大学（岡山県岡山市）

〔図書〕（計１件）

① 卯城祐司. (編著). (2014).『英語で教える英文法 : 場面で導入、活動で理解』. 東京: 研究社. <238 頁>

〔その他〕
ホームページ等
http://www.u.tsukuba.ac.jp/~ushiro.yuji.gn/

６．研究組織
(1)研究代表者
卯城　祐司（USHIRO, Yuji）
筑波大学・人文社会系・教授
研究者番号：６０２７１７２２