

平成 21 年 6 月 7 日現在

研究種目：基盤研究（C）

研究期間：2006～2008

課題番号：18520423

研究課題名（和文）中学・高校の日本人教師のための信頼性・妥当性の高いライティング  
の評価規準作成研究課題名（英文）The Construction of Reliable and Valid Criteria for Writing Tests  
for Japanese Teachers of English at Junior and Senior High Schools

研究代表者

望月 昭彦 (MOCHIZUKI AKIHIKO)

大東文化大学・外国語学部・教授

研究者番号：30219969

研究成果の概要：中学・高校の日本人教師のための信頼性・妥当性の高いライティングの評価規準作成を目的として、高校、高専7校、中学校3校を対象として英語熟達度テストとライティング・テストを実施した。分析の結果、(1)ライティング・テストの評価者間信頼性は高く、熟達度テストとの併存的妥当性は高いこと、(2)熟達度を見るには文の数よりも語数を見るほうが有効であること、(3)複雑さの指標と熟達度の相関は弱いこと等がわかった。

(金額単位：円)

	直接経費	間接経費	合計
2006年度	1,400,000	0	1,400,000
2007年度	1,300,000	390,000	1,690,000
2008年度	700,000	210,000	910,000
年度			
総計	3,400,000	600,000	4,000,000

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：(1) 英語教育 (2) 評価論 (3) ライティング・テスト (4) 評価規準  
(5) タスク実現度1 (6) タスク実現度2 (7) 正確さ (8) 流暢さ (9) 複雑さ

## 1. 研究開始当初の背景

(1) 1998年、文科省告示の新学習指導要領でコミュニケーション能力の育成を引き続き重視した結果、コミュニケーション・ランゲージ・ティーチングを意識的に取り入れようという努力がなされるようになった。その結果、オーラルの面では日本の中学生、高校生が以前より上達してきたと言われていたが、書く能力については全くと言ってよいほど英文が書けないという意見が現場の教師から出されていた。(2) まとまった談話レベル

のライティングをさせる TOEFL の Test of Written Test (TWE) の 2004 TWE 報告書 ETS s 5 版によると、2001 3 年の成績について 1 から 6 までの 6 段階の平均点で、日本と同じ EFL の韓国は 3.6 点、タイは 3.2 点、中国は 3.8 点に対して日本は 3.1 点で東南アジアで最下位であった。(3) ライティングといっても、単文レベルの作文が中学校、高校では大半である(宮田 2001)し、また、談話レベルのライティングについては日本人教師はその指導法がわからず、採点に時間がが

かるので生徒に書かせないという実態があった(宮田 2001)。たとえ、数少ない日本人教師が生徒に談話レベルのライティングを書かせても年間常駐または不定期に訪問するALTに採点を一任しているので、総合的評価、分析的評価方法をどのようにするべきかについては全く又は殆どわからない状態であった。(4)採点到時間のかかるライティングはプロの教育機関に任せてしまえばよいという考えの基に上述のTWEを、中学校または高校で生徒にクラス単位または学校単位で契約して受けさせる形式をとっている学校もあるが、日本人教師がそれにより評価者訓練を受けて上達するわけではない。また、生徒にライティングをさせた後、その答案を米国の大学のライティング・センターに送り、謝金を払って採点して返してもらっている学校もあった。このALT任せ、ETS、米国の大学のライティング・センター任せは日本の英語教育にとって大変残念な選択であった。

## 2. 研究の目的

(1) 中学・高校の日本人教師のために、日本人中学生・高校生の談話レベルのライティング能力を測定する信頼性が高く妥当性のある複数のタスクからなるライティング・テストを開発すること

(2) 中学校及び高校の、日本人英語教師が評価するために使う総合的評価規準及び、分析的評価規準を作成し、既に作成したライティング・テストを実施した後で、作成した評価規準に従い、評価を行い、それらの評価者間信頼性が高いものなるようにすること

## 3. 研究の方法

(1) 1年目 文献でライティングの評価方法について先行研究の検討、理論的研究について、2ヶ月に1回の割合で5, 7, 8, 10, 12, 2月の計6回、筑波大学において分担研究者、代表者4人で、各々、勉強してきたことを印刷物にまとめて研究発表会を実施した。予備実験実施について、54項目の45分間の熟達度テストを茨城県の公立高校2年生1クラス、茨城県公立中学校2年生1クラスを対象にして、又、3つのタスクからなる45分間のライティング・テスト(WTと略、中学生高校生共通の同一問題)を北海道公立高校2年生1クラス、茨城県公立中学校3年生1クラスを対象に、3月初旬に実施した。分析の結果、54項目の熟達度テストの信頼性係数( )は公立高校の場合、0.79であり、公立中学校の場合、0.94であることがわかり、

高い信頼性であったのでこれらのテストを2年目の本番のテストで使うことに決めた。また、WTについて、文レベルの作文能力を測定する3問からなるタスク1.1~1.3は制限時間10分でよいことがわかったが、談話レベルで個人的なトピックのタスク2と談話レベルで社会的なトピックのタスク3の時間配分については過去の先行研究との比較の観点から制限時間を各々15分とし、タスク1.1~1.3, 2~3で合計40分のテストに変更することに決めた。

(2) 2年目 1年目と同様に、研究発表及び、WTの評価規準のレベル、記述子、模範例の作成と確定、評価者訓練のために、研究会を年間6回にわたり、開催した。高校6校、高専1校で熟達度テスト、WTを実施し、高校1年生と2年生の計650名から641名のデータを入手した。

(3) 3年目 2年目と同様に、研究会を年間6回、更に、統計的技術の向上のためにワークショップを2回を開催した。中学校3校で熟達度テスト、WTを実施し、2年生、3年生、計450名から408名のデータを入手した。

## 4. 研究成果

パート1高校の場合と2中学校の場合に分ける。まず、パート1. 高校の場合。

2007年6-7月に、計7校、650名を対象に英語熟達度テスト(英検2級程度、52問、多肢選択式、45分)WT(文レベルで3問、10分、談話レベル2問各15分で30分、記述式、計40分)を実施した。これら2つのテストを両方とも受験した641名のデータを使いShaw & Weir (2007)の妥当性検証枠組

(1) テスト受験者の特徴、(2) 認知的妥当性、(3) 文脈妥当性、(4) 採点妥当性、(5) 基準関連妥当性、(6) 結果妥当性の6つの観点からなる枠組に従って妥当性検証を行った。(1) テスト受験者の特徴としては、英語熟達度が上位群から下位群に至るまで、上下の幅が大きかったこと、地域的にも北海道から関東地方まで広範囲に渡ったことが挙げられた。(2) 認知的妥当性(=タスク自体を超えた文脈で即ち、現実生活のタスクを行うことに関わる認知的処理を如何に密接に表しているかの尺度)については、文レベルの作文タスク1.1~1.3、談話レベルの作文タスク2,3を記述して「知識告知」(knowledge telling)、「知識変形」(knowledge transforming)の観点から分類した。(3) 文脈妥当性(=文脈妥当性は、成功したタスク実現のために合致しなければならない言語学的及び文脈の要求及び、要求される言語運用を記述するのに役立つ

タスク状況の特性に関係している)については、応答形式、生徒の応答の長さ、解答時間の制限設定したことを述べ、文レベルのタスク1.1~1.3の内、1.1の感謝を述べる英文でも下位群の3割が書けないこと、1.2の「分遅れる」の表現は上位群でも3割が書けないこと、1.3の勧誘の表現は、上位群で63.7%が書けないこと、談話レベルのタスク2~3の内、「旅の行き先」とその理由を聞く2では、全体の74.4%が不合格(5段階の2)、タスク3の「田舎と都会ではどちらに住みたいか」を聞くタスクでは全体の66.1%が不合格であった。

(4) 採点妥当性 (=テスト得点の信頼性に影響するテスト過程のすべての面に関連している特性)については、採点のために文レベルのタスク1.1~1.3のためのタスク実現度I(総合的評価、TR1)、談話レベルのタスク2~3のためのタスク実現度II(分析的評価、TR2、意味の一貫性、パラグラフ・ライティングの書き方ができるかを測る評価尺度)、記述子、模範例を作成し、評価者訓練を3回を実施した。その結果、評価者間信頼性は0.9を超えた。その後、研究者4人が各々、1クラス分を採点した結果、評定者間信頼性は0.9を越えた。その後は、各自、160~210余名を採点した。熟達度テストの信頼性は、クロンバックで0.912で高かった。

(5) 基準関連妥当性については、併存的妥当性を扱った。以下のようなリサーチ・クエスション(RQ)を設定した。

RQ1 英語熟達度テスト(語彙・文法・作文・読解)と作文のタスク実現度I・IIは、どのような関係にあるか

RQ2 英語熟達度テストは正確さとどの程度の関係がみられるのか。

RQ3 英語熟達度テストは流暢さとどの程度の関係がみられるのか。

RQ4 英語熟達度テストは複雑さとどの程度の関係がみられるのか。

RQ5 英語熟達度テストは語彙的特徴とどの程度の関係がみられるのか。

RQ6 英語熟達度テスト(語彙・文法・作文・読解)のどの要素が作文のタスク実現度I・IIに大きな影響を与えるのか

RQ7 作文のタスク実現度I・IIと正確さはどのような関係がみられるのか。

RQ8 作文のタスク実現度I・IIと流暢さはどのような関係がみられるのか。

RQ9 作文のタスク実現度I・IIと複雑さはどのような関係がみられるのか。

RQ10 作文のタスク実現度I・IIと語彙的特徴はどのような関係がみられるのか。

RQ11 正確さ、流暢さ、複雑さ、語彙的特徴のうち、どの指標が作文のタスク実現度I・IIに大きな影響を与えるのか

研究者間の連絡上の誤解からRQ5、RQ10は、相関を調べることが出来なかった。また、RQ6、RQ11は将来的な課題とすることになった。研究質問1、熟達度と作文のタスク実現度I、IIとの相関については、熟達度テストの作文整序問題から文レベルの作文のタスク実現度I(総合的評価)を予測する際に、作文整序問題よりも熟達度テストの文法問題、更には、熟達度テストの合計点の方がタスク実現度Iを予測するのに有用でありそうであること、談話レベルの作文でタスク実現度IIを予測する際、文法問題、熟達度テストの合計点からは強く予測できないようであること等であった。

研究質問2、7の正確さと英語熟達度テストと正確さ、作文のタスク実現度I、IIと複雑さの相関については、タスク1.1、タスク1.2において、ほぼ熟達度の上位群、中位群、下位群間に殆ど差がなかったこと、EFTGT(文法的グローバルエラーの無いTユニット(Tユニット全体の)割合)では上位、中位、下位群間の差は有意ではなかったこと、談話レベルの作文のタスク実現度IIにおけるEFTLT(語彙的グローバルエラーの無いTユニットの(Tユニット全体の)割合)について、上位群、下位群に有意差が無かったことは興味を引くこと等である。

研究質問3の流暢さと英語熟達度テスト、タスク実現度と流暢さの相関については、流暢さの語数と熟達度テストの合計点とは一貫して有意な相関であったこと、熟達度を見るには文の数よりも語数を見るほうが有効らしいこと、特に談話レベルの作文においてそれが言えること等が示唆されたことである。

研究質問4の英語熟達度テストと複雑さ、作文のタスク実現度I、IIと複雑さの相関については、全ての文法的複雑さが熟達度の差を明確に生み出したタスクはタスク1.2であり、タスク1.2は英語の実力差を産み出しやすいタスクであること、長文問題の従属節の使用には殆ど実力差を生み出さないこと、複雑さの指標と熟達度とは殆ど相関がない又は、弱い相関であることである。

研究質問5、10の英語熟達度と語彙的特徴、作文のタスク実現度I、IIと語彙的特徴の相関については、ギロー指数(Guiraud Index)を使って語彙の豊かさを比べると熟達度に差が出たこと、高校生の語彙の豊かさを調べる上で、ギロー指数は他の指標と比較して、

ある程度、有効であるとみなせること等である。

#### パート2 中学校の場合

2008年6月7月に、計3校、450名を対象に英語熟達度テスト(英検3級程度、52問、多肢選択式、45分)WT(高校と同一の問題)を実施した。これら2つのテストを両方とも受験した408名のデータを使い、高校の場合と同様にShaw & Weir(2007)の妥当性検証枠組に従って妥当性検証を行った。(1)「テスト受験者の特徴」としては、英語熟達度が上位群から下位群に至るまで、上下の幅が大きかったこと、地域的には関東地方、中部地方に限られたことが挙げられた。(2)認知的妥当性については、高校の場合と同様に、文レベルの作文タスク1.1~1.3、談話レベルの作文タスク2,3を記述して知識告知、知識変形の観点から分類した。(3)文脈妥当性については、応答形式、生徒の応答の長さ、解答時間の制限設定について述べ、文レベルのタスク1.1~1.3の内、1.1の感謝を述べる英文でも下位群の94.5%が書けないこと、1.2の「分遅れる」の表現は全体の約8割が書けないこと、1.3の勧誘の表現は、全体の93%が書けないこと、談話レベルのタスク2~3の内、「旅の行き先」とその理由を聞くタスク2と「田舎と都会ではどちらに住みたいか」を聞くタスク3では、不合格の率がタスク2の方が高かった。これは、タスク2は選択の幅が無限にあり、先ず受験者が1つの目的地を選ぶのに苦労し、次にその理由付けを考えなくてはならず、認知的負荷が高く、うまく文章を構成できなかったことが考えられる。

(4)採点妥当性については、高校のWTで評価者間訓練を実施後、程なく、中学校のWTが実施され、評定者にすぐに答案が配布された。そこで、評価規準、模範例、すでに行った評価者訓練で使用したかつての答案を練習用に使って思い出してもらい、中学生の答案の採点をしてもらった。従って、高校の場合と同様、タスク1.1~3まで、評定者間信頼性は0.9以上をキープしたと思われる。

(5)基準関連妥当性については、併存的妥当性を扱った。高校の場合と同様のリサーチ・クエスチョン(RQ)1~11を設定した。

研究質問1、熟達度と作文のタスク実現度I, IIとの相関については、熟達度テストの文法下位テストとタスク1.1~1.2で相関が有意で0.6を越えており、一文レベルのタスク1.1~1.2は文法能力を測定しているように思われること、熟達度テストの合計とタスク1.1~1.3の相関は、タスク1.1は強い相

関、タスク1.2,1.3とは有意で0.6を越える相関であること、従って、熟達度テストの合計点で一文レベルの作文の評価をかなり予測できそうであること、しかし、タスク2~3は、意味の一貫性を測定しており、熟達度の合計点とは測っている能力が異なるので、中程度の相関であったことである。

研究質問2,9の正確さと英語熟達度テスト、タスク実現度と正確さの相関について、EFTG(文法的グローバルエラーの無いTユニットの数)については、上位群は有意差がなく、中位群が下位群より上回っていたこと、即ち、熟達度が高くて必ずしも誤りが少ないということは無かったこと、EFTL(語彙的グローバルエラーの無いTユニットの数)について、タスク1.2、談話レベルの作文タスク2,タスク3において中位群、上位群が下位群より高かったが上位群、中位群に有意差がなかった。文法的誤りのないTユニットの割合は、タスク1.2,1.3において下位群が上位群より高かったこと、談話レベルの作文では上位群、下位群間で差が出なかったこと等が興味深い。

研究質問3,8の流暢さと英語熟達度テスト、タスク実現度と流暢さの相関については、熟達度テストの合計点とタスク1.2の語数の相関が有意で高かった( $r = .79$ )こと、中学校の場合も高校の場合と同様に、熟達度を見るには文の数よりも語数の指標を見るほうが有効であること等である。

研究質問4,9の英語熟達度テストと複雑さ、作文のタスク実現度I,IIと複雑さの相関については、中学生の方が高校生よりもどのタスクにおいても複雑さにより熟達度の差が生み出されたこと、特に、節の数で英語の実力差が明確にされるが従属節が多く表出されるタスク2,3に英語の実力差が現れる傾向があること等である。

研究質問5,10の英語熟達度と語彙的特徴、作文のタスク実現度I,IIと語彙的特徴の相関については、ギロー指数(Guiraud Index)を使った場合、文レベルの作文タスク1.1で熟達度順に高い数値を示した。談話レベルの作文のためのタスク実現度IIでもギロー指数が熟達度の順に高く計測されたこと等である。

(6)結果妥当性(=一般に「波及効果」といわれる妥当性)については、英語熟達度テストとWTを受験した中部地方の公立中学校2年生6クラス全員212名を対象として2008年12月に意識調査を実施した。分析の結果、今回のWTの構成概念(測るべき能力・技能)がテスト項目に反映していたと言える。波及

効果と表面的妥当性については、中程度であり、全体として、結果妥当性は、テストの学習、教授への影響の点から高いとは言えなかった。

得られた成果の国内外の位置づけとインパクト、今後の展望について

本研究では、Messick (1989)の妥当性検証枠組と異なる Shaw & Weir(2007)の枠組を採用し、我が国で初めて妥当性検証を試みた。テスト分野で今後、妥当性検証がますます重要視されていくので、本研究がその一端を担えれば幸いである。今後、6つの要素からなる妥当性を更に細かく分析していきたい。

#### 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)

望月昭彦・長橋雅俊「高校における英語ライティング・テストの妥当性検証  
Shaw and Weir (2007)の枠組を基にして抽出100名のデータ - 」『語学教育研究論叢』第26号、大東文化大学語学教育研究所、103～131.2009. 審査有、  
<http://www.daito.ac.jp/gakubu/gaikoku/goken/>

[学会発表](計4件)

望月昭彦・長橋雅俊、「中学・高校の日本人教師のための信頼性・妥当性の高いライティングの評価規準作成 高校生のライティング・テストのタスク実現性、語彙について」第34回全国英語教育学会東京研究大会、昭和女子大学、2008年8月9日

久保田章、鈴木基伸、「高等学校におけるタスクを活用したライティング能力の評価」第14回日英・英語教育学会研究大会、成蹊大学、2008年9月27日

久保田章「英作文の評価におけるより効果的な客観的指標について」外国語教育メディア学会関東支部第120回研究大会、東洋学園大学、2008年6月7日

鈴木基伸「英語を内在化させる授業実践」第38回中部地区英語教育学会長野大会、清泉女学院大学、2008年6月28日

[図書](計1件)

望月昭彦、久保田章、鈴木基伸、平成18年度～20年度、文部科学省 科学研究費補助金(基盤研究(C))研究成果報告書、課題番号 18520423 『中学・高校の

日本人教師のための信頼性・妥当性の高いライティングの評価規準作成』The Construction of Reliable and Valid Criteria for Writing Tests for Japanese Teachers of English at Junior and Senior High Schools、いなもと印刷、2009.1～266. メールアドレス：  
mochizua@ic.daito.ac.jp

#### 6. 研究組織

##### (1)研究代表者

望月 昭彦 (MOCHIZUKI AKIHIKO)  
大東文化大学・外国語学部・教授  
研究者番号：30219969

##### (2)研究分担者

久保田 章 (KUBOTA AKIRA)  
筑波大学・大学院人文社会科学部  
研究者番号：30205132  
鈴木 基伸 (SUZUKI MOTONOBU)  
豊田工業高等専門学校・一般学科  
研究者番号：50321443