多肢選択式項目の項目形式が文法テストパフォーマンスに与える影響について

The Effects of Multiple-Choice Item Formats on Grammar Test Performance

熊澤孝昭（くまざわたかあき）
関東学院大学

Item formats are a facet of testing that influences examinees' test performance. In this study, six types of item formats were adopted, and 55 multiple-choice grammar items were developed and administered to 608 first-year university students for placement and diagnostic purposes. The research questions were: to what extent do items function for placement and diagnostic purposes, to what extent do item formats differ in terms of difficulty, to what extent do item formats contribute to the total score variance, and to what extent is the grammar test reliable. Based on the item analyses, most items functioned for placement and diagnostic purposes. FACETS analysis revealed that the six item formats differed in terms of difficulty. The generalizability study showed that 3% of the variance components was due to the item formats. The decision study indicated that the generalizability coefficient and dependability index were satisfactory for placement and diagnostic purposes. Based on the results, implications are discussed.

Keywords: grammar item format, multiple-choice items, test methods, placement test, diagnostic test, generalizability theory, multifaceted Rasch model

項目形式はテストパフォーマンスに影響する一つの要因だといわれている。本研究では6種類の項目形式を用いて55項目の多肢選択式項目を文法テストとして作成し、608名の大学一年生にプレイスメントテストと診断テストとして実施した結果を報告する。研究目的は、項目の機能、項目形式の困難度、項目形式などの分散成分の割合、文法テストの信頼性の検証である。項目分析の結果、ほとんどの項目がプレイスメントと診断テストの項目として機能していたことがわか

った。FACETS分析の結果、6種類の多肢選択式項目の形式はそれぞれ困難度が異なることがわかった。一般化可能性研究の結果、項目形式の違いによって生じた分散成分が若干あり、文法テストパフォーマンスに影響する要因であることも明らかとなった。決定研究の結果、一般化可能性係数は.81で、信頼度指数は.75であった。上述の結果を踏まえ、教育的示唆について論じた。

キーワード: 文法項目形式、多肢選択項目、テスト方法、プレイスメントテスト、診断テスト、一般化可能性理論、多相ラッシュモデル

はじめに

テストを受験している際のパフォーマンスは得点という結果になるが、その得点に影響すると思われる要因は様々であり、言語テストの分野ではその要因の解明や影響の度合いが調査されてきた。その要因の中でも、テスト環境、テスト指示、インプット、予期する応答、インプットと応答の関係であるテスト方法（test method）は受験者のテストパフォーマンスに大きく影響する相（facet）だといわれている（Bachman, 1990）。予期する応答についてはさらに選択形式応答の項目（selected response item）と応答構築式型の項目（constructed response item）の二つの形式に分けられる。真意判定項目（true/false item [T/F]）、多肢選択式項目（multiple choice item [MC]）、整合項目（matching item）などは受験者がいくつかある選択肢から正解を一つ選ぶ形式なので、選択形式応答の項目である。空欄箇所補充項目（fill-in item）、短答形式の項目（short-response item）、クローズ項目（cloze item）などは受験者が英単語などの言語的なアウトプットを記述することで正解することができるため、応答構築式型の項目である。これらの項目形式は受験者のテストパフォーマンスに影響するため、テスト作成者はテストの目的に沿うようもっとも妥当な形式を選ぶ必要がある。

応答構築式型の項目より選択式応答の項目のほうが入念に作成する必要があるが、後者は採点がマークシートリーダーなどを用いることで容易に採点処理ができるのでより実用的であり、利害関係が大きい（high stakes）テストであるセンター試験や入学試験から比較的に利害関係が小さい授業内テストまでよく用いられる。その中でも選択肢がいくつか設けられるMCがもっとも一般的に用いられている。MCにも多様な形式があり、どのMC項目形式が受験者には難しかったかとMC項目形式という要因がどの程度得点に影響するかを調査することはテスト開発者に重要な示唆を与えると考える。例えば、難しいと思われる項目形式を用いることでテストの難易度を上たり、項目形式が与える影響を最小限にすることでより妥当性が高いテストを開発できると考えられる。本研究では文法テストにある6種類のMC項目形式の困難度、MC項目形式が得点に与える影響の度合い、文法テストの信頼性を調査することを主な目的とする。

研究の背景

心理測定の分野ではMCを含む項目作成方法について書かれた文献がある（Haladyna, 1992, 2004; Haladyna, Downing, & Rodriguez, 2002等）。言語テストの分野でも、項目作成の際に避けるべき事項などの指針が記してある文献がある（Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 1996; Brown, 2005等）。MC

に関する実証研究もあり、Shizuka, Takeuchi, Yashima, and Yoshizawa（2006）は、大学入学試験で多用される4択MCを3択MCと比較し、3択MCでも十分に機能すると示唆した。Dudley（2006）は語彙と読解としてのMCと多肢真意判定項目（multiple true/false item）と比較した結果、多肢真意判定項目はMCとほぼ同様に機能すると示唆した。

項目形式が受験者の読解テストパフォーマンスに与える効果についての研究はShohamy（1984）、Kobayashi（2002）がある。Shohamy（1984）はMCと自由回答式項目（open-ended items）からなる読解テストを2000人の学習者に実施した結果、MCのほうが自由回答式項目より易しいという結果を得た。自由回答式項目の場合、英語を記述しないと正解にならないため、発表技能（productive skill）が試されるので難しいとの理由である。また、英語力が高い受験者はさほど項目形式にパフォーマンスは影響されなく、逆に英語が低い受験者はより影響されると主張している。Kobayashi（2002）はテキストの種類と項目形式が読解テストのパフォーマンスに与える影響について調査した。項目形式は3種類あり、クローズ、自由回答式項目、要約ライティング（summary writing）で、信頼性は高かった。平均値をもとに項目形式の難易度を比較したところ、自由回答式項目、要約ライティング、クローズの順に難しかった。しかし、テキストの段落構成がなっていない場合、クローズが難しかった。教育的示唆としては、それぞれの項目形式は違う読解能力を測っているようで、自由回答式項目と要約ライティングはどちらかといえば大まかな内容を理解したかを測っているとのことであった。

聴解テストのパフォーマンスに与える影響に関しての研究として、Shohamy and Inbar（1991）、Yi’an（1998）、Brindley and Slatyer（2002）がある。Shohamy and Inbar（1991）では、テスト形式の他に、テキストの種類も一つの要因として調査している。結果として、ニュース放送が最も難しく、講義と会話は比較的易しいことがわかった。また、本文の内容にある事柄を理解したかを問う自由解答式項目は全体の内容を理解したかを問う項目より易しいことがわかった。Yi’an（1998）はMCからなる聴解テストを受験しているEFL学習者に回顧的言語化報告（retrospective verbal report）を用いてテストを受けている際の頭脳処理を調査した。結果として、基本的には言語知識の処理を行っているが、言語知識を補うため非言語知識が使われ、MCは英語力が高い学習者が消去法などを駆使し正解できる場合があるので妥当性に欠けると主張した。Brindley and Slatyer（2002）は話す速度（speech rate）、テキスト、項目形式が聴解タスクと項目困難度に与える影響について調査した。用いられた項目形式は空欄箇所補充項目（sentence completion）、短答形式の項目（short response）、表に必要情報を記入する表記述式項目（table answer）で、すべて応答構築式型の項目であった。結論として、会話速度を変えるなどして難しいと思われた項目は易しい場合があり、一概に特定の相が項目を難しくするというわけではなく、相の交互作用が項目を難しくすると主張している。

In’nami and Koizumi（2009）はMCと自由回答式項目が第一言語（L1）の読解、第二言語（L2）の読解、L2の聴解テストパフォーマンスに与える影響について言及した56もの研究を用いてメタ分析（meta-analysis）を行っている。MCはL1読解とL2聴解テストにおいて自由回答式項目より易しいという結果になった。L2読解では特定の研究を用いて分析を行えば同様の結果が得られた。

語彙テストの研究として、Morimoto（2006）は3種類の40項目を含むMCからなるテスト形式（test form）を英語力が同等だと思われる三つのグループに語彙知識を測る

目的として実施した。タイプAは英文中にある下線が引かれた英単語の意味にもっとも近い英単語を選ぶ形式で、文脈独立的（context-independent）項目で、文脈から答えを導き出せないものである。タイプBももっとも近い意味を選ぶ形式だが、文脈依存的（context-dependent）項目で、文脈から答えを導き出せるものである。タイプCは英文中にある空欄箇所にもっともあてはまる英単語を選ぶ形式で、文脈依存的項目である。三つのテスト形式の信頼性は高いといえる。平均値が最も高かったテスト形式はタイプBであった。つまり、文脈からより情報が得られるテスト形式の平均値が高かった。

文法テストの研究として、David（2007）は新たな文法項目形式であるマルチトラック（multitrack）を考案し、従来よく用いられる3種類の項目形式である文中にある空欄箇所補充MC（sentence-based）、クローズMC（text-based MC cloze）、文中にある空欄二箇所補充MC（double-blank sentence-based）の項目形式困難度を調査するためFACETS分析を用いた。その結果、項目形式困難度の推定値のばらつき度合いを示す分離指数（separation index）は13.00となり、そのばらつきが有意かを示すカイ二乗は有意な数値になり、マルチトラック、空欄二箇所補充MC、空欄箇所補充MC、クローズMCの順に困難度の推定値が高く、項目形式は文法テストのパフォーマンスに影響する要因であることがわかった。

研究の目的

テスト方法についての既存研究では予期する応答についての研究が主であり、選択形式応答と応答構築式型の項目の難易度など項目形式がパフォーマンスに与える効果について検証され、選択形式応答の項目形式のほうがより易しいとの結果となっている（In'nami & Koizumi, 2009; Shohamy, 1984）。また、MCの項目形式に関する研究は少数みられ、それぞれ難易度も違いがあることが指摘されている（David, 2007; Morimoto, 2006）。MCは授業内テストのような利害関係が小さいテストから入学試験のような利害関係が大きいものまで幅広く用いられる項目形式であり、テストのパフォーマンスに影響するので、さらにその相が与える影響について調査をする必要がある。しかし、その相が文法テストパフォーマンスに与える影響について言及した研究はDavid（2007）のみなので、さらなる調査が必要である。よって、本研究は以下の4点を研究目的とする。

1. 本研究で用いた文法MCはプレイスメントテスト目的と診断テスト目的として受験者にどの程度機能しているか。
2. 本研究で用いた文法テストにある6種類の項目形式はどの程度困難度の推定値には違いがあるか。
3. 受験者、項目、項目形式の相は文法テスト得点の分散に対してどの程度の割合であるか。
4. 本研究で用いた文法テストはどの程度信頼性があるか。

方法

対象者

本研究は関東地方にある私立大学で行われた。対象者はテストを受験する前に研究目的として結果を用いられることに同意した工学部に所属する2009年度入学した一年生で、608名の英語学習者である。対象者の中には中等教育で習った学習内容を完全に習熟していない学生がいるため、リメディアル英語教育にも力を入れている。よって、カリキュラムの主要な目標の一つは英語の基礎である文法知識を学習することである。

文法テスト

この文法テストの開発目的はプレイスメントテストと、到達度テスト（achievement test）としても実施できるテストを開発することである。よって、新学期がはじまる4月にプレイスメントテストとして実施し、テスト得点をもとにクラス分けを行い、一年後に同じテストを実施し、授業目的を学習したかを調査する目的で開発された。つまり、このテストはプレイスメントテストとしても用いられる目標規準準拠テスト（criterion-referenced test　[CRT]）であるといえよう。やはり、プレイスメントテストの内容はカリキュラムの内容にある授業目的と合致したものがよいと考えた（Bachman, 1990; Brown, 1989）。

試行テスト

2008年12月に試行テストを行った。試行テストの作成手順は、まずリメディアル英語教育で使用する教科書にある文法項目や練習問題をみて、教科書の内容を補足する練習問題を作成し、練習帳として仕上げ冊子にした。扱った16の文法項目はbe動詞、一般動詞、未来形、助動詞、冠詞、代名詞、前置詞、接続詞、比較、進行形、to不定詞、動名詞、受動態、現在完了、関係詞、仮定法である。教科書と練習帳でよく用いられた項目形式は空欄箇所補充、英和訳、和英訳、並べ替え、MCなどである。試行テストでは教科書にある文法項目と項目形式を用いてテスト項目を作成することとした。セクションは四つあり、それぞれ項目形式は空欄箇所補充MC、英和訳MC、和英訳MC、間違い探しMCで、項目数はそれぞれ32、16、16、16問である。

試行テストは、工学部の9クラスと法学部の4クラスを履修した合計374名の2008年度入学した一年生に授業内で実施された。受験時間は一時間程度であった。文法テスト以外にも65問からなる語彙テストもあり、語彙から文法の順に受験したためか、受験者の中には時間が足りなかった人がいたとの報告もあった。しかし、項目についての良質なデータを得ることができた。分析の結果、平均点は80点満点中38点で、項目数が十分にあったためかそれぞれのセクションの信頼性は高いほうであった（α＝.80, .74, .76, .70）。

本試験

2009年4月に本試験であるプレイスメントテストを実施した。そのテストを開発するため、試行テストの結果をもとにテスト項目を改良した。まず、項目容易度（item facility）、項目弁別力（item discrimination）など古典的テスト理論にもとづき項目

分析を行うコンピュータソフトであるITEMAN [version 3.6] (Assessment Systems Corporation, 1996)を用いた。項目の是非を決める基準は項目容易度が.70以下で項目弁別力が.20以上である。項目容易度が高いということはすでに受験者はその項目を学習済みであり、学習度合いを測るのに適していないといえる。また、プレイスメントテストは得点のばらつきを生じさせたいので、項目弁別力が高いほうがよい。セクション1は16の文法項目につき２項目ずつ作成したので、二つの基準をより満たした項目を二つのうちから一つ選び、撹乱肢分析（distractor analysis）の結果なども参考にし、選択肢にある文言を変更するなどをして改良した。セクション2、3、4は試行テストでは16の文法項目について順に1項目ずつ作成した。練習帳では並べ替えの項目形式が用いられていたが、試行テストではなかったので、新たに偶数の文法項目についての項目を8問作成し、セクション5とした。セクション6ではクローズMCを7問作成した。他のセクションにある項目数は8項目、または16項目なので、各セクションの項目数を均等にすべきだが、語彙テストが25問と文法テストが55問で、合計項目数を80問としたかったため、セクション6のみ7問とした。文法テスト以外に語彙テストとして25問出題されたが、本稿の主題に外れるので触れないこととする。

セクション1は空欄箇所補充MC（k = 16）で、文中の空欄にもっともあてはまる選択肢を一つ選ぶもので、例題は以下の通りである。

There (　) a pen on the table.

①　is　　②　am　　③　are　　④　aren't

この項目はbe動詞についてで、正解は①である。セクション2は英和訳MC（k　=　8）で、英文の意味にもっとも合う和文を選択肢①～④から一つ選ぶもので、例題は以下の通りである。

There isn't a pen on the table.

①　ペンは机の上にある。
②　ペンは机の上にあった。
③　ペンは机の上にありましたか。
④　ペンは机の上にない。

これもbe動詞についての項目で、正解は④である。セクション3は和英訳MC（k＝8）で、和文の意味にもっとも合う文を選択肢①～④から一つ選ぶものであり、例題は以下の通りである。

彼は多くのレポートを書いた。

①　He is writing many reports.
②　He is written many reports.
③　He writes many reports.
④　He wrote many reports.

これは一般動詞についての項目で、正解は④である。セクション4は間違い探しMC（k = 8）で、英文中にある文法的に間違いだと思う選択肢を①~④から一つ選ぶという形式で、例題は以下の通りである。

① My sons ② is ③ not at home ④ yesterday.

これはbe動詞についての項目で、正解は②である。セクション5は並べ替えMC（k = 8）で、文中にあるカッコ内の①～④を意味の通るよう文に並べ替え、3番目にくる選択肢を①-④から一つ選ぶという形式であり、例題は以下の通りである。

Meg（① did ② cook ③ well ④ not）.

これは一般動詞についての項目で、正解は②である。セクション6はクローズMC（k = 7）で、会話文中にある空欄箇所にあてはまる選択肢を①-④から一つ選ぶというもので、例題は以下の通りである。

A: I went to Tokyo last week.

B: Really? How ( ) it?

A: It was nice.

① is ② are ③ was ④ were

これはbe動詞についての項目で、正解は③である。

分析の手順

まず、テスト結果を把握するためSPSS [version 17.0]を用いて記述統計と内部一貫性信頼性（internal consistency reliability）であるクロンバックアルファ係数（α係数）を求めた。項目が受験者にプレイスメントテストとしてと、目標規準準拠テストとしても機能しているかを古典的テスト理論の観点からみるため、項目容易度、項目弁別力、*B*-指数（*B*-index）をエクセルで算出した。*B*-指数とはいかに合格者と不合格者をある項目が区別しているかを示す指数である。

次に、FACETS [version 3.41] (Linacre, 2002)を用いて受験者能力値、項目困難度、項目形式困難度の三つの相の推定値を求めた。このソフトは項目応答理論の一種である多相ラッシュモデル（multifaceted Rasch model）にもとづいている（Linacre, 1989）。ラッシュモデル（Rasch model）とはある項目の項目困難度の推定値とある受験者の正解や不正解の応答にもとづき受験者能力値を推定する統計モデルである（Bond & Fox, 2001）。多相ラッシュモデルとは項目困難度や他の相の推定値とある受験者の得点をもとに受験者能力値を推定する統計モデルである（Linacre, 1989）。また、このFACETS分析はより受験者の能力や項目の困難度などに影響されない項目形式困難度の推定値を求めることができる。一般的にこのモデルは評定者が受験者の能力を採点するパフォーマンステストで用いられる（Coniam, 2008; Norris, Brown, Hudson, & Bonk, 2002; Schaefer, 2008等）。しかし、項目形式の困難度を求める目的でDavid（2007）もFACETS分析を行っている。

一般化可能性理論（generalizability theory [G theory]）を用いて、テスト得点の分散に対して項目形式の相の分散の割合を一般化可能性研究（generalizability study [G研究]）で推定し、古典的テスト理論では信頼性係数（reliability coefficient）に相当する一般化可能性係数（generalizability coefficient）と信頼度指数（dependability index）を決定研究（decision study [D研究]）で求めた。一般化可能性理論に関してはBrennan（2001a）、ShavelsonとWebb（1991）が解説書を書いている。この理論は一般的には分散分析と古典的テスト理論を用いて算出する信頼性係数を合わせたようなものだといわれている（Gebril, 2009）。つまり、分散分析では各相の分散成分を推定し、その分散成分をもとに一般化可能性係数と信頼度指数を推定する。この理論もパフォ

ーマンステストの分析によく用いられる（Lee & Kantor, 2007; Yamanishi, 2005; Xi, 2007等）。Gebril（2009）はパフォーマンステストにおけるライティングタスクの違いによって分散が生じ、また一般化可能性係数に影響するとの結果を示している。また、MC型TOEFLにおける受験者の第一言語の違いがどの程度テストの分散を生じさせているかを検証するため、Brown（1999）はこの理論を用いている。目標規準準拠テストの信頼度を調査するため、この理論を用いた研究はBrown（1993）、Kumazawa（2009）、Kunnan（1992）がある。本研究でのG研究は受験者（*p*）、項目形式（*f*）、項目が項目形式に組み込まれており（i:f）ネスト（nest）しているので三つの相を分析対象とする。よって、デザインは*p X (i:f)*となる。また、各項目形式に組み込まれている項目数が異なるためアンバランスデザインとなる。使用したソフトはmGENOVA [version 2.1] (Brennan, 2001b)である。D研究のデザインは*p X (I:F)*で様々な項目数の組み合わせによる一般化可能性係数と信頼度指数の変動を検証するため行った。

結果

表1は記述統計の結果を記す。この文法テストは六つのセクションからなり、各セクションは異なる項目形式が用いられている。項目数はセクション1がもっとも多く、16個すべての文法項目についての内容を出題している。セクション2から5は8項目ずつで、セクション6のみ7項目である。最小値は全セクションとも0点で、最大値は満点である。もっとも平均値が高いのはセクション1の8.37で、これは主に項目数が多いためである。平均値が低いのはセクション6の2.37で、主に項目数が少ないためである。項目数がセクションごとに異なるため、平均値ではどのセクションが受験者にとって難しかったかなどを解釈するのは不可能である。標準偏差は得点のばらつきを示す。もっとも値が大きいのはセクション1で次にセクション5、セクション3の順である。α係数はプレイスメントテストとして機能しているかを判断するための重要な統計で、得点のばらつきなどに係数は左右される。セクション6のα係数は.27と低く、測定誤差が大きいことを示す。これは項目数が少ないのが主な理由である。セクション4のα係数も.36で、思わしくない。しかし、全体のα係数は.84で、プレイスメントテストとしては十分だと考える。また、平均値は中央値とほぼ同じの27.14で、歪度と尖度も誤差の範囲以内でほぼ標準分布をなしている。

表1. 記述統計の結果（$N = 608$）

セクション	項目形式	*k*	最小	最大	*M*	*SD*	歪度	尖度	α
1	空欄箇所補充	16	0	16	8.37	3.18	.13	-.58	.70
2	英和訳	8	0	8	5.17	1.48	-.60	.05	.41
3	和英訳	8	0	8	4.15	1.76	-.09	-.48	.46
4	間違い探し	8	0	8	3.13	1.59	.30	-.39	.36
5	並べ替え	8	0	8	3.95	1.86	-.11	-.57	.56
6	クローズ	7	0	7	2.37	1.41	.34	-.27	.27
合計		55	7	55	27.14	8.16	.17	-.55	.84

表2は項目分析の結果を記す。項目容易度はプレイスメントテストの場合、値が.30から.70の範囲にある項目がよいとされている（Brown, 2005）。17項目がその範囲以外で思わしくない値を示す。目標規準準拠テストで事前テストとして実施された場合、項目容易度は低いほうが望ましい。つまり、一定の値以上を示す項目は既に授業を受ける前に学習済みということで、学習する必要がない項目となり、機能していないとみなす。8項目が.70以上の値を示し、すでに学習済みであり機能していない。項目弁別力はいかにある項目が受験者の能力を弁別しているかを示し、プレイスメントテストの項目として機能しているかを判断するうえで重要な統計であり、.20以上を示す項目がよいとされる（Brown, 2005）。9項目の値が低く改善の余地がある。B-指数は合否分割点（cut-off score）を設定した目標規準準拠テストの項目がどの程度合格者と不合格者の能力を区別しているかを示す統計である。合否分割点を60点とした場合、指数が.20以下で機能していないと思われる項目は13項目あった。一方で、全体的にプレイスメントテストと事前診断テストとして機能している項目も多くあった。

図1はFACETS分析で求めた受験者、項目形式、項目の推定値をグラフ化したものである。グラフの左側から受験者、項目形式、項目の順に分布があり、グラフの上にいくほど能力値が高い、項目形式が難しい、項目が難しいということになる。受験者、項目形式、項目の分離指数はそれぞれ2.28、8.02、9.65で、ラッシュ信頼性係数は.84、.98、.99となった。つまり、難しい項目であれば難しく、簡単であれば簡単であるというばらつきを再生できる可能性が高いということを意味する。項目形式の分離指数が高いことから六つの項目形式には困難度のばらつきがあるということがわかる。また、このグラフからもわかるが、項目形式困難度の推定値にはばらつきがある。セクション6のクローズMCがもっとも難しいようである。また、その反対に簡単な項目形式はセクション2の英和訳MCであることがわかる。

FACETS分析を行い求めた項目困難度の推定値、誤差、インフィット平均二乗、アウトフィット平均二乗の結果が表2にある。項目困難度の推定値は－2.79から2.10で、広範囲の能力を測定できる項目がある点から、能力が多様な受験者が受けるプレイスメントテストとしては望ましい。項目1のbe動詞についての項目の困難度推定値は1.48で、これは難しい項目だといえる。be動詞は初期段階である中学一年に学習する項目で、易しいと思われる文法事項であるが、難しい項目となった。一方で項目17もbe動詞についてであるが、項目困難度の推定値は－2.79となり、易しい項目であるといえる。測定誤差は.19以下で誤差が少ない。インフィット平均二乗は実際のデータとモデルとの適合度（fit）を表し、.80から1.20以内であれば項目がモデルに適合しているといえる（Bond & Fox, 2001）。このテストにおいては全項目ともにモデルに適合している。アウトフィット平均二乗ははずれ値に敏感なので、設定した範囲から外れ、不適合（misfit）となった項目が7項目ある。これらの項目の項目弁別力をみると値が極端に高い、または低いことがわかる。どちらかといえば、アウトフィット平均二乗の値が1.20以上になり適合不足（underfit）と判断された5項目がより問題視すべきである。なぜならば文法知識が豊富な受験者がやさしい項目に不正解になったということを示すからである。

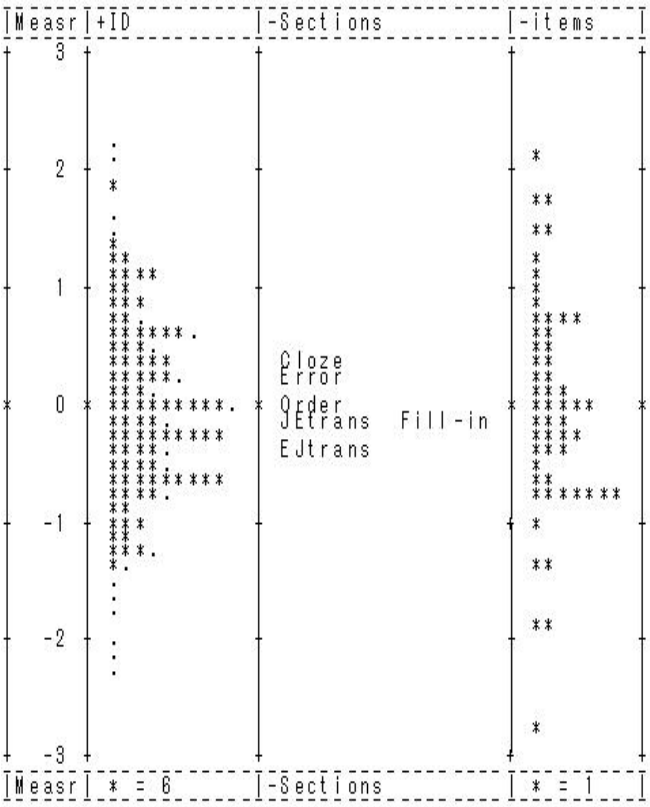

図1. グラフ化したFACETS分析の結果.

表2. 古典的テスト理論とFACETS分析を用いての項目分析の結果（*N* = 608）

		古典的テスト理論			FACETS分析			
項目	文法項目	IF	ID	*B*-指数(.6)	Diff	*SE*	Infit MS	Outfit MS
1	be動詞	.22	.44	.38	1.48	.10	.90	.80
2	一般動詞	.64	.44	.34	-.61	.09	.90	.90
3	未来形	.29	.32	.31	1.08	.09	1.00	1.00
4	助動詞	.59	.43	.33	-.35	.09	.90	.90
5	冠詞	.67	.18	.16	-.75	.09	1.10	1.20
6	代名詞	.51	.33	.32	.03	.09	1.00	1.00
7	前置詞	.18	.36	.29	1.76	.11	.90	1.00

表2. 古典的テスト理論とFACETS分析を用いての項目分析の結果（$N = 608$）

		古典的テスト理論			FACETS分析			
項目	文法項目	IF	ID	*B*-指数(.6)	Diff	*SE*	Infit MS	Outfit MS
8	接続詞	.79	.40	.23	-1.41	.10	.90	.80
9	比較	.53	.39	.36	-.08	.09	1.00	1.00
10	進行形	.57	.45	.37	-.27	.09	.90	.90
11	to不定詞	.85	.34	.15	-1.83	.12	.90	.90
12	動名詞	.37	.37	.34	.69	.09	1.00	1.00
13	受動態	.59	.48	.42	-.34	.09	.90	.90
14	現在完了	.67	.33	.26	-.73	.09	1.00	1.00
15	関係詞	.43	.32	.34	.40	.09	1.00	1.00
16	仮定法	.48	.46	.40	.16	.09	.90	.90
17	be動詞	.95	.27	.07	-2.79	.19	.90	.70
18	未来形	.71	.40	.27	-.62	.09	.90	.90
19	冠詞	.18	.06	.05	2.10	.11	1.10	1.40
20	前置詞	.69	.38	.30	-.52	.09	1.00	.90
21	比較	.45	.34	.38	.61	.09	1.00	1.00
22	to不定詞	.88	.40	.14	-1.85	.13	.90	.70
23	受動態	.57	.17	.11	.09	.09	1.10	1.20
24	関係詞	.74	.37	.24	-.77	.10	.90	.90
25	一般動詞	.55	.52	.46	-.19	.09	.90	.80
26	助動詞	.59	.29	.26	-.37	.09	1.00	1.00
27	代名詞	.67	.39	.27	-.77	.09	.90	.90
28	接続詞	.73	.42	.31	-1.05	.10	.90	.90
29	進行形	.31	.19	.14	.97	.09	1.10	1.10
30	動名詞	.50	.32	.27	.05	.09	1.00	1.00
31	現在完了	.46	.34	.28	.23	.09	1.00	1.00
32	仮定法	.35	.20	.18	.74	.09	1.10	1.10
33	be動詞	.59	.37	.37	-.70	.09	1.00	.90
34	未来形	.61	.44	.37	-.80	.09	.90	.90
35	冠詞	.16	.05	.05	1.53	.11	1.10	1.60
36	前置詞	.20	.07	.04	1.27	.11	1.10	1.40
37	比較	.32	.24	.21	.57	.09	1.10	1.10
38	to不定詞	.47	.23	.21	-.16	.09	1.10	1.10
39	受動態	.42	.39	.40	.05	.09	1.00	1.00

表2. 古典的テスト理論とFACETS分析を用いての項目分析の結果（*N* = 608）

		古典的テスト理論			FACETS分析			
項目	文法項目	IF	ID	*B*-指数(.6)	Diff	*SE*	Infit MS	Outfit MS
40	関係詞	.37	.35	.34	.28	.09	1.00	1.00
41	一般動詞	.77	.45	.30	-1.35	.10	.90	.80
42	助動詞	.40	.26	.19	.44	.09	1.10	1.10
43	代名詞	.49	.44	.42	.04	.09	.90	.90
44	接続詞	.18	.08	.10	1.71	.11	1.10	1.40
45	進行形	.66	.50	.33	-.77	.09	.90	.80
46	動名詞	.35	.30	.28	.70	.09	1.00	1.10
47	現在完了	.55	.40	.31	-.25	.09	1.00	.90
48	仮定法	.56	.39	.34	-.28	.09	1.00	.90
49	be動詞	.33	.28	.26	.41	.09	1.00	1.10
50	仮定法	.38	.27	.23	.15	.09	1.10	1.10
51	関係詞	.32	.31	.28	.46	.09	1.00	1.00
52	代名詞	.24	.23	.20	.90	.10	1.10	1.10
53	接続詞	.40	.16	.21	.05	.09	1.10	1.20
54	比較	.26	.01	-.06	.77	.10	1.20	1.30
55	to不定詞	.44	.25	.24	-.14	.09	1.10	1.10

注. IF = 項目容易度、ID = 項目弁別力、Diff = 項目困難度、*SE* = 標準測定誤差、Infit MS = インフィット平均二乗、Outfit MS = アウトフィット平均二乗

表3はFACETS分析を用いての項目形式困難度の推定値を示す。分離指数とラッシュ信頼性係数は8.02と.98で、高い値を示した。また、カイ二乗も有意な値を示し、項目形式困難度の推定値は有意なばらつきがみられた。平均値をみるだけでは項目数が違うなどの理由からどの項目形式が難しいかを解釈することはできないが、FACETS分析を用いるとすべての相の推定値が同じ間隔尺度上にあるので解釈が容易である。項目形式困難度の推定値をみると、クローズ、間違い探し、並べ替え、和英訳、空欄箇所補充、英和訳の順で難しい項目形式であるといえる。測定誤差も少なく、モデルの適合度もよい。

表3. 項目形式困難度の推定値

セクション	項目形式	セクション困難度	*SE*	Infit MS	Outfit MS
1	空欄補充	-.10	.02	1.00	1.00
2	英和訳	-.43	.04	1.00	1.00

表3. 項目形式困難度の推定値

セクション	項目形式	セクション困難度	*SE*	Infit MS	Outfit MS
3	和英訳	-.08	.03	1.00	1.00
4	間違い探し	.27	.03	1.00	1.00
5	並べ替え	-.03	.03	1.00	1.00
6	クローズ	.36	.03	1.00	1.00

注. SE = 標準測定誤差、Infit MS = インフィット平均二乗、Outfit MS = アウトフィット平均二乗

表4は*p X (i:f)*アンバランスデザインによるG研究の結果を記す。受験者の文法知識の差によって生じた分散成分の割合はテスト得点の分散成分を100%とした場合、7%であった。項目形式の違いによって生じた分散成分の割合は3%であった。項目によって生じた割合はより高く13%であった。受験者と項目形式の交互作用の分散成分の割合は1%であった。つまり、受験者によって得意不得意な項目形式が若干あり、テストパフォーマンスに影響したといえる。受験者と項目の交互作用の割合は76%で、受験者は項目によって異なるパフォーマンスをしたことになるが、これは受験者の特定の項目に対する得意不得意などから生じているのではないかと推測される。

表4. *p X (i:f)*アンバランスデザインによるG研究の結果

相	分散成分推定値	分散成分%
受験者(*p*)	.01789	7%
項目形式(*f*)	.00655	3%
項目(*i:f*)	.03186	13%
受験者X項目形式(*pf*)	.00347	1%
受験者X項目(*p X i:f*)	.19202	76%
合計	.25179	100%

表5は*p X (I:F)*によるD研究の結果を表し、図2はその結果をグラフ化したものである。母得点(universe score [$\sigma^2(\tau)$])は古典的テスト理論では真値(true score)に値するが、その分散は.01789であった。集団基準準拠テストに用いる誤差は相対誤差(relative error($\sigma^2(\delta)$))というが、その誤差は.00413であった。目標規準準拠テストに用いる誤差は絶対誤差(absolute error [$\sigma^2(\Delta)$])というが、その誤差は.00593であった。母得点分散と相対誤差をもとに算出するのが一般化可能性係数(generalizability coefficient [ρ^2])であり、集団基準準拠テストの信頼性の解釈の際に用いられ、その

係数は.81であった。母得点分散と絶対誤差をもとに求めるのが信頼度指数（index of dependability [Φ]）であり、目標規準準拠テストの信頼度の解釈に用いられ、その指数は.75であった。六つのセクションとも4項目の場合、一般化可能性係数は.68で、信頼度指数は.63で、もとの係数と指数より大きく下回る。セクション1から8項目を削除し、セクション6に1項目追加し、全セクションともに8項目ずつにすると、係数は.80となり、もとの係数である.81とさほど変わらない。係数が.90に達するには144項目も必要になる。48項目からなるテストでは係数は.80になり、それ以降項目を追加しても係数には若干の変動しかみられない。

表5. *p X (I:F)*デザインによるD研究の結果

D研究	空欄補充	英和訳	和英訳	間違い探し	並べ替え	クローズ	合計	ρ^2	Φ
1	4	4	4	4	4	4	24	.68	.63
2	5	5	5	5	5	5	30	.72	.66
3	6	6	6	6	6	6	36	.75	.69
4	7	7	7	7	7	7	42	.78	.72
5	8	8	8	8	8	8	48	.80	.74
6	9	9	9	9	9	9	54	.82	.75
7	10	10	10	10	10	10	60	.83	.77
8	11	11	11	11	11	11	66	.84	.78
9	12	12	12	12	12	12	72	.85	.79
10	13	13	13	13	13	13	78	.85	.80
11	14	14	14	14	14	14	84	.86	.80
12	15	15	15	15	15	15	90	.87	.81
13	16	16	16	16	16	16	96	.87	.82
14	24	24	24	24	24	24	144	.90	.85
15	32	32	32	32	32	32	192	.92	.86

注. ρ^2 = 一般化可能性係数、Φ = 信頼度指数

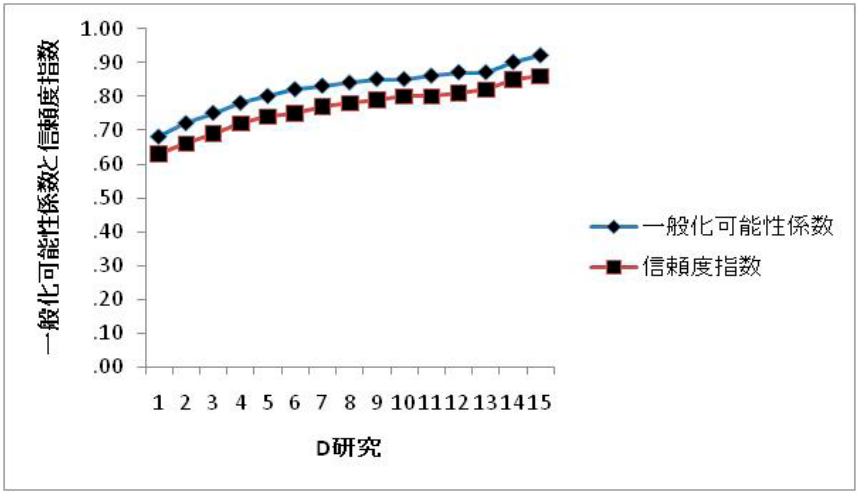

図2. グラフ化したD研究の結果

考察

本章では4点の研究目的について考察する。研究目的1は文法MCは受験者にどの程度機能しているかについてである。項目容易度の値から55項目中17項目がプレイスメントテストとしてはやさしすぎるまたは難しすぎるということとなった。しかし、プレイスメントテストにおいてより重視すべき統計は項目弁別力である。55項目中9項目のみ受験者の文法知識を弁別していないという結果になった。市販の集団基準準拠テストの一種である英語運用能力テスト(English proficiency test)をプレイスメントテストとして用いた場合についての既存研究(Culligan & Gorsuch, 1999; Westrick, 2005)と比較しても、84%の項目が機能している本テストはかなり受験者に合った項目からテストが構成されているといえる。これはやはり受験者の英語力を想定し項目を作成し、一度試行テストとして実施し、その結果をもとに厳密に項目を選抜したからであろう。

この文法テストはプレイスメントテストとして実施したが、学習者がどの程度カリキュラム内容を学習しているかを測る目的でも実施しており、つまり診断テスト(diagnostic test)でもある。その目的を担う項目は項目容易度の値が高くなく、B-指数が高いものがよいが、その条件を満たしていない項目は項目11、17、22の3項目のみであった。これも試行テストの結果をもとに項目容易度が高い項目を削除したからであろう。いかに項目がカリキュラム内容の学習度合いを測定しているかを検証するには到達度テストとして事後テストを実施し、事後テストでの正答率から事前テストでの正答率の差を算出する差異指数(difference index)を求める必要がある。

項目困難度は文法項目に左右されるとは思うが必ずしもそうではなく、どちらかといえば項目の特性によるものという結果であった。例えば、be動詞に関しての項目は項目1、17、33、49と4項目あるが、項目困難度の推定値はそれぞれ1.48、−2.79、

－.70、.41で、値は一定ではない。また、be動詞は文法項目の中でも易しい部類だと思うが、項目1の推定値は高く、難しい項目である。よって、同じ文法項目についての項目であっても撹乱肢の出来具合などの要因によって困難度は左右されるのではないか。今回、マークシートリーダーからは正解誤答を表すデータのみ入手できたので、撹乱肢がいかに機能しているかを調査することができなかったが、今後、受験者がある撹乱肢を選んだ%などを調査する目的で撹乱肢分析（distractor analysis）を行うことで、より撹乱肢の出来具合が項目困難度に与える影響がわかるであろう。

研究目的2は項目形式の困難度の推定値はどの程度違いがあるかであった。クローズMC、間違い探しMC、並べ替えMC、和英訳MC、空欄箇所補充MC、英和訳MCの順で難しい形式であるといえる。MCの項目形式は多種あり、既存研究（David, 2007; Morimoto, 2006）の結果と同様にそれぞれ困難度は異なるといえる。Morimoto （2006）によると語彙テストの項目は文脈がよりあり、前後の内容についての情報量がよりあったほうが正答率は上がると言っている。David（2007）の結果では文脈についての情報量がよりあるクローズMCがもっとも易しい形式であった。しかし、本研究ではクローズMCがもっとも難しい形式となった。理由として考えられるのは項目形式に不慣れであったという点である。間違い探しMCは過去TOEFLにあったが、日本の教育機関ではあまり使われない項目形式である。よって、この形式に慣れていない受験者がおり、困難度が上がったのではないか。逆に並べ替えMCは選択肢を文法的に正しい順に並べ替えないと正解を得られないので、より情報処理量が多いと思うが、この形式は入学試験でも用いられるので慣れていたためクローズMCより易しい形式になったのではないかと思われる。しかし、受験者が項目形式に慣れているかなどの練習効果（practice effect）が項目形式困難度に与える影響はさらなる調査が必要であろう。

研究目的3は受験者、項目、項目形式の相は文法テストの分散に対してどの程度の割合であるかである。まず、受験者の分散成分の割合は7%を占めた。つまり、プレイスメントテストを受験した学生の英語力には若干差があったといえる。これは55点中最小値は7点、最大値は48点であり、得点差が大幅にあることからもいえる。次に、項目形式の違いによって生じた分散成分の割合は3%で、これは項目形式が6種類あり、それぞれの難易度が異なることから生じたといえる。項目の分散成分の割合は項目形式のものより大きくなった。これは、項目数は項目形式数より断然多く、また難易度の違いが項目は広範囲になっているため、よりテストパフォーマンスに影響したことを意味する。受験者と項目形式の交互作用は1%で、さほど大きくない。しかし、受験者と項目の交互作用は76%と大きい。これはある特定の相ではなく、いくつかの相の交互作用が大きくパフォーマンスに影響していると主張したBrindley and Slatyer（2002）を支持している。部分的測定テスト項目（discrete-point item）からなるテストを一般化可能性理論を用いて分析を行った既存研究の結果と比較すると、この文法テストでも受験者の分散成分の割合が少なく、項目と交互作用の割合が大きくなっており、同様の結果が得られた。Kumazawa（2009）では英語力がほぼ同じ学生に語彙診断テストとして実施したため、受験者の分散成分の割合が低く、代わりに項目の割合が大きく占めた。Brown（1999）の研究では英語運用力テストであるTOEFLの文法セクションを分析したが、やはり交互作用の割合が大きく占めている。Zhang（2006）の研究はTOEICについてで、測定している技能が聴解と読解だが、交互作用の割合が大きくなっている。この交互作用の割合の大きさが意味することは、言語テストでは多様な要因がテストパフォーマンスに影響しているということで

あろう。よって、受験者などの他の相の分散成分の割合が比較的少なくなるのであろう。今後、より多くの相（文法項目、文中にある語数等）を分析の対象とし、それらがパフォーマンスに与える影響を調査する必要がある。

研究目的4は文法テストはどの程度信頼性があるかである。表1にあるセクションごとのα係数には高低の差はあるが全体では.84で、比較的高いといえる。D研究ではセクションごとの項目数の設定を変えた場合の一般化可能性係数と信頼度指数の変動を検証した。実施時間が項目数に対して短いことを考えると全セクションを8項目で統一し、項目数を減らしたほうがよい。48項目とした場合、係数と指数はそれぞれ.81、.74となり、係数を.80以上に確保できる。この値は既存研究（Culligan & Gorsuch, 1999; Westrick, 2005）の結果より上回っている。もっとも妥当な項目数を決めるには信頼性と同時にテストの実用性（practicality）も考慮する必要がある。つまり、係数も確保でき、受験者の負担を軽減し、テスト実施時間内に終える項目数を出題すべきである。しかし、このテストはテスト開発の時間と労力はかかったが、マークシートリーダーで採点処理でき、教員が開発した学内テストなので、費用はさほどかからないため実用的ではある。

結論

本研究ではリメディアル教育の一環として行われる授業用の教科書と練習帳に準拠した目標規準準拠テストをプレイスメントテストとして実施した結果をもとにMCの機能、MCの項目形式の困難度、分散成分の割合、テストの信頼性について論じた。総括すると、この文法テストのほとんどの項目は機能していたこと、MCの項目形式には困難度のばらつきがあること、項目形式の分散成分が抽出され文法テストパフォーマンスに影響する要因であること、信頼性は比較的高かったことなどが主な結果として挙げられる。

教育的示唆としては、プレイスメントテストは集団基準準拠テストであり、事前診断テストのような目標規準準拠テストとは異なるが、その二つの実施目的を満たすテストを開発することは可能である点がまず挙げられる。この点についてはBrown（1989）でも支持する結果を示している。次に、MCの項目形式には多種あり、それぞれの困難度は異なるが、本研究での結果を踏まえ若干ではあるがテストの難易度を調節できる可能性が示された。例えば、文法テストの難易度をより上げたい場合、比較的易しい項目形式であった空欄箇所補充MCではなく、間違い探しMCを用いるということだ。最後に、カリキュラムに適したプレイスメントテストを実施するための最善策は独自にカリキュラム内容と受験者に合ったテストを開発することだと考える。その際、明確なテスト設計をし、試行テストを実施し、改良することは必須である。市販のテストを採用する場合でも必ず試行テストを行い、テストが受験者に適しているか検証することが推奨される。

本研究の欠点として挙げられるのは妥当性について具体的に言及していない点である。Kane（1992）が推奨する論証型アプローチ（argument-based approach）によると、妥当性を論証するには、観測（observation）、一般化（generalization）、外挿（extrapolation）、理論（theory-based inference）、決定（decision）について言及すべきだとある。本研究では、方法にある内容通りテスト開発過程、実施方法、実施目的などについて述べ、これは観測に相当するが、テスト細目（test specification）をもとに妥

当性論証に用いたほうがより適切ではある。一般化については、D研究で一般化可能性係数を求め、比較的値が高いことで論証できる。外挿とは文法テストにある項目が実際の授業で行う問題と類似しているかについてであるが、授業で使用する指定教材にある練習問題の種類などを調査したうえでその形式と同じ、または似たテスト項目を作成したことを一つの論証として用いることができる。理論とは文法テストが文法知識を測定していると思われる推論が妥当かということであるが、この点についてはさらなる統計分析を行い、それぞれの項目形式がどのような文法知識を測定していると思われるかの構想概念について調査する必要がある。決定とは文法テスト得点をもとに決定した事柄が妥当であったかということである。このテストはプレイスメントテストと診断テストとして実施された。プレイスメントテスト得点をもとにリメディアル教育を受けるかを判断することが妥当なのか、さらなる検証が必要である。また、プレイスメントテスト得点でリメディアル教育を受ける必要があると判断された場合、それが学習者の英語学習にどのような影響を与えるかの波及効果も調査する必要がある。

謝辞

本研究は、関東学院大学授与学習支援GPプロジェクト（「校訓に基づく入学前～卒業後までの総合支援」）の一つである「リメディアル用教材開発」の一環として実施されたテスト結果を報告したものである。貴重なコメントをくださった工学部と法学部の先生方、及び査読者の方々に感謝いたします。

参考文献

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Assessment Systems Corporation. (1996). *ITEMAN* (version 3.6) [computer software]. St. Paul, MN: Assessment Systems Corporation.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2001b). *mGENOVA* (version 2.1) [computer software]. Iowa City, IA: The American College Testing Program.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19,* 369-394.

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

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing: Collaboration and cooperation* (pp. 163-184). Ann Arbor, MI: University of Michigan Press.

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16,* 216-237.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill College Press.

Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *JALT Journal, 30,* 69-84.

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,* 7-25.

David, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing, 24,* 65-97.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing, 23,* 198-228.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26,* 507-531.

Haladyna, T. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 51,* 73-88.

Haladyna, T. (2004). Developing and validating multiple-choice test items (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26,* 219-244.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19,* 193-220.

Kumazawa, T. (2009). Revision of a criterion-referenced vocabulary test using generalizability theory. *JALT Journal, 31,* 81-100.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing, 9,* 30-49.

Lee, Y., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for new ESL writing test through G-theory. *International Journal of Testing, 7,* 353-385.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA.

Linacre, J. M. (2002). *FACETS* (Version 3.41) [computer software]. Chicago: MESA.

Morimoto, Y. (2006). Comparison between matching- and supplying-format in multiple choice vocabulary tests. *JLTA Journal, 9,* 73-85.

Norris, J., Brown, J. D., Hudson, T., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19,* 395-418.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25,* 465-493.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23,* 35-57.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1,* 147-170.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing, 8,* 23-40.

Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27,* 71-94.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing, 24,* 251-286.

Yamanishi, H. (2005). 一般化可能性理論を用いた高校生の自由英作文評価の検討. (Using generalizability theory in the evaluation of L2 writing) *JALT Journal, 27,* 169-185.

Yi'an, W. (1998). What do tests of listening comprehension test?-A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15,* 21-44.

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing, 23,* 351-369.