学内開発プレイスメントテスト得点解釈と使用の妥当性の評価について

Evaluating Validity for In-House Placement Test Score Interpretations and Uses

熊澤孝昭（くまざわ　たかあき）
Takaaki Kumazawa
関東学院大学
Kanto Gakuin University

Validity theory has evolved dramatically in the past few decades. The most prominent theory in recent years is an argument-based validity framework, proposed by Kane (1992, 2004, 2006). To evaluate test score interpretations and uses based on Kane's framework, test developers first need to provide interpretive arguments and then validity arguments by proving sound warrants for the following four inferences: (a) scoring from observation to an observed score, (b) generalization from the observed score to the universe score, (c) extrapolation from the universe score to a target score, and (d) decision from the target score to use.

In the field of language testing, a number of studies have been conducted to investigate the validity of test score interpretations and uses, especially for the ones considered to be high-stakes such as the TOEFL (Chapelle, 2008; Chapelle, Enright, & Jamieson, 2010). However, not many studies have been conducted to validate in-house placement test score interpretations and uses, and no study has evaluated the validity of such low-stakes tests using Kane's validity framework. Regardless of

whether the tests are high or low stakes, test developers need to be responsible for validating their test score interpretations and uses in order to attest to the validity.

This study uses Kane's (2006) argument-based validity framework to evaluate the validity of in-house placement test score interpretations and uses. The research questions are as follows: (a) to what extent do examinees get placement items correct and high-scoring examinees get more placement items correct; (b) to what extent are placement items consistently sampled from a domain sufficient in number so as to reduce measurement error; (c) to what extent do the difficulty of placement items match the objectives of a reading course; and (d) to what extent do placement decisions made to place examinees in their proper level of the course have an impact on washback in the course?

An in-house placement test made up of 40-item grammar, 40-item vocabulary, and 10-item reading sections was developed and administered to 428 first-year private-university students in April 2010. The item format adopted was all multiple-choice so the answer sheets could be easily scored with a reader. Based on their test scores, about 60 high-scoring students and 50 low-scoring students were placed into one of two advanced or one of two basic reading classes. The remaining students were placed into one of several intermediate classes. A 55-item grammar achievement test was administered twice (once as a pretest and then again as a posttest) to the two basic and two intermediate classes. In addition, a 51-item class evaluation survey was administered to investigate students' participation in the reading classes and to gauge students' satisfaction with the classes and study support.

Warrants for a validity argument of score inference were based on the results of the item analysis. A warrant for a validity argument of generalization inference was based on the composite generalizability coefficient of .92. A warrant for a validity argument of extrapolation inference was based on FACETS analysis, showing that difficulty estimates of learning levels were in an expected difficulty order. A warrant for validity arguments of decision inference was based on the basic-level students' score gain on an achievement test and their positive reactions to a class evaluation survey. All the validity arguments presented in this study support the validity of the placement test score interpretations and uses. However, to further improve the validity of the test score interpretations and uses, it is necessary to investigate washback effects of the placement test in the reading classes and to revise the test to make grammar, vocabulary, and reading sections with 30 items each.

本研究では論証型妥当性枠組み（Kane, 2006）を用いて大学一年生必修リーディング科目用のプレイスメントテスト得点の解釈と使用についての妥当性を評価した。その結果、項目分析の結果は良好であった（得点化の妥当性論証の論拠）。測定領域より項目は抽出されており、多変量一般化可能性理論を用いて合成一般化可能性係数を求めた結果、.92と高い結果が得られた（一般化の妥当性論証の論拠）。FACETS分析を用いて学習レベル困難度推定値を求めた結果、難易度推定値は予想通りであった（外挿の妥当性論証の論拠）。基礎クラスと中級クラ

ス履修者を対象に、到達度テストを事前・事後テストとして実施し、その平均値差を検証した結果、基礎クラス履修者は得点を上昇させることがでた。また、授業評価アンケート結果によると、学習支援など指導が効果的だったということがわかった（決定の妥当性論証の論拠）。

> Validation is simple in principle, but difficult in practice. The argument-based framework provides a relatively pragmatic approach to validation. (Kane, 2012, p. 15)

日本の高等教育では集団基準準拠テスト（norm-referenced tests）の一種である一般入試やAO入試を含む入学試験制度の多様化などの要因から、学生の英語力の差が学部内でも顕著になり、学習者の英語力により適した指導ができるよう習熟度別カリキュラムが重要視されるようになった。それに伴い、クラス分けに用いられるプレイスメントテスト（placement tests）の関心が高まり、その研究（Culligan & Gorsuch, 1999; Westrick, 2005等）が行われてきた。また、よりカリキュラムの内容を反映した目標規準準拠テスト（criterion-referenced tests）の特性を持ち合わせたテスト開発についての研究も行われてきた（Brown, 1989; 熊澤, 2010）。

妥当性の評価方法は様々であり、妥当性理論も進化してきた。市販のテストや学内開発プレイスメントテストが多用される中、実際にそれらのテスト得点の解釈（interpretations）と使用（uses）がどの程度妥当であったかを評価する研究はさほど行われてこなかった。しかし、実施されたテスト得点の解釈と使用が妥当であったということを検証することは開発者の義務ではないかと考える。また、妥当性を論証することにより改善点が明白になり、さらに妥当性が高いものに改訂することができる。よって、本研究では妥当性理論の変遷について述べ、Kane（1992, 2004, 2006）が考案した論証型妥当性枠組み（argument-based validity framework）をもとに、ある大学で一般教養一年生必修リーディング科目のクラス分けを目的に開発されたプレイスメントテスト得点の解釈と使用の妥当性を評価する。

妥当性について

妥当性は「テスト得点の解釈と使用がいかに論拠と理論によって支持されるか」と定義されている（American Educational Research Association/American Psychological Association/National Council on Measurement in Education, 1999, p. 9）。心理測定の分野では妥当性について書かれた文献は多々ある（e.g., Lissitz, 2009; Wainer & Braun, 1988）。その中でも妥当性という概念がいかに進化してきたかを詳述しているのはKane（2001）である。それによると妥当性理論の変遷は（a）基準的妥当性（criterion-based validity）と内容的妥当性（content-based validity）、（b）構成概念妥当性（construct validity）、（c）妥当性の現代的見解（current view of validity）と三つの時期を経ている。第一の時期では、基準的妥当性を検証するにはあるテストと外的基準となるテストとの相関関係を検証する。例えば、TOEFLとあるテストとの相関係数が高いことにより妥当性が高いとする。しかし、外的基準となるテストが無く基準的妥当性を検証することが困難な場合、テスト項目と測りたい領域との一致度を専門家が判断することで内容的妥当性が検証されていた。

第二の時期では、Cronbach and Meehl（1955）があるテストが何の構成概念をどの程度測っているかを検証する構成概念妥当性という概念を提唱した。その後、複数の構成概念とテスト形式の相関関係を算出することで妥当性を検証する多特性多方法（multitrait-multimethod; Campbell & Fiske, 1959）など構成概念妥当性を検証する方法が考案された。その後、Cronbach（1971）はテスト自体が妥当であるかではなくテスト得点の解釈と使用が妥当であること、テスト得点の解釈と使用を妥当化するには複数の証拠（evidence）が必要であることを説いた。また、Cronbach（1980）とMessick（1981）は妥当性の評価方法と概念を確立すべきだと主張した。

第三の時期では、Messick（1989）が単一的構成概念妥当性（unitary construct validity）を発表し、証拠基準（evidential basis）と影響基準（consequential basis）を含むテストの正当化（test justification）、および解釈と使用を含むテストの機能（test function）の二つの関連した相（facet）を妥当性の枠組みとした。また、構成概念妥当性には内容的側面（content aspect）、本質的側面（substantive aspect）、構造的側面（structural aspect）、一般化的側面（generalizability aspect）、外的側面（external aspect）、影響的側面（consequential aspect）の六つの側面があるとした（Messick, 1995）。単一的構成概念妥当性は特に信頼性の概念とテストの影響（consequence）について妥当性理論に大きな影響を与えた。以前は、信頼性は妥当性の必要条件として独立したかたちで捉えられていたが、信頼性は一般化という概念に代わり妥当性の一部として捉えられるようになった。テスト得点は受験者に影響を与え、利害が大きい（high-stakes）テストほど影響も大きくなる。Messick（1989）はこのような影響も妥当性の枠組みに入れることを主張した。

Kane（1992, 2004, 2006）は妥当化の実践的でなおかつ系統的な方法を示すため論証型妥当性枠組み（argument-based validity framework）を考案した。Kane（2006）によると妥当性を評価するためには解釈的論証（interpretive argument）と妥当性論証（validity argument）を提示する必要がある。前者はテスト結果の解釈と使用に内在する連鎖的な推論（inference）に対する前提条件を明示することであり、後者は複数の分析結果を用いてその解釈的論証を評価するための論拠（warrant）を提示することである。図1にあるように、観測（observation）から観測得点（observed score）、観測得点から測定領域得点（universe score）、測定領域得点から目標得点（target score）、目標得点から用途（use）をそれぞれ結ぶ推論は得点化（scoring）、一般化（generalization）、外挿（extrapolation）、決定（decision）と四つがある。プレイスメントテストの解釈的論証の場合、得点化についての推論は、ある項目がいかに採点されるかを述べ、それが適切であることを前提とする。一般化についての推論は、測定領域を代表する項目が抽出されており（representative of the universe）、また誤差を最小に抑えられるよう項目数は十分であることを前提とする。外挿についての推論は、項目は授業目標の到達度を示すのに適切であり、誤差はテスト方法などの要因の影響をさほど受けていないことを前提とする。決定についての推論は、低得点者に適切な難易度の授業目標を学習させることで、より適切な成績評価ができるなど利益があることを前提とする。

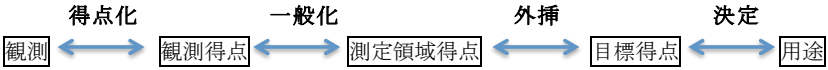

図1. 解釈的論証における推論の連鎖について
（Kane, Crooks, & Cohen, 1999, p. 9参考）

プレイスメントテストをもとに、いかに解釈的論証を評価するための妥当性論証の論拠を提示するかを一例として挙げる（Kane, 2006）。得点化の妥当性論証の論拠としては古典的テスト理論であれば項目容易度（item facility）と項目弁別力（item discrimination）、項目応答理論であれば項目困難度（item difficulty）と適合度（fit）を含む項目分析をもとにすべての項目は正確に得点されていることが論拠となる。一般化の妥当性論証の論拠としては内部一貫性信頼性クロンバック α 係数を用いて論拠とすることができる。または、一般化可能性理論（Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001a; Shavelson & Webb, 1991）を用い、一般化可能性研究（generalizability study [G研究]）で各相の分散成分を求め、それをもとに決定研究（decision study [D研究]）で集団基準準拠テストであれば一般化可能性係数（generalizability coefficient）を、目標規準準拠テストであれば信頼度指数（dependability index）を論拠とする。外挿の妥当性論証論拠を提示するには分析型と実証型の方法がある。分析型の方法では、測定領域と目標領域（target domain）との関係を調べるため、どのように問題を解いたかを発話させ、それを分析すること（think-aloud protocol analysis）で論拠を得る。実証型の方法では従来の基準的妥当性や因子分析の結果を論拠に用いることができる。決定の妥当性論証にはテスト得点をもとに下した決定の影響や波及効果（Cheng & Watanabe, 2004）について述べ、下した決定が適切であったことを示すことを論拠とする。

言語テストの分野では妥当性について記述した文献は多々ある（Cumming & Berwick, 1996; Kunnan, 1998; Weir, 2005等）。Bachman and Palmer（1996）は信頼性（reliability）、構成概念妥当性（construct validity）、真正性（authenticity）、相互性（interactiveness）、影響（impact）、実用性（practicality）というテストの質からなるテストの有用性（test usefulness）という概念を発表した。Bachman（2005）はKaneの論証型妥当性枠組みとBachman and Palmer（1996）のテストの有用性を踏襲し、評価妥当性論証（assessment validity argument）と評価用途論証（assessment utilization argument）からなる評価使用論証（assessment use argument）という概念を発表した。評価妥当性論証とはKane（2006）でいう解釈に関する論証であり、得点化から外挿までの推論と同意である。評価用途論証はKane（2006）では決定の推論にあたり、Bachmanはさらに決定の推論を細分化し、関連性（relevance）、用途（utility）、意図した影響（intended consequences）、充足（sufficiency）とした。また、論拠に対する反論（rebuttal）を挙げることも重要視した。関連性とは得点解釈が得点をもとに下す決定と関連性があるかをいう。用途とは得点解釈が意図した決定を下すのに役に立つかをいう。意図した影響とは得点解釈をもとに下した決定が受験者、教育機関、会社など社会全体に有益な影響を及ぼすかをいう。充足とは行われた評価が決定を下すのに十分な情報を提供するかをいう。よって、Bachman（2005）の枠組みは得点化から充足まで七つの推論（得点化、一般化、外挿、関連性、用途、意図した影響、充足）にもとづいている。

Bachman and Palmer（2010）はさらに新たな論証型妥当性枠組みを考案した。まず、タスクをもとにパフォーマンスが観測される。そして、評価記録（assessment records）は一貫性があるかの論拠と反論を検証する。次に、解釈は意義があるか（meaningful）、公平であるか（impartial）、一般化できるか（generalizable）、関連性があるか（relevant）、充足するものか（sufficient）の論拠と反論を検証する。次に、決定は価値があるか（values sensitive）、公平か（equitable）の論拠と反論を検証する。最後に影響は有益なものか（beneficial）の論拠と反論を検証する。よって、この枠組みは一貫性から有益性まで九つの点（評価記録に一貫性があるか、解釈に意義・公平性・一般化・関連性・充足があるか、決定に価値・公平性があるか、影響が有益か）についての論拠と反論を検証することになる。

Chapelle（2008）は主にKane（2006）の枠組みを踏襲し、測定領域記述（domain description）、評価、一般化、説明（explanation）、外挿、用途（utilization）の推論について論証すべきとしている。測定領域記述とは測っている概念を記述することである。評価と用途はKane（2006）では得点化と決定と同意である。説明とは分析結果をもとに測っている概念はなにかを明示することである。よって、Kane（2006）の枠組みに測定領域記述と説明を加え六つの推論（測定領域記述、評価、一般化、説明、外挿、用途）にもとづくのがChapelle（2008）の枠組みである。

表1. 妥当性枠組みの比較について

	得点解釈に関する推論	得点使用に関する推論
Kane (2006)	得点化、一般化、外挿	決定
Bachman (2005)	得点化、一般化、外挿	関連性、用途、意図した影響、充足
Bachman & Palmer (2010)	評価記録は一貫性があるか、解釈には（意義があるか、公平であるか、一般化できるか、関連性があるか、充足するものか）	決定には（価値があるか、公平であるか）、影響は（有益か）
Chapelle (2008)	測定領域記述、評価、一般化、説明、外挿	用途

言語テストの分野ではテスト得点解釈と使用の妥当性を評価した実証研究がある。まず、Pardo-Ballester（2010）はBachman（2005）の枠組みを用いてウェブ上で実施するスペイン語リスニングテストの有用性（usefulness）の特性である信頼性、構成概念妥当性、真正性の論拠を挙げている。信頼性にはラッシュ信頼性係数（Rasch reliability coefficient）、項目適合度、受験者適合度、主成分分析（principal component analysis）の結果を提示している。構成概念妥当性にはタスクの難易度を三レベルに区分けし、項目応答理論にもとづくそれぞれのレベルの項目困難度平均値差の結果を挙げている。真正性はテスト項目が授業で行った問題と類比しているかなどを調査するアンケートの結果を提示した。

Chapelle, Chung, Hegelheimer, Pendar, and Xu（2010）ではコンピュータ上で受験できる文法テストを開発し、Chapelle（2008）の妥当性評価の枠組みである六つの解釈的論証の推論のうち測定領域記述、評価、一般化、説明、外挿の五つの妥当性検証の論拠を挙げている。領域記述に対しての妥当性検証の論拠にはその文法テストの内容、実施目的、開発過程などの記述を用いている。Chapelle（2008）がいう評価というのはKane（2006）がいう得点化にあたるが、この論拠として採点方法の記述を用いている。一般化の論拠ではクロンバックα係数が、説明の論拠には第二言語習得理論にもとづく文法習得過程と項目困難度との一致が、そして外挿の論拠には他の文法テストとの高い相関関係および習熟度別の三グループの平均値に有意な差があったことが挙げられている。

Beglar（2010）はMessick（1995）の単一的構成概念妥当性の枠組みを用いてラッシュモデル（Rasch model）から得られた結果をもとに語彙サイズテスト（Vocabulary Size Test [VST]）得点解釈の妥当性を評価している。内容的側面には項目困難度（item difficulty）と項目適合度を用いている。本質的側面としては語彙の頻度レベル（frequency level）とそれぞれのレベルの項目困難度が一致しているかと能力別に語彙数を調査した結果を提示している。構造的側面としてはVST項目の一次元性（unidimensionality）についての結果を示した。一般化的側面としてはラッシュ信頼性係数を提示した。

Koizumi et al.（2011）は日本人学習者対象に文法診断テストを開発し、項目弁別力、錯乱肢分析（distracter analysis）、信頼性、名詞節の難易度の結果を妥当性論証の論拠として挙げた。項目弁別力は比較的高い、選択肢は機能している、信頼性は高い、名詞節の難易度はほぼ予想通りであったことからそのテスト得点解釈の妥当性を評価している。

研究課題

妥当性理論は進化を遂げ、また言語テストの分野でもTOEFLなどのテスト得点解釈と使用の妥当性の評価がされてきた（Chapelle, 2008; Chapelle, Enright, et al. 2010）。しかし、教員がカリキュラムの内容を反映するプレイスメントテストを開発し、その妥当性を評価した研究はさほど行われてこなかった。ましてや、現代の妥当性理論であるKane（2006）の妥当性枠組みを用いてプレイスメントテスト得点解釈と使用の妥当性を評価した論文はほぼ皆無であろう。

本研究ではKane（2006）の枠組みを用いるが、その理由は明確さと簡潔さである。まず、得点化から決定までの推論である一連の関連性が非常に明確であることが挙げられる。つまり、解答が得点化され観測得点となり、その観測得点が一般化され測定領域得点となり、その測定領域得点が外挿され目標得点となり、その目標得点をもとに決定が下され使用されるという一連の推論の関係性が明確であるということである。次に、論証型妥当性枠組みは簡潔であるべきであり、Kaneのものがもっとも簡潔であるといえる。枠組みが複雑になることで、論拠をもとに妥当性検証をすることが教員にとってより困難になるのではないか。妥当化を行うことは容易いことではないので、枠組みが明確で簡潔なほうが教員にとって取り組みやすいということである。

本研究の目的は大学一年生対象一般教養必修リーディング科目のクラス分けを目的に開発されたプレイスメントテスト得点解釈と使用の妥当性をKane（2006）の論証

型妥当性枠組みを用いて評価することである。つまり、得点化、一般化、外挿、決定の推論に付随する前提条件（assumption）を挙げることにより解釈的論証を行い、それに対する論拠をもとに妥当性論証をすることである。以下にある四つの推論に対する解釈的論証の前提条件を研究課題とする。

1. プレイスメントテスト項目はどの程度意図した正解を受験者が解答でき、高得点者がより正解できるか。
2. プレイスメントテスト項目はどの程度一貫して測定領域から抽出されていて、また項目数は誤差が最小になる程度あるか。
3. プレイスメントテストにおける学習レベルはどの程度リーディング科目の授業目標と難易度が一致しているか。基礎レベルの学習目標は文法であれば中等教育で学習した事項、語彙であれば2000語程度、読解であれば基礎的な英文理解ができる程度とする。
4. プレイスメントテスト得点をクラス分けの判断材料として使用し、習熟度別クラス編成し、受講者のレベルにより適した授業内容を提供することにより、どの程度正の波及効果が生じるか。

方法

対象者

本研究の対象者は関東地方にある私立大学法学部に在籍するプレイスメントテストを受験した428名の2010年度入学一年生である。外国人学習者一名を除いてはすべて日本人学習者であった。法学部開講必修英語は一年次には四科目あり、そのうちの二科目はリーディング、それ以外はコミュニケーション中心の授業であった。前者の科目のみ習熟度別に編成され、プレイスメントテスト高得点者の60名程度が上級クラスである二クラス、低得点者の50名程度が基礎クラスである二クラスへ得点順に振り分けられた。それ以外の受験者は中級クラスである十クラスのいずれかに自動的に振り分けられた。上級クラスでは速読と語彙学習が中心で、基礎クラスではリメディアル教育の一環として中等教育で学習した文法事項を復習するのが主な授業内容であった。基礎クラスの学習者はさらなる教育支援が必要と思われるので授業外でも教員から指導を受けられる体制が整っていた。

プレイスメントテスト

プレイスメントテストを開発した理由は、一年生必修リーディング科目を習熟度別にクラス編成するためである。学部で独自にテスト開発をすることにより、よりカリキュラムを反映する内容になると考えた。また、履修生に迅速に指定クラスを掲示するためマークシートリーダーで採点できるテストが必要であった。リーディング科目のプレイスメントテストということで、最終的には文法（k = 40）、語彙（k = 40）、読解（k = 10）の三つのセクションから成るテストとした。

文法項目を作成するにあたり、リーディング科目では中等教育で学習したことをもとにさらに読解力を伸ばすということが目標であることを考慮し、まず中学校・高等学校の教科書（霜崎他, 2006, 高橋他, 2005）にあるすべての重要文法項目を調べ、リス

トを作成した。中学校および高等学校の学習文法事項を参照するためにその教科書を用いた理由は一般的に広く用いられていると思われるため、また教科書によって扱う学習事項が若干異なるので統一したほうがいいと判断したためである。その後、重要文法事項を用いて英文を作成し、次に文法事項の個所を空欄にし、そして正解を含む選択肢を四つ作成した。よって、項目形式は多肢選択式項目である。2006年に実施した試行テストでは70項目を20名の受講者が受験し、項目容易度の値が極端に高いまたは低い項目で、項目弁別力の値が低い30項目を除外した。特に、中学校一年生で学習する文法事項は受験者にとって易しすぎたので削除した。残った40項目は143名の受講者対象に試行テストとして再度実施され、その結果をもとに改良したものが本試験で用いられた。本試験の前に再度四人の教員が項目の文言などをみて修正を加えた。2007年以降毎年ほぼ同じ項目を用いこのプレイスメントテストを実施しているが、2010年度の実施では文言と選択肢を何点か変更した。例題は以下の通りである。

Hi, I (　) Ken.

1. am　2. are　3. is　4. be

受験者は英文を読んで空欄個所にあてはまるもっとも文法的に正しいものを選ぶ形式であり、正解は1である。配点は各設問につき一点である。

語彙セクションの開発過程では、まず大学英語教育学会基本語リストJACET　List of 8000 Basic Words（大学英語教育学会基本語改訂委員会, 2003）より1,000語から2,000語の頻度にある48語を選んだ。この語彙リストを選んだ理由としては日本人学習者に即しているという点である。まず、受験者の語彙レベルを把握するため20名の受講者にテストを実施した。その結果、得点結果にはばらつきがさほどみられなかったため、本試験では1,000から3,000語の頻度範囲から40語を選んで40項目を作成した。そして、その40項目は試行試験として実施され、その結果をもとに改良を加えたものが本試験で用いられた。また、文法セクション同様2007年以降、複数の文言と選択肢に変更を加えた。例題は以下の通りである。

Bring

1. 送る　2. 持ってくる　3. 鳴る　4. 購入する

受験者が英単語を見てもっとも意味が近い和訳を選ぶという形式であり、正解は2である。配点は各設問につき一点である。

読解セクションは2008年より追加された。その主な理由はやはりリーディング科目のプレイスメントテストであるにも関わらず、読解項目がなく語彙と文法のみだったためである。本文は基礎クラスで使用されている教科書にある123語から成るものと、上級クラスで用いられている教科書にある341語から成るものを選んだ。低得点者は基礎クラスで用いている教科書にある本文の理解度が低く、高得点者はほぼ理解できると想定した。それぞれの本文に対して五項目ずつ英文で書かれた多肢選択式項目を作成した。項目数を増やしたかったのだが文法と語彙項目数だけですでに80項目あり、試験時間も45分であるので断念した。文法と語彙項目を減らし読解項目を増やすことも検討したが、項目を減らすことによって、前年度の結果と比較できなくなるとの理由で行わなかった。四名の教員と一名の英語母語話者教員が英語を確認し、不自然な文言などを修正した。試行テストは行われなかった。例題は以下の通りである。

How do they travel?

1. by plane 2. by bus 3. by car 4. by train

受験者は本文と設問を読み、その解答を選択肢から選ぶという項目形式である。配点は各設問につき二点である。その理由は満点を100点にするため、またリーディング科目のプレイスメントテストとして用いるのでリーディングの配点を大きくしたいためである。

実施の手順

教員と職員とでテスト問題やマークシートを用意し、教員が試験を監督した。試験監督者は監督の手引きを試験前に受験者に読み上げ、このプレイスメントテストの結果はクラス分けと研究目的で使用されるとの説明がなされた。試験時間が45分に保たれるように監督者は留意した。マークシートはリーダーによりデータ化され、エクセルで採点がされた。未解答は不正解とみなした。テスト結果は受験者へは返却されなかった。

上級クラス、中級クラス、基礎クラスともに通年でおよそ30回の授業が行われた。上級クラスでは速読と語彙学習を中心に授業が行われた。複数ある中級クラスの内、二クラスのみ、プレイスメントテストとは異なる55項目から成る文法到達度テストを2010年四月事前テストとしてと、2011年一月事後テストとして実施した。このテストは基礎クラスでの学習内容の到達度を測る目的とプレイスメント目的で開発された（熊澤, 2010）。また、最終授業内で51項目から成る授業評価アンケートを実施した。六件法（（6）つよくそう思う、（5）そう思う、（4）まあそう思う、（3）あまりそうは思わない、（2）そうは思わない、（1）まったくそうは思わない）を採用し、授業の満足度、履修者の参加度合い、学習支援内容などを評価した。例えば、項目1は「この授業でよくできたのは学習支援室でのサポートがあったからだ」である。中級クラス授業内では文法学習や訳読などが行われた。

基礎クラスでは中級レベルの二クラスと同様に文法到達度テストが実施された。しかし、基礎クラスのみ文法到達度テストの結果が成績に考慮された。また、同クラスでは授業評価アンケートも最終授業内で実施された。基礎クラス授業内では基礎的な文法を学習すること、英文を訳すこと、学習内容の習熟をみるため小テストなどが行われた。授業外では文法練習帳を課題として出題し、わからない点や答え合わせなど教員から指導を受ける支援があった。

分析の手順

得点化の妥当性論証の論拠を提示するため、古典的テスト理論および項目応答理論（item response theory）を用いた。古典的テスト理論にもとづく分析では項目容易度および項目弁別力の値を求めた。項目応答理論を用いた分析は以下で述べる。一般化の妥当性論証の論拠を提示するため、多変量一般化可能性理論（multivariate generalizability theory）を用いて文法、語彙、読解セクションそれぞれの受験者（p）、項目（i）、受験者X項目（$p \times i$）の分散成分推定値を一般化可能性研究（generalizability study [G研究]）で求め、その結果をもとに各セクションの一般化可能性係数（generalizability coefficient）と合成一般化可能性係数（composite

generalizability coefficient）を決定研究（decision study [D研究]）で求めた。よって、このG研究のデザインは$p \bullet X I^{o}$でD研究は$p \bullet X I^{o}$である。•は全セクションともに受験者は同じであることを意味し、^{o}は各セクションに設けられた項目は異なるということを意味する。多変量一般化可能性理論を用いた理由はこのテストは三セクションから成るので、各セクションの項目数を変更することにより、いかに一般化可能性係数が変動するかを検証したいがためである。mGENOVA（Brennan, 2001b）をG研究とD研究に用いた。多変量一般化可能性理論についてはBrennan（2001a）が解説書を書いている。また、言語テストの分野でも多変量一般化可能性理論を用いた論文がある（Lee, 2006; Sawaki, 2007; Xi, 2007等）。

得点化および外挿の妥当性論証の論拠を提示するため、項目応答理論の一種である多相ラッシュモデル（multifaceted Rasch model; Linacre, 1989）をもとにした統計ソフトであるFACETS（Linacre, 2002）を用いた。インフィット平均二乗に用いる基準は0.80～1.20（Bond & Fox, 2001）とし、この項目の値が基準を満たしている場合、一次元性が保たれたと仮定することができる（Bond & Fox, 2001）。分析対象とした相は受験者能力、項目困難度、学習内容困難度で、それぞれの推定値を求めた。得点化の論拠には項目困難度推定値とその標準誤差を用いた。外挿の論拠としては一次元性と学習レベル困難度を用いた。学習レベル困難度推定値は高校文法レベルより中学文法レベル、JACET3000語彙レベルよりJACET1000語彙レベル、上級読解レベルより基礎読解レベルのほうがより低くなると仮定された。ラッシュモデルについてはBond and Fox（2001）が解説書を書いており、言語テストの分野でも多相ラッシュモデルを用いた研究は多々ある（e.g., Coniam, 2008; 熊澤, 2010; Schaefer, 2008）。

そして、決定の論拠を求めるため、プレイスメントテスト、文法到達度テスト、および授業評価アンケートの記述統計を求めた。まず、習熟度別にクラスが編成されているかを確認するため全体、上級クラス1、上級クラス2、中級クラス、基礎クラス1、基礎クラス2ごとのプレイスメント得点平均値と標準偏差値を求めた。この記述統計では読解項目のみ配点を各項目二点とした。次に、中級群と基礎群ではどの程度教育効果があったかを検証するため事前テスト時と事後テスト時の文法到達度テスト得点平均値と標準偏差値を求めた。最後に、中級群と基礎群がどのように授業を評価したかを調査するために実施されたアンケートのそれぞれの平均値と標準偏差値を求めた。

結果

表2は古典的テスト理論とFACETS分析による項目分析の結果についてである。項目容易度（IF）の値には大きなばらつきがみられる。例えば、項目41の値は.99で項目28の値は.15となっている。プレイスメントテストでは、値が.30から.70の範囲にある項目がよいとされる（Brown, 2005）。その基準をもとに項目の良否を判断すると表中にある太字で表記されている28項目は値が適切ではなく、受験者にとって易しい、または難し過ぎるということになる。しかし、高得点者が低得点者より正解できるかを示す項目弁別力をみると、不良項目と判断される基準である.19以下の項目は表中にある太字で表記されている四項目のみである。

項目困難度推定値のばらつきは大幅にあり、分離指数（separation index）は6.87でラッシュ信頼性係数は.98であった。これは項目困難度推定値の最大値と最小値がそ

れぞれ2.33と-3.79で、項目困難度推定値のばらつきが広範囲になっていることを意味する。FACETS分析の利点は図2にあるように項目困難度と受験者能力値などの推定値が同一の間隔尺度上に示され、視覚的にその推定値のバランスを検証できることである。大半の受験者能力推定値は－1.00から1.00の範囲にあり（74％）、またその範囲を測る項目も多くある（82％）。しかし、能力推定値が1.00以上となった受験者は若干いる（20％）のにもかかわらず1.00以上の能力推定値を測る項目がさほどない（7％）。また、能力推定値が-2.00以上になった受験者はいないが、項目41と項目44は大幅にその推定値を下回っているので削除してもいいかもしれない。標準誤差は最大値が項目41の.50で、項目が易しく99％の受験者が正解したため、この項目の推定値を算出するのにデータが足りず、誤差が大きくなったことを示す。他の値は問題なく推定値の誤差はさほどない。インフィット平均二乗は基準である0.80～1.20（Bond & Fox, 2001）に項目67以外のすべての項目の値が収まりモデルに適合しているといえる。項目67の出題単語はacquireで選択肢は1から4の順に、「合唱する、必要とする、取得する、支払う」で正解は3であるが、高得点者が不注意でrequireと間違い2を選択してしまったなどの理由が挙げられる。しかし、アウトフィット平均二乗ははずれ値に敏感なため0.80~1.20の基準の範囲以外の項目が14項目もある。特に問題視すべきなのは項目50で、項目弁別力が-.05、アウトフィット平均二乗の値が2.20になった。これは、高得点者が誤答してしまう確率があり、選択肢などを改訂すべきであることを示唆している。この項目内容をみると英単語がruleで、選択肢が「役割、規定、命令、事実」となっており、正解は規定である。高得点者も不注意でroleと見違えてしまい、選択肢の役割を選んでしまったことが推測される。一方、受験者については、インフィット平均二乗値が0.80~1.20の基準からはずれ、不適合（misfit）と判定された者は16名（4%）であった。その内14名は1.20以上になり適合不足（underfit）と判定され、集中力が持続しなかったなどの理由が挙げられる。

表2. 古典的テスト理論とFACETS分析による項目分析の結果について（*N* = 428）

		古典的テスト理論		FACETS分析			
項目	項目内容	IF	ID	Diff	SE	Infit MS	Outfit MS
1	過去形不規則変化 (中)	**.81**	.37	-0.74	0.13	1.00	0.90
2	過去進行形 (中)	.52	.25	0.79	0.10	1.20	**1.30**
3	接続詞when (中)	**.81**	.45	-0.74	0.13	0.90	0.80
4	～あるthere (中)	**.78**	.45	-0.54	0.12	0.90	0.90
5	接続詞and (中)	**.80**	.42	-0.71	0.13	0.90	**0.70**
6	命令形Will (中)	.58	.47	0.51	0.11	0.90	0.90
7	比較 (中)	.65	.43	0.20	0.11	1.00	1.00
8	受身 (中)	**.76**	.39	-0.43	0.12	1.00	0.90
9	have to=must (中)	.51	.25	0.87	0.10	1.10	1.20
10	仮定法 (中)	.67	.49	0.06	0.11	0.90	0.90

項目	項目内容	古典的テスト理論 IF	古典的テスト理論 ID	FACETS分析 Diff	FACETS分析 SE	FACETS分析 Infit MS	FACETS分析 Outfit MS
11	疑問詞How long (中)	**.82**	.45	-0.83	0.13	0.90	0.80
12	現在完了形・経験 (中)	**.73**	.37	-0.24	0.12	1.00	0.90
13	形式主語 (中)	.67	.50	0.09	0.11	0.90	0.80
14	so that構文 (中)	.59	.29	0.51	0.11	1.10	1.10
15	動名詞形容詞的用法 (中)	.60	.40	0.43	0.11	1.00	1.00
16	過去分詞形容詞的用法 (中)	**.23**	.21	2.33	0.12	1.10	**1.40**
17	関係代名詞Who (中)	.67	.53	0.12	0.11	0.90	0.80
18	関係代名詞Which (中)	.69	.38	-0.01	0.11	1.00	1.00
19	how to (中)	**.81**	.47	-0.69	0.13	0.90	0.80
20	because of (中)	.52	.31	0.83	0.10	1.10	1.00
21	不定詞の名詞的用法 (高)	.64	.52	-0.68	0.11	0.90	0.90
22	現在完了形 (高)	.39	.36	0.52	0.11	1.00	1.10
23	関係代名詞what (高)	.56	.47	-0.29	0.11	1.00	0.90
24	分詞の形容詞的用法 (高)	.62	.49	-0.59	0.11	0.90	0.90
25	seemの使い方 (高)	.54	.54	-0.20	0.11	0.80	0.80
26	分詞 (高)	**.73**	.47	-1.15	0.11	0.90	0.80
27	受動態の完了形 (高)	.39	.30	0.53	0.11	1.10	1.10
28	分詞構文 (高)	**.15**	.34	2.03	0.15	0.90	1.00
29	時制の一致 (高)	**.71**	.36	-1.05	0.11	1.00	1.10
30	with+名詞+形容詞の使い方 (高)	.32	.38	0.88	0.11	1.00	1.00
31	関係副詞(非制限用法) (高)	.50	.42	-0.04	0.10	1.00	1.00
32	形式目的語 it (高)	.43	.21	0.29	0.11	1.20	1.20
33	強調構文 (高)	.41	.32	0.39	0.11	1.10	1.10
34	過去完了進行形 (高)	.51	.43	-0.05	0.10	1.00	1.00
35	未来完了形 (高)	.66	.41	-0.80	0.11	1.00	0.90
36	未来進行形 (高)	**.19**	**.18**	1.64	0.13	1.10	**1.30**
37	倒置 (高)	**.29**	.23	1.01	0.12	1.20	**1.30**
38	助動詞 would の用法 (高)	**.75**	.40	-1.29	0.12	1.00	1.00
39	不定詞の完了形 (高)	.55	.39	-0.28	0.11	1.00	1.00
40	動名詞の意味上の主語 (高)	.34	.26	0.78	0.11	1.10	1.20
41	Country (139)	**.99**	**.14**	-3.79	**0.50**	1.00	**0.50**

項目	項目内容	古典的テスト理論 IF	古典的テスト理論 ID	FACETS分析 Diff	FACETS分析 SE	FACETS分析 Infit MS	FACETS分析 Outfit MS
42	Problem (179)	**.95**	.22	-1.92	0.21	1.00	0.90
43	Bring (281)	**.79**	.42	-0.32	0.12	0.90	0.80
44	Company (294)	**.97**	.24	-2.75	0.31	1.00	**0.50**
45	Century (335)	**.95**	.31	-1.97	0.22	0.90	**0.60**
46	National (368)	.61	.35	0.66	0.11	1.10	1.10
47	Certain (428)	.54	.55	1.03	0.11	0.90	0.80
48	Consider (453)	.67	.40	0.38	0.11	1.00	0.90
49	Law (489)	.64	.39	0.55	0.11	1.00	1.00
50	Rule (495)	**.89**	**-.05**	-1.15	0.16	1.10	**2.20**
51	Population (558)	.59	.47	0.79	0.11	0.90	0.90
52	Various (647)	**.74**	.43	0.01	0.12	0.90	0.80
53	Development (786)	**.78**	.45	-0.24	0.12	0.90	0.80
54	Standard (825)	**.93**	.24	-1.67	0.19	1.00	1.00
55	Industry (873)	.69	.53	0.31	0.11	0.90	0.80
56	Reduce (878)	.66	.44	0.44	0.11	1.00	0.90
57	Represent (956)	.55	.36	0.99	0.11	1.10	1.10
58	Particular (966)	.66	.36	0.43	0.11	1.10	1.10
59	Tend (986)	.70	.48	0.25	0.11	0.90	0.90
60	Worth (994)	.52	.58	1.10	0.10	0.80	0.80
61	Solution (1186)	.50	.54	0.13	0.10	0.90	0.90
62	Available (1299)	.48	.59	0.24	0.11	0.80	0.80
63	Prison (1336)	**.73**	.28	-1.00	0.11	1.10	1.10
64	Environmental (1356)	.65	.50	-0.59	0.11	0.90	0.80
65	Victim (1486)	.47	.50	0.25	0.11	0.90	0.90
66	Status (1571)	.64	.20	-0.55	0.11	1.20	**1.40**
67	Acquire (1646)	.46	**.01**	0.34	0.11	**1.30**	**1.40**
68	Typical (1648)	.66	.42	-0.64	0.11	1.00	0.90
69	Poverty (1851)	.46	.36	0.33	0.11	1.10	1.10
70	Legal (1874)	.51	.39	0.09	0.10	1.00	1.00
71	Promote (1898)	.41	.48	0.54	0.11	0.90	0.90
72	Wage (1949)	.43	.51	0.46	0.11	0.90	0.90

項目	項目内容	古典的テスト理論 IF	古典的テスト理論 ID	FACETS分析 Diff	FACETS分析 SE	FACETS分析 Infit MS	FACETS分析 Outfit MS
73	Convince (1964)	.38	.31	0.73	0.11	1.10	1.10
74	Actual (1989)	.54	.48	-0.08	0.11	0.90	0.90
75	Interpret (2000)	.33	.28	0.96	0.11	1.10	1.20
76	Criminal (2003)	.50	.43	0.17	0.10	1.00	1.00
77	Discrimination (2411)	.50	.30	0.19	0.11	1.10	1.10
78	Superior (2622)	.47	.41	0.30	0.11	1.00	1.00
79	Punish (2859)	.51	.56	0.09	0.10	0.90	0.80
80	Equality (2974)	.62	.55	-0.40	0.11	0.90	0.80
81	基礎	.58	.54	-0.38	0.11	0.90	0.90
82	基礎	.55	.50	-0.24	0.11	0.90	0.90
83	基礎	.50	.56	-0.02	0.10	0.90	0.90
84	基礎	.52	.50	-0.08	0.10	1.00	1.10
85	基礎	.30	.45	0.99	0.11	1.00	1.00
86	上級	.49	.45	-0.36	0.11	1.10	1.10
87	上級	.52	.40	-0.50	0.10	1.10	1.10
88	上級	**.26**	.27	0.80	0.12	1.20	**1.40**
89	上級	**.25**	.34	0.86	0.12	1.10	**1.30**
90	上級	**.26**	.38	0.81	0.12	1.10	**1.30**

注.（中）＝中学文法学習レベル、（高）＝高校文法学習レベル、（139）＝JACET8000による出現頻度、基礎＝基礎読解クラスレベル、上級＝上級読解クラスレベル、IF＝項目容易度、ID＝項目弁別力、Diff＝項目困難度、*SE*＝標準誤差、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$p \cdot X\ i^o$デザインによるG研究の結果を示す。受験者の分散成分％はテスト得点の分散成分を100％とした場合、文法、語彙、読解、それぞれ11％、12％、11％であった。つまり、受験者の能力には10％程度のばらつきがあり、受験者の能力はテスト得点の分散の10％を占めることがわかった。受験者の文法と語彙、文法と読解、語彙と読解の共分散はそれぞれ.02、.02、.02で相関係数（disattenuated correlation）は.83、.79、.79となった。つまり、三つのセクションの得点は高い相関関係にある。項目の分散成分％は文法、語彙、読解、それぞれ13％、13％、7％で、項目の難易度には10％程度のばらつきがあったことになる。受験者X項目交互作用の分散成分％は76％、75％、82％で、受験者と項目以外にも様々な要因が得点のばらつきを生じさせていることを示す。

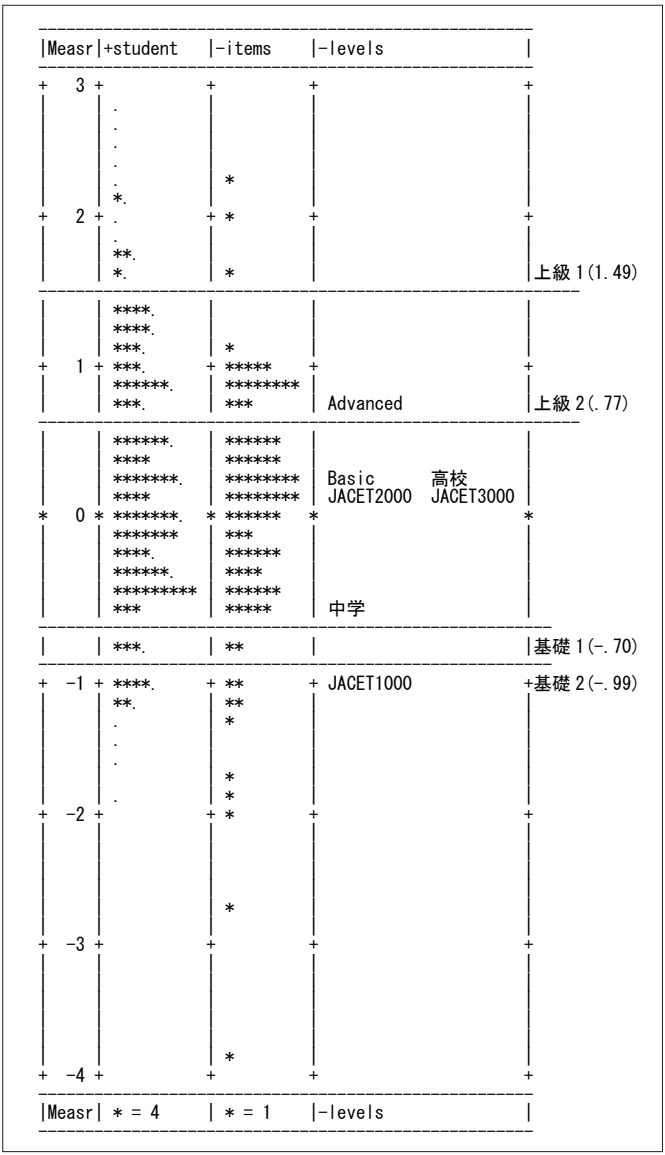

図2. グラフ化したFACETS分析の結果. 中学＝中学文法レベル、高校＝高校文法レベル、JACET1000＝JACET1000語彙レベル、JACET2000＝JACET2000語彙レベル、Basic＝基礎読解クラスレベル、Advanced＝上級読解クラスレベル.

表3. p • X i°デザインによるG研究の結果について（*N*＝428）

	文法		語彙		読解	
相	分散成分	分散成分％	分散成分	分散成分％	分散成分	分散成分％
受験者（*p*）	**.02709**	11%	.83205		.78810	
	.02288		**.02791**	12%	.79338	
	.02144		.02191		**.02732**	11%
項目（*i*）	**.03281**	13%				
			.03030	13%		
					.01781	7%
受験者X項目（*p X i*）	**.18603**	76%				
			.17753	75%		
					.19925	82%
分散成分合計	.24593		.23574		.24438	

注. 太字＝分散成分、対角線上＝相関係数、対角線下＝共分散.

表4は*p* • X *I*°デザインによるD研究の結果を表している。文法、語彙、読解の項目数がそれぞれ40、40、10の場合、母得点（universe score [$\sigma^2(\tau)$]）は受験者分散成分推定値と同じで.02709、.02791、.02732となり、これは古典的テスト理論では真値（true score）に対応する。プレイスメントテストは集団基準準拠テストの一種であるので、目標規準準拠テストに用いる信頼度指数ではなく、一般化可能性係数を提示すべきである。そのためにはまず相対誤差（relative error[$\sigma^2(\delta)$]）を求める必要があり、これは文法の場合、受験者X項目分散成分推定値である.18603を項目数である40で割ることで求めることができ、.00465（.18603/40）となる。語彙と読解の相対誤差はそれぞれ.00444、.01992である。そして、母得点を相対誤差に母得点を足したもので割ると一般化可能性係数を求めることができ、文法、語彙、読解それぞれの係数は.85（.02709/[.00465 + .02709]）、.86、.58となる（表4中にあるD研究7参照）。読解のみ係数が十分ではなく、内部一貫性が欠如していることになる。合成一般化可能性係数（composite generalizability coefficient [$p^{\wedge 2}$]）を求めるにはまず合成母得点（composite universe score [$\sigma^2(\tau)$]）と合成相対誤差（composite relative error [$\sigma^2(\delta)$]）を求める。その値はすべてのセクションの母得点の値と重み付け（weight）および相対誤差の値と重み付けで求められる。ここでは特定の重み付けを指定し算出する名義重み付け（nominal weight）ではなく統計的に求めた効果的重み付け（effective weight）を用いた（Brennan, 2001a）。項目数がセクションごと40、40、10の場合、効果的重み付けは.44、.44、.11であった。合成母得点は.02452（[(.44)(.44)(.0271)] + [(.44)(.44)(.0280)] + [(.44)(.11)(.0273)] + [2(.44)(.44)(.0229)] +[2(.44)(.44)(.0219)] +[2(.44)(.11)(.0214)]）で、合成相対誤差は.00204であった。そして、合成母得点を合成母得点と合成相対誤差を足した値で割ると合成一般化可能性係数が求められ、.92（.02452/

[.02452+.00204]）であった。合成一般化可能性係数は75項目の場合は.91（表4中にあるD研究3参照）で120項目の場合でも.94で（表4中にあるD研究19参照）さほど変化はなく、プレイスメントテストには十分高い値である。各セクションの一貫性を保つためには一般化可能性係数で.80は必要である。各セクション25項目だと.80には達せず、30項目であれば達することがわかった。やはり、各セクションの一般化可能性係数が十分に高く、そのうえ合成一般化可能性係数も高いほうが望ましい。

表4. p • X I°デザインによるD研究の結果について（*N*＝428）

D研究	文法		語彙		読解		合計	
	k	ρ^2	*k*	ρ^2	*k*	ρ^2	*k*	$\hat{\rho}^2$
1	35	.84	30	.83	10	.58	75	.91
2	30	.81	35	.85	10	.58	75	.91
3	25	.78	25	.80	25	.77	75	.91
4	35	.84	35	.85	10	.58	80	.91
5	40	.85	35	.85	10	.58	85	.92
6	35	.84	40	.86	10	.58	85	.92
7	40	.85	40	.86	10	.58	90	.92
8	30	.81	30	.83	30	.80	90	.92
9	45	.87	40	.86	10	.58	95	.93
10	40	.85	45	.88	10	.58	95	.93
11	40	.85	40	.86	15	.67	95	.93
12	45	.84	45	.83	10	.58	100	.91
13	40	.85	40	.86	20	.73	100	.93
14	40	.85	40	.86	25	.77	105	.93
15	50	.88	50	.89	10	.58	110	.94
16	40	.85	40	.86	30	.80	110	.93
17	40	.84	35	.85	35	.83	110	.93
18	40	.85	40	.86	35	.83	115	.94
19	40	.85	40	.86	40	.85	120	.94

注. *k*＝項目数、ρ^2＝一般化可能性係数、$\hat{\rho}^2$＝合成一般化可能性係数.

表5はFACETS分析による学習レベル困難度推定値の結果についてである。図2でも視覚化された結果がある。学習レベル困難度の分離指数は14.37でラッシュ信頼性係数は1.00なので推定値にはばらつきがある。受験者にとってもっとも容易だったのはJACET1000語彙レベルであり、順に中学校文法レベル、JACET3000語彙レベ

ル、JACET2000語彙レベル、高校文法レベル、基礎読解レベル、上級読解レベルであった。本文を読んで設問に解答するには、文法や語彙などより多くの知識を必要とするため読解項目の難易度が高かったと思われる。標準誤差の値は低く、推定値の誤差は少なかった。インフィット平均二乗は基準の範囲以内であるが、アウトフィット平均二乗は上級レベル読解の値が基準範囲以外となり不適合となった。これは、アウトフィット平均二乗は外れ値に敏感であるためだと思われる。

表5. FACETS分析による学習レベル困難度推定値の結果について（*N*＝428）

学習レベル	Diff	*SE*	Infit MS	Outfit MS
中学文法レベル	-0.65	0.03	1.00	1.00
高校文法レベル	0.29	0.02	1.00	1.00
JACET1000語彙レベル	-0.94	0.03	1.00	0.90
JACET2000語彙レベル	0.15	0.03	1.00	1.00
JACET3000語彙レベル	0.12	0.05	1.00	1.00
基礎読解レベル	0.30	0.05	1.00	1.00
上級読解レベル	0.73	0.05	1.10	1.30

注. Diff＝学習レベル困難度、*SE*＝標準誤差、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.

図2にあるグラフ化されたFACETS分析の結果をみると学習レベルと級ごとの習熟度合いが把握できる。上級1の分割点は受験者能力推定値の1.49であった。上級1の履修者には高校までの学習レベルは容易に正解できることがわかる。例えば、もっとも推定値が高かった上級読解レベル（.73）でも50％以上正解できる確率があり、また基礎読解レベルには75％以上正解できる確率がある。その逆に基礎2の履修者にとってJACET1000語彙レベルは50％程度の確率で正解を導きだすことはできるが、それ以外のレベルは正解できる確率が低く、上級読解レベルに正解する確率は0％に等しい。この結果から、基礎クラス履修者は中学文法学習レベルにも到達していないことがわかる。

表6はクラス別のプレイスメントテストの得点結果についてである。全体的に上級クラスと判定された受講者は平均値が高く、文法であれば中等教育の学習内容、語彙であればJACET3000語彙レベル、読解であれば基礎レベルの本文はほぼ理解できることがわかる。他方、基礎レベルと判定された受講者は、平均値が低く習熟していないということがわかる。この結果から、上級クラス履修者が基礎レベルの授業内容を学習することはすでに学習済みの内容なので不利益になるといえ、基礎クラス履修者には上級クラスの授業内容は難しすぎるので不利益になるといえる。また、上級1と上級2の平均値差は顕著であるが、基礎1と基礎2の平均値差はさほどない。今後、

この基礎二クラスをより分別する項目を入れたほうがいいかもしれない。しかし、級ごとに得点順になっており、また得点差も顕著であることから習熟度別に編成されているといえる。

表6. クラス別プレイスメントテスト得点結果について

		文法		語彙		読解		合計	
クラス	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
上級1	36	33.53	3.38	35.94	2.44	15.22	2.72	84.69	4.17
上級2	33	30.33	2.39	34.09	1.84	12.42	2.59	76.85	1.37
中級	310	22.48	5.41	24.13	5.58	7.72	3.55	54.31	10.99
基礎1	25	13.04	2.68	16.76	3.35	4.32	2.55	34.12	0.76
基礎2	24	11.17	2.91	14.42	2.89	3.33	2.55	28.92	3.46
全体	428	22.82	7.13	24.91	7.19	8.27	4.35	56.00	16.25

表7は中級クラスの二クラスを中級群とし、学習支援およびリメディアル教育を受けた基礎クラスの二クラスを基礎群とした場合の文法事前テストと文法事後テストの結果についてである。まず、事前テストの結果をみると中級群と基礎群の平均値差は十点ほどある。事後テストの結果は中級群の場合、事前テストの平均値と比較すると、十点ほど値が下がっている。これはテスト結果が成績に考慮されないのでテスト受験に集中しなかったということが主な原因として挙げられる。標準偏差の値が大きく、真剣に受験した履修生とそうでない者がいたため得点のばらつきが生じた可能性がある。基礎群の事後テスト平均点は、事前テスト平均点よりも六点ほど上昇している。いずれにしろ、中級群と基礎群の得点低下または得点上昇の理由にはさまざまな要因があるため学習効果のためと断定することはできない。最終授業内で六件法の授業評価アンケートを実施した結果、中級群と基礎群の平均値が1.00ほど差があった項目は学習支援についてのもので、基礎クラス履修者にとって学習支援を受けることは重要であると「まあそう思った」（*M* [基礎群]＝4.11; *M* [中級群]＝2.16）、役に立ったと「まあそう思った」（*M* [基礎群]＝4.32; *M* [中級群]＝2.63）という結果になった。しかし、中級群と基礎群の授業満足度を調査する「全体的にこの授業は満足できた」という項目における平均値差（*M* [基礎群]＝4.63; *M* [中級群]＝4.51）はさほど顕著ではなく、全体的にやや満足したようである。

表7. クラス別文法到達度テスト得点結果について

	文法事前テスト（α=.85）			文法事後テスト（α=.92）		
クラス	*n*	*M*	*SD*	*n*	*M*	*SD*
中級1	26	30.38	6.34	21	12.14	2.50
中級2	25	32.36	8.47	24	28.63	7.93
中級群	51	31.35	7.45	45	20.93	10.24

	文法事前テスト（α=.85）			文法事後テスト（α=.92）		
クラス	*n*	*M*	*SD*	*n*	*M*	*SD*
基礎1	25	20.80	5.09	22	26.82	5.21
基礎2	24	19.88	4.29	23	26.78	5.95
基礎群	49	20.35	4.69	45	26.80	5.53

考察

考察では解釈的論証に対しての妥当性論証の論拠を挙げることを主な目的とする。まず、得点化から決定までの推論に対する解釈的論証の前提条件と妥当性論証の論拠を表8に示す。

表8. プレイスメントテストの解釈的論証の前提条件と妥当性論証の論拠について

推論	解釈的論証（前提条件）	妥当性論証（論拠）
得点化	• プレイスメント項目は意図した正解を受験者が解答でき、高得点者がより正解できる	• ほぼすべての項目が機能していたことから受験者の解答は適切に観測得点へ得点化された
一般化	• プレイスメント項目は一貫して測定領域から抽出されていて、また項目数は誤差が最小になる程度ある	• 全項目は測定領域から抽出されており、合成一般化可能性係数が高かったことから誤差が少ないため観測得点は一貫性をもって測定領域得点へ一般化された
外挿	• プレイスメントテストレベルはリーディング科目の授業目標と難易度が一致している	• 学習レベルの難易度は予想通りであったため測定領域得点から適切に目標得点へ外挿された
決定	• プレイスメントテスト得点をクラス分けの判断材料として使用し、習熟度別クラス編成したため、受講者のレベルにより適した授業内容を提供できることにより、正の波及効果が生じる	• 得点順に習熟度別にクラスが編成されており、基礎群の授業内容に対する評価が高く、到達度テストで得点上昇が生じたことから目標得点から用途へ決定が下された

研究課題1はプレイスメントテスト項目はどの程度意図した正解を受験者が解答でき、高得点者がより正解できるかであった。古典的テスト理論の観点からみると、受験者にとって項目容易度の値が高いまたは低いということで、易しすぎるまたは難しすぎるという項目が90項目中28項目あった。プレイスメントテストは得点のばらつきを

生じさせることでより的確なレベル判別ができるので、受験者の能力をいかに弁別しているかの項目弁別力が重要であり、その値が低かったのはわずか四項目のみであった。よって、項目容易度の値が不適切な項目があったが、集団基準準拠テストの一種であるプレイスメントテストの主たる目的である得点のばらつきを生じさせるという点では項目弁別力が高い項目が大半であったことから、目的は達成できたといえる。また各項目には正解が明確にあり、高得点者がより確実に正解できるものであったといえる。

項目応答理論の観点からみると、本テストは項目分離指数が高く項目困難度推定値にはばらつきがあり広範囲の能力推定値を測定できるので、適切である。しかし、項目困難度推定値が低く受験者の能力値を測るにはさほど必要ない項目が二項目あった。また、現時点では合計得点が高いまたは低い順に上級、または基礎クラスに振り分けているが、今後項目応答理論を用い分割点（cut-off point）を設けることで振り分ける場合、分割点の推定値あたりに項目が多くあるほうがより適切に受験者の能力を判別できるのでよいかもしれない（Hudson, 1991）。例えば、－.50から－1.00くらいの推定値に収まる項目がより多くあれば、基礎クラス1か基礎クラス2を履修したほうがいいのかがより正確に判断できる。インフィット平均二乗は項目67以外すべての項目は基準範囲以内であり、これはほぼすべての項目は作成時に意図していた正解に受験者が解答でき、高得点者がより正解できたということがいえる。したがって、研究課題1は達成でき、受験者の解答は適切に観測得点へ得点化された。

研究課題2はプレイスメントテスト項目はどの程度一貫して測定領域から抽出されていて、また項目数は誤差が最小になる程度あるかであった。まず、このプレイスメントテスト開発手順に記述したが、測定領域である文法であれば中学・高校の教科書にある文法事項、語彙であればJACET1000~3000レベル、読解であれば基礎クラスと上級クラスで実際に用いられた本文より項目を抽出しているので代表性は保たれていると判断する。次に、多変量一般化可能性理論を用い、どの程度の項目数があれば誤差が最小になるかを検証した。合成一般化可能性理論は文法、語彙、読解の項目数が本テストと同じでそれぞれ40、40、10項目の場合、.92になり十分な値ではあるが、読解の項目数が十項目と少なく、一般化可能性係数が.58と低い結果となった。よって、全体的には項目は一貫して測定領域から抽出されており、誤差が少ないが、読解の項目数のみ少なく誤差がより大きい。誤差を最小にし、全セクションともに一般化可能性係数を確実に.80以上にするにはテストを改訂し、各セクションの項目数をそれぞれ30項目にする必要がある。しかしながら、研究課題2は十分に達成でき、観測得点は一貫性をもって測定領域得点へ一般化された。

プレイスメントテストの項目分析を行い信頼性を検証した既存研究と本研究結果を比較する。Culligan and Gorsuch（1999）では読解と聴解150項目を含むSLEPテスト（Secondary Level English Proficiency Test）を日本人大学生に実施したところ、項目弁別力が.20以上であった項目は66項目のみであった。また、信頼性係数（KR-20）は.81であった。Westrick（2005）は日本人大学生161名に読解、文法知識、語彙知識を測る目的に開発された60項目から成るQPT-PPT（the Quick Placement Test－Pen and Paper Test）を実施したところ、項目弁別力が.20以上であった項目はわずか23項目で、信頼性係数（KR-20）は.55であった。既存研究結果と比較しても不良項目がほとんどなく、一般化可能性係数が高い結果を示す本研究結果のほうが数段とよい。また、Beglar（2010）でのラッシュ項目信頼性（.95以上）やKoizumi et al.（2011）で提示さ

れた項目弁別力平均値（M = .45）と合成一般化可能性係数（ρ^2 = .93）と比べ、本研究結果は項目弁別力平均値が.40で合成一般化可能性係数が.92であることからほぼ同等だといえる。Chapelle, Chung et al.（2010）で17項目からなる発信型文法テストをもとに算出されたクロンバック係数（α = .82）より本研究の合成一般化可能性係数のほうが高い。これはやはり市販のテストをそのまま使用するのではなく、カリキュラムを考慮し本学部独自のテストを作成し、試行テストの結果をもとに改訂したためであろう。

研究課題3はプレイスメントテストレベルはどの程度リーディング科目の授業目標と難易度が一致しているであった。文法学習レベルは中学校より高等学校、語彙学習レベルはJACET1000レベルよりJACET3000レベル、読解学習レベルは基礎レベルより上級レベルのほうが高いことが予想された。FACETS分析により学習レベル困難度推定値を算出した結果、JACET2000レベルとJACET3000レベルにおいて学習レベル困難度推定値がほぼ同等になった以外は、他の推定値は予想通りであった。基礎クラスではbe動詞から分詞構文までを学習することになっていた。図2をみると、基礎クラスの履修生はJACET1000の語彙レベルにはほぼ到達しているがそれ以外の学習レベルにはさほど到達していないことから、適切な難易度の分割点で、基礎クラスのクラス分けがなされていたと考えられる。また、上級クラスの履修生は中等教育での学習事項は習熟していることを前提とし、それにもとづき読解指導が行われる。このテスト結果から上級クラス履修生は上級レベルの読解設問以外は到達していることがわかった。よって、適切な難易度の分割点で、上級クラスのクラス分けがなされていたと考える。したがって、研究課題3は達成でき、測定領域得点から適切に目標得点へ外挿された。

Beglar（2010）は語彙出現頻度順に項目困難度推定値の平均値を算出した。高頻度の語彙レベルの値は低く、順に高くなっており、Kane（2006）の枠組みでは外挿にあたる本質的側面を評価している。本研究でもJACET1000レベルよりJACET2000レベルとJACET3000レベルのほうが学習レベル困難度推定値は高く、Beglarと同様の結果が得られたため、高頻度の語彙ほど習熟しやすいということがいえる。Chapelle, Chung et al.（2010）では第二言語習得理論の分野で発表された文法習得順序をもとに、文法項目を初級レベル、中級レベル、上級レベルと三段階に区分けをした。項目容易度平均値は予想通り初級レベル、中級レベル、上級レベルの順に高くなり、初級レベルの文法項目がもっとも習得しやすいという結果になり、Chapelle, Chung et al.はKaneの枠組みでは外挿にあたる説明の妥当性論証の論拠として用いている。本研究でも文法学習レベルを区分けした結果、中学校学習レベルのほうが推定値は低くなったことから、中学校学習レベルのほうが習熟しやすいということがいえる。

研究課題4はプレイスメントテスト得点をクラス分けの判断材料として使用し、習熟度別にクラス編成したため、受講者のレベルにより適した授業内容を提供できることにより、どの程度正の波及効果が生じるかであった。基礎クラスの履修者は学習支援などを含むリメディアル教育を受け、また彼らの習熟度に合った学習目標が設定されていたので、プレイスメントテストで基礎クラスに振り分けられ、そのクラスを受講したことにより、基礎クラスの履修生にはより高い波及効果があったといえるだろう。その論拠として30回の授業を受けた結果、事前テストと比べ事後テストの得点平均値は上昇していた。また、アンケート結果でも学習支援は役に立ったとの意見があり、波及効果の論拠として挙げられるであろう。しかし、基礎群の得点上昇は顕著ではな

く、また教育効果以外にも多様な要因があると思われるので決定の論拠としては若干弱いと考える。また、到達度テスト得点とアンケート結果では直接プレイスメントテストの波及効果を検証しているわけではなく、間接的な影響なので若干弱い論拠ではある。したがって、研究課題4は論拠としては若干弱いが達成されたとみなし、目標得点から用途へ決定が下された。

結論

教育的示唆としては第一にKane（2006）の論証型妥当性枠組みは明確であり簡潔であるということが挙げられる。つまり、本研究のように教員が項目分析、信頼性の算出などを行って期末テストなどを含むテスト得点の解釈と使用の妥当性を評価することが可能である。また、妥当性を評価することで問題点なども明確になることから、Kane（2006）の枠組みを用いて妥当性を評価することは重要である。本研究では決定の推論に対する論拠が弱かったため、習熟度別に編成することは各レベルを履修している履修者に利益または不利益があるか、また正または負の波及効果があるかを今後さらに直接的に検証する必要がある。第二に多変量一般化可能性理論の有用性である。一般化可能性理論ではセクションごとにG研究を行わないといけないことに対し、多変量一般化可能性理論を用いることで複数のセクションから成るテストでも一回のG研究だけで済み、またセクション間の相関関係や合成一般化可能性係数も求めることができるため、非常に応用の幅が大きい理論であるといえる。第三は多相ラッシュモデルの有用性である。このモデルを用いることでいくつかの相の推定値を同一上の間隔尺度に変換でき、また、ある受験者がある相に属す項目に何パーセントの確率で正解できるかを推定することができる。これは目標設定が受験者に一致しているかなどを判断する上では最適な判断材料になるのではないか。

本研究ではKane（2006）が提唱する論証型妥当性枠組みを用い大学一年生対象必修リーディング科目のクラス分け目的で開発したプレイスメントテストの解釈的論証を挙げ、それに対する妥当性論証を挙げ、テスト得点の解釈と決定についての妥当性を評価した。ほぼすべての項目に対し受験者は意図した正解に解答し、高得点者が正解できる項目がほとんどであったことが得点の妥当性論証の論拠である。また、合成一般化可能性係数が.90以上になったことから、項目は測定領域から一貫して抽出されており、項目数は誤差を最小にするために十分であったことが一般化の妥当性検証の論拠である。FACETS分析を用いて学習レベル困難度推定値を求めた結果、推定値は予想通りの難易度を示し、また授業目標は履修生に適切に設定されているということを外挿の論拠とした。到達度テストを基礎クラスと中級クラス履修者を対象に事前・事後テストとして実施し平均値差を検証した結果、基礎クラス履修者は得点を上昇させることができた。また、授業評価アンケートにより学習支援など指導が効果的だったという意見が示された。したがって、プレイスメントテストの波及効果があったことが示され、それを決定の妥当性論証の論拠とした。このようにプレイスメントテスト得点の解釈と決定についての妥当性を評価することができた。今後、D研究の結果にある通り、各セクションの項目数を30項目になるようテスト改良をすることで各セクションの誤差を最小にする、実証型のみで外挿の妥当性論証論拠を提示したので分析型で求めた論拠も提示する、および直接授業を観察するなどして波及効果を検証することでよりテスト得点の解釈と使用の妥当性を向上させることができるであろう。

謝辞

この研究ができたのも多くの教員と事務員のおかげです。ここに感謝を述べたいと思います。編集長と査読者にこの研究について貴重なご指摘をいただけたことに感謝をいたします。また、佐藤敬典氏（メルボルン大学）にも丁寧にみていただきましたことに感謝を述べます。本研究に残る誤りはもちろんわたくしの責任です。

熊澤孝昭（*Takaaki Kumazawa*）は現在関東学院大学法学部所属准教授である。Temple Universityより教育学（博士）を取得した。専門分野は英語教育学で、特にテスト理論、第二言語習得理論、カリキュラム開発に興味がある。

引用文献

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34. doi:10.1207/s15434311laq0201_1

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*, 101-118. doi:10.1177/0265532209340194

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

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.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*, 65-83. doi:10.2307/3587508

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105. doi:10.1037/h0046016

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-352). London: Routledge.

Chapelle, C. A., Chung, Y., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*, 443-469. doi: 10.1177/0265532210367633

Chapelle, C. A., Enright, M., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*, 3-13. doi: 10.1111/j.1745-3992.2009.00165.x

Cheng, L., & Watanabe, Y. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

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.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Selection theory for a political world. *Public Personnel Management, 9*, 37-50.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. doi:10.1037/h0040957

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.

Cumming, A., & Berwick, R. (Eds.). (1996). *Validation in language testing*. Clevedon, UK: Multilingual Matters.

大学英語教育学会基本語改訂委員会. (2003). *JACET list of 8000 basic words.* Tokyo: JACET.

Hudson, T. (1991). Relationships among IRT item discrimination and item fit indices in criterion-referenced language testing. *Language Testing, 8*, 160-181. doi:10.1177/026553229100800205

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535. doi:10.1037/0033-2909.112.3.527

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*, 135-170. doi:10.1207/s15366359mea0203_1

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: Greenwood Publishing.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*, 3-17. doi:10.1177/0265532211417210

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17. doi:10.1111/j.1745-3992.1999.tb00010.x

Koizumi, R., Sakai, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Development and validation of a diagnostic grammar test for Japanese learners of English. *Language Assessment Quarterly, 8*, 53-72. doi:10.1080/15434303.2010.536868

熊澤孝昭. (2010). 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. *JALT Journal, 32*, 169-188.

Kunnan, A. J. (Ed.). (1998). *Validation in language assessment*. Mahwah, NJ: Erlbaum.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131-166. doi:10.1191/0265532206lt325oa

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

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

Lissitz, R. W. (Ed.). (2009). *The concept of validity*. Charlotte, NC: Information Age.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*, 9-20. doi:10.3102/0013189X010009009

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*, 137-159. doi:10.1080/15434301003664188

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*, 355-390. doi:10.1177/0265532207077205

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493. doi:10.1177/0265532208094273

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

霜崎實他. (2006). *New Crown English Series I & II*. 東京:三省堂出版.

高橋貞雄他 (2005). *New Crown English Series 1, 2, & 3*. 東京:三省堂出版.

Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

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. doi:10.1177/0265532207076365