指標の妥当性をどう捉え，妥当性検証をどのように行うべきか

語彙の豊かさ指標「S」の解釈的論証

Towards exploring a suitable way to validate measurements in second language acquisition studies
Interpretive arguments for validation of the lexical richness measure S

小島ますみ
Masumi　KOJIMA

Abstract

The purpose of this paper is to examine the concept of validity in tests and measurements and to investigate several approaches to validation through a survey of previous studies and discussion.　Next, it explores a suitable approach to validate the lexical richness measure S, which is proposed by Kojima (2010).　S is developed to assess the productive vocabulary of learners in their written or spoken texts.　Validity is considered as an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989).　The argument-based approach of Kane (2006) is employed, and interpretive arguments for the validation of S are discussed.

Keywords : 妥当性，妥当化，論証に基づくアプローチ，語彙の豊かさ，S

1. はじめに

テストや指標の妥当性 (validity) とは，テストや指標が測定しようとしていると主張し，「意図しているもの」を測定している度合いと考えられている (Brown, 1996)。Messick (1989 池田央訳 1992, p. 21) はテストの妥当性について，「テスト得点またはそれに類する他の評価法をもとにして行う推論 (inferences) と行為 (actions) の相応性 (adequacy) ならびに適切性 (appropriateness) について，それを支持する経験的証拠と理論的理由づけの度合いを示す総合的な評価判断」と定義している。Kane (2006) は，提案するテストの解釈と使用のもっともらしさを評価する過程を妥当化 (validation) と呼び，提案するテストの解釈と使用を支持または反駁する証拠の度合いを示す妥当性 (validity) と区別した。テストの妥当性について，これまでさまざまな議論がなされ，妥当性のとらえ方が変化してきた。本研究ではこれまでの妥当性研究について概観するとともに，言語発達指標の妥当性をどう捉え，妥当性検証をどのように行うべきかについて考察し，小島 (2010) で提案した語彙の豊かさ指標「S」の妥当性検証アプローチを提案する。

2. 学習者のテクストにおける語彙の豊かさ

語彙の豊かさ (lexical richness) 指標とは，学習者の産出する語彙が，どの程度多様で広範なものかを量的に示す指標である (Laufer & Nation, 1995)。前提として，学習者の持つ語彙の広さが，産出される語彙に反映されると考えられている (Laufer & Nation, 1995; Meara & Bell, 2001)。語彙の豊かさを量的に表す指標は，母語習得，第二言語習得，医学臨床等の分野で，言語の発達指標として用いられてきた (Malvern & Richards, 1997)。

語彙の豊かさ指標は，ライティングやスピーキングの評価に影響を与える要因の研究や，語彙知識と語彙使用の関係についての研究など，幅広い研究に応用されてきた (e.g. Arnaud,1984; Daller & Phelan, 2007; Engber, 1993; 石川, 2008; Koizumi, 2005b; Linnarud, 1986; Malvern & Richards, 2002; 水本, 2008; 杉浦, 2008; Richards & Malvern, 2000; Vermeer, 2000)。このような研究において，語彙の豊かさをどのように測定するかにより，結果は異なるものとなる。妥当性の低い指標を使用すれば，そこから得られる研究結果も妥当性の低いものとなり，研究を間違った方向へ導きかねない。これまで学習者の言語産出における語彙の豊かさを測定するさまざまな指標が提案されているが，スコアがテクストの長さの影響を受ける，明らかに熟達度の異なる学習者群を区別しないなど，問題点も多い。また指標の妥当性が十分に検討されないまま

に使用されているという実態もある。

小島 (2010) は，学習者の語彙の豊かさをより適切に評価することを目指し，語彙の豊かさ指標 S を開発した。それまで提案されていた語彙の豊かさ指標の多くは語彙を基本語と低頻度語に 2 分するのみであったのに対し，語彙の豊かさ指標 S は語彙の頻度順位を連続データとして扱い，学習者の発表語彙の頻度分布全体をとらえる指標である。具体的には，テクストから 50 語ずつ取った各サンプルにおける高頻度語の累積カバー率をデータとし，データに最も近似するモデルを求め，累積カバー率が 100%に達する単語の頻度順位を推定する。例えば S が 2,016 の場合，そのテクストにおける学習者の発表語彙レベルは 2,016 語と考えられるため，スコアの解釈が容易であるという利点を持つ。小島 (2010) の結果より，学習者のエッセイから得られた累積頻度のデータは S のモデルに適合しており，熟達度の異なる 2 つの学習者群と母語話者群で S のスコアに有意差があった。また，総語数 200-600 語程度のテクストであれば，S はテクストの長さの影響を受けず，安定した結果を返すことが示された。しかし，小島 (2010) は既存の学習者コーパスを使用したため，学習者の語彙サイズと S の関係などについて調査を行なうことができなかった。より体系的に S の妥当性検証を行う必要があるが，妥当性とはそもそもどのようなものなのだろうか。次節では，これまでの妥当性研究を概観し，妥当性概念について検討する。

3. 妥当性のとらえ方の変遷

3.1 基準関連モデル (criterion model)

Kane (2006) によると，1920～1950 年代は基準関連妥当性 (criterion validity) が妥当性の黄金基準 (gold standard) であった。妥当なテストとは，外部基準との相関が高いテストであると考えられていた。基準関連妥当性は併存的妥当性 (concurrent validity) と予測的妥当性 (predictive validity) の 2 種類に分けられた。併存的妥当性の検証は，基準となるテストのスコアと問題となるテストのスコアを調べ，2 つのテストの相関が高ければ，問題のテストは妥当性があると考えられた。予測的妥当性の検証は，問題となるテストと将来のパフォーマンスの相関を調べ，相関が高ければ問題のテストは妥当性があると考えられた。例えば，入社試験で使用する適性検査とその後の会社での実績を調べ，2 者の相関関係を分析することで，適性検査の妥当性を検証するなどである。

基準関連モデルは，優れた外部基準が存在すれば，シンプルで有効な妥当化のアプローチとなる (Kane, 2006)。Kane (2006) は，基準関連モデルには 2 つの主な利点があると述べている。1 つ目は，多くの場面で，問題のテストが外部基準に関連するという証拠は，そのテストスコアの解釈や利用法のもっともらしさを支える証拠となる。例えば，ある適性試験のスコアと入社後の実績の相関が高ければ，その適性試験が妥当だという主張は説得力のあるものとなる。2 つ目は，外部基準に基づいて問題のテストの妥当化を行うことは，客観的な手続きにより行うことができるという点である。

しかし，基準関連モデルには問題点もある。Kane (2006) は，適切な外部基準を見つけることは，多くの場合難しいと述べている。問題となるテストよりも明らかによいテストは存在しないかもしれない。例えば入社後の実績など，外部基準を概念化すること自体が難しい可能性がある。外部基準の妥当性を検討するために，他の適切な外部基準が必要となるなど，議論が循環する。Kane (2006) はこのような問題について，基準関連モデルの根本的な問題 (fundamental problem) であると述べている。

3.2 内容的モデル (content model)

妥当性の内容的モデルは，外部基準に寄らない妥当化を目指し，1940～1960 年代に強調された。妥当なテストとは，問題や質問の内容が測定したい領域を適切に反映し，代表性のあるテストであると考えられた (Kane, 2006)。したがって妥当化の観点は，領域適切性や内容代表性の吟味となり，専門家の判断に基づき検討された。もしテスト項目が，測定したい領域の範囲内から偏りなく選ばれており，項目数が十分大きければ，テストで観察されるパフォーマンスは，そのテスト領域の適切性や内容代表性を持つと考えられる (Kane, 2006; 村山, 2006)。テスト領域全体をカバーする項目群は，項目ユニバースと呼ばれる (図 4 参照)。

このような妥当性の内容的モデルを語彙習得研究に当てはめて考えてみよう。例えば，生徒が高頻度 1,000 語のうち何語知っているか調べたいとする。しかし，1,000 語すべてをテストする時間はないし，生徒の集中力もそう長くは続かないだろう。そこでテスト項目を高頻度 1,000 語から選ぶことになる。テスト項目は高頻度 1,000 語からランダムに偏りなく選ぶ必要がある。偏りがあっては，高頻度 1,000 語を代表するテストとはならない。また，テストの項目数も 1 問や 2

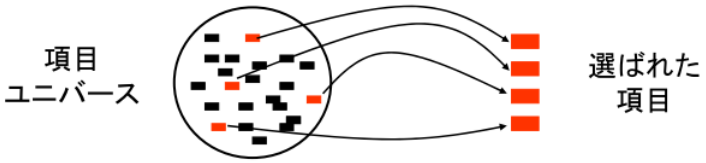

図 1　領域適切性と内容代表性の考え方 (村山, 2006, p. 8)

問では，高頻度 1,000 語のうち何割の単語を知っているかを適切に推測することができない。テスト項目が高頻度 1,000 語から偏りなく選ばれ，項目数も十分に大きければ，例えば生徒がテスト項目の約 8 割正解した場合，教師は生徒が高頻度 1,000 語のうち約 8 割の 800 語を知っていると推測できるであろう。

内容的モデルの長所としては，外部基準に寄らずに妥当化を行うことができる点である。Kane (2006) は，特定のスキルを測定するテストでは，内容的モデルのアプローチはうまく機能すると述べている。内容的モデルの欠点としては，領域適切性や内容代表性の検討は専門家の判断に基づいて行われるため，主観的になりやすいという点である (Kane, 2006; Messick, 1989)。Kane (2006) は，テストの妥当化を行うのは，テストの開発者自身であることが多いため，テストの妥当性を認める方向へのバイアスがかかる傾向があると指摘している。

3.3 構成概念モデル (construct model)

Cronbach & Meehl (1955) は，科学理論の仮説演繹モデル (hypothetico-deductive model) の考え方に基づき，構成概念妥当性の枠組みを提案した。仮説演繹モデルでは，ある構成概念は，他の構成概念や観察可能なパフォーマンスのネットワークにより定義される。このネットワークは法則定立ネットワーク (nomological network) と呼ばれる (図 2 参照)。

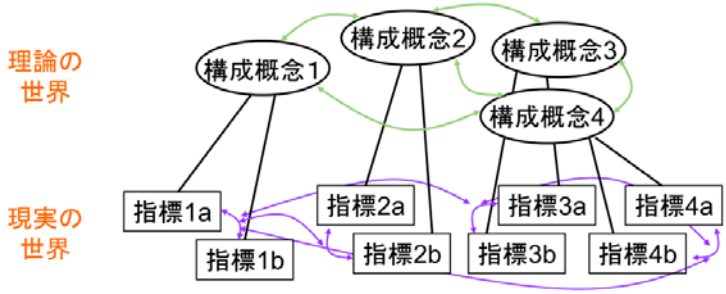

図 2　法則定立ネットワーク (村山, 2006, p. 10)

Cronbach & Meehl (1955) は，テストで測定するものは，理論的・仮説的な構成概念であると述べている。彼らのアプローチでは，複数の観察から想定される法則定立ネットワークを検証することで，問題となるテストの妥当化を行う。例えば，問題となるテストは，理論的に関連の強い構成概念を測定するテストとは相関が高いはずであり，逆に理論的に関連の弱い構成概念を測定するテストとの相関は低いはずである。したがって妥当化は，仮説検証の繰り返しのプロセスとなり，研究者は理論から実証可能な予測を複数立て，それを検証していく (図 3 参照)。

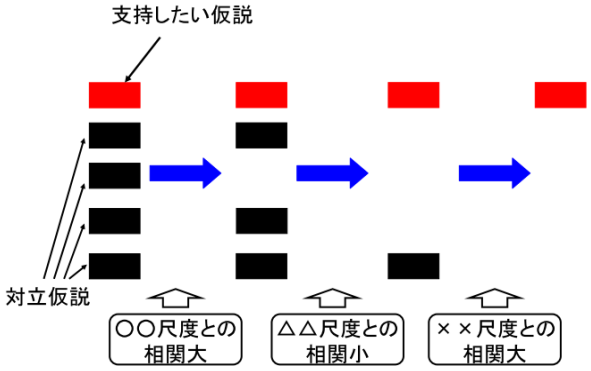

図 3　仮説の実証例 (村山,2006, p. 13)

3.4 構成概念モデルの発達

Cronbach & Meehl (1955) は構成概念妥当性を提案したが，妥当性の包括的な枠組みとしては提示しなかった。Kane (2006）によると，1960 年代に発行された APA (American Psychological Association), AERA (American Educational Research Association), NCME (National Council on Measurement in Education) のテストスタンダードでは, 構成概念妥当性は基準関連妥当性や内容的妥当性と区別され，特に適切な外部基準が存在しないときに構成概念妥当性の検討が適切とされた。1970 年代に入っても，これらのテストスタンダードは，基準関連妥当性・内容的妥当性・構成概念妥当性を妥当性の 3 つの種類として記述し，3 つの妥当性の関係についての詳細な考察は行われなかった (Kane, 2006)。

これに対し，1970 年代後半から，より多くの研究者が妥当性を単一の概念としてとらえる必要があると考えるようになり, 1980 年代初めには, 構成概念妥当性は妥当性の一般的なアプローチとして広く受け入れられるようになった (Kane, 2006)。近年では，構成概念妥当性は妥当性の下位概念でなく，妥当性そのものと捉えられている (Messick, 1996)。

Messick (1989) は，構成概念妥当性とは，テスト得点に基づき構成概念に対する推論・解釈を行うとき，その推論・解釈を支える証拠の適切性に対する統合的な評価であると述べている。構成概念妥当性は，内容妥当性，基準関連妥当性，信頼性を包含するものであり，それらは構成概念妥当性を示す内容的な証拠，外的証拠，一般化可能性の証拠と考えられるようになった。

3.5 妥当性の統合的なモデルとしての構成概念妥当性

表 1　妥当性評価の 6 つの基準 (Messick (1989, 1996) と小泉(2005a) に基づき作成)

基準の種類	検討する側面	分析手法の例
内容的	・測定しようと意図した領域が明確に定義されているか。 ・テスト項目の形式と内容が，測定しようと意図した領域と一致しているか。 ・テスト項目は，測定しようと意図した領域の代表性を持つか。 ・技術的な質は保たれているか（例： 指示文は明確か)。	・専門家の判断
実質的	・テスト項目は，測定しようと意図した領域のプロセスを適切に引き出すものか。 ・理論的に予測されるプロセスが実際のテスト中に見られるか。	・　観察・質問紙・面接 ・　プロトコル分析
構造的	・テスト得点の構造は，理論的に仮定された構成概念の構造に適合するか。	・項目応答理論，因子分析
一般化可能性的	テスト得点と解釈が，異なるグループ・時間・状況・タスクでも一般化できるか。	・信頼性
外的	複数の外部基準との比較において，収束的・弁別的な証拠が得られるか。	・相関分析，因子分析 ・多特性多方法行列の分析
結果的	得点の解釈と使用について，意図した影響と意図しない影響が長期的・短期的にあるか。	・観察・質問紙・面接

Messick (1989) は，より包括的に構成概念モデルを定義し，構成概念妥当性を妥当性の統合的なモデルとして確立させた。Messick (1989) は，妥当性はいくつかの側面が統合された単一の概念であるとし，妥当性は種類で分けられるのでは

なく，妥当性を示す証拠の種類が複数あるとした。妥当性は，程度の問題であって，有り無しの問題ではないと主張した。また，妥当性はテストの属性ではなく，解釈や使用の妥当性であることを強調し，社会的な影響の側面も妥当性に含まれるとした。妥当性を示す証拠は常に不完全であり，妥当性の確認は絶え間なく続けられる過程であるとした。さまざまな基準から，多面的に妥当性を検討する必要性を主張した。表 1 は，Messick (1989, 1996) が提案した妥当性を評価する 6 種類の基準を筆者が表にまとめたものである。分析手法については，Messick (1989, 1996) で詳しく述べられていなかったため，小泉 (2005a) を参考にした。

図 4 は，表 1 中の多特性多方法行列の例である。例えば，特性 A は依存性，特性 B は温かさ，特性 C は社交性とする。これらの特性を測ることを意図して，アンケート項目が作成されたとする。行列中の対角線上の数値は，アンケート項目の内的一貫性を表し，信頼性の数値となる。また，自己評定で依存性が高いと答えた被験者は，他者評定でも依存性が高いと評価される可能性が高いと考えられるので，同じ特性を測る自己評定と他者評定はある程度の相関があると考えられる。このように，理論的に関連の強い構成概念を測定する指標との相関が高い場合，収束的妥当性があると考えられる。これに対し，異なる特性間の相関は，自己評定・他者評定にかかわらず，信頼性や収束的妥当性ほど高くないと考えられる。このように，理論的に関連の弱い構成概念を測定する指標との相関が低い場合，弁別的妥当性があると考えられる。

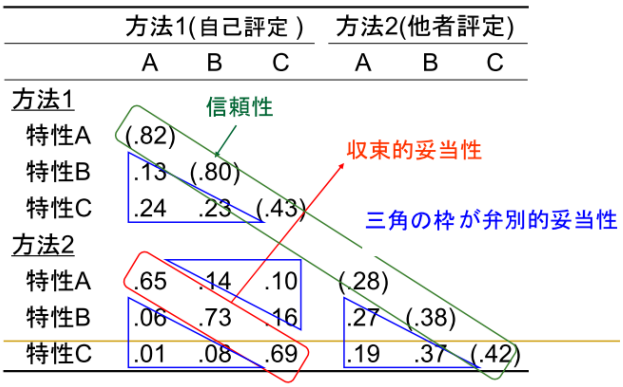

図 4　多特性他方法行列の例 (村上, 2006, p. 33)

信頼性の検討方法について，S の妥当化とも関連が深いと考えられるため，より詳しく検討する。Brown (1996 和田稔訳 1999) によると，テストの信頼性を判定するための基本的な方法は，再テスト法，等価形式法，内部一貫性法の 3 つがある。再テスト信頼性とは，同じテストを同じ受験者集団に 2 回実施し，2 組の得点間のピアソン積率相関係数を求め，この相関係数をテストの信頼性係数と解釈するものと説明している。再テスト信頼性の欠点として，同じテストを同じ受験者集団に 2 回実施するというのは，教育上あまり好ましくないという点を指摘している。等価形式法とは，2 つの異なっているが等価のテストをひとつの受験者集団に対して実施し，2 組の得点間のピアソン積率相関係数を求め，この相関係数をテストの信頼性係数と解釈するものと説明している。等価形式法の欠点として，2 つの形式は異なる等価なテストを開発し，その等価性を証明するのは困難な作業であることを指摘している。それに対し内部一貫性法は，ひとつの形式のテストをただ一度だけ行ってテストの信頼性を測定できるという長所があると述べている。内部一貫性法には，折半法により信頼性を求める方法や，クロンバックの α を求める方法があるが，ここでは概念的に最も理解しやすい折半法について，Brown (1996 和田稔訳 1999) に基づき説明する。この方法ではまず 1 つのテストを 2 つに折半し，2 組の得点について相関係数を算出する。ここで求めた相関係数は，テストが半分の長さの時の信頼性を表す。通常，長いテストの方が短いテストよりも信頼性が高いため，テスト全体の信頼性を推定するためには，スピアマン・ブラウンの予言公式を使用する。この公式は次のとおりである。

$$r' = \frac{n \times r}{(n-1)\, r + 1}$$

上記式で，r' はテスト全体の信頼性を表し，r は半分のテスト間の相関を表し，n はテストの長さを何倍にするべきか

という度数を表す。例えば，2 つに折半したテスト間の相関係数が .6 であった場合，公式の *r* に .6 を代入する。全体のテストは折半されたテストの 2 倍であるため，*n* は 2 となる。このような計算の結果，全体の信頼性は.75 と算出される。*r'* は 1 つのテストを 1 回実施して得たデータに基づく内部一貫性信頼性係数の推定値となる。

4. 妥当化の具体的手続

4.5 仮説検証による妥当化

構成概念妥当性を妥当性の統合的なモデルとして確立させた Messick (1989, 1996) の考え方は，今日的な妥当性のとらえ方である (日本教育心理学会, 2002)。しかし，Messick (1989, 1996) に対する批判もあり，Bachman (2005) は，実際のテスト開発の中で, 妥当化をどのように行えばよいか, ガイドラインを全く提示していないと述べている。Kane (2006) は，Messick (1989, 1996) の妥当性を評価する 6 種類の基準がどのように関連しているか不明確であるとし，どのような妥当性の証拠がより重要で，具体的にどのような順番で妥当化を行っていけばよいか，明確な指針が提示されていないと述べている。

Chapelle (1999) は, Messick (1989) の妥当性を評価する 6 種類の基準に基づき仮説を立て, その仮説を検証することで, テストの妥当化を行うことを提案した。Chapelle が提案する妥当化の手順は，3 段階に分かれる。問題のテストが何を測定しているか, またそのテスト得点をどのように使用できるかについて, テスト作成者は前提を持っていると考えられる。妥当化は，そのような前提に基づき仮説を立てることから始まる。次に，仮説を検証するためのさまざまな証拠を提示する。最後に，仮説検証に使用した証拠や理論的解釈を統合し，証拠に基づく推論とテストの使用に適した妥当性の論証を行うというものである。

Chapelle (1999) の妥当化の手続きに関する提案は，次節で述べる Kane (1992) の論証に基づくアプローチより影響を受け，言語テストを作成・使用する人向けに提案されたものである。この論証に基づくアプローチは Kane (2006) で体系化されているため，今日テストの妥当化を行う研究者が Chapelle (1999) のアプローチを特に採用する必要はないと筆者は考えるが，仮説検証により妥当化を行うことを提案した点が参考になる。

4.6 論証に基づくアプローチ (argument-based approach)

Kane (1992) は, 妥当化の具体的な手続きとして, 論証に基づくアプローチ (argument-based approach) を提案した。Kane, Crooks & Cohen (1999)，Kane (2001)， Kane (2002) で上記の論証に基づくアプローチを発展させ，Kane (2006) で同アプローチを体系化し，妥当化の具体的な手続きについて，詳細かつ具体的に論述した。

Kane (1992) は，妥当性はテストやテスト得点の妥当性ではなく，テスト得点解釈の妥当性であるとし，テスト得点の解釈・使用やそれらの前提・推論を明確にすることからテストの妥当化が始まると説いた。意図する得点の解釈と，それに反駁する解釈を挙げ，意図する得点の解釈を支持し，反駁する解釈を棄却する証拠を挙げることで，テストの妥当化を行うアプローチを提案した。

Kane (2006) は, 論証に基づくアプローチを体系化し, 妥当化の具体的な手続きについて, 詳細かつ具体的に論述した。このアプローチでは，妥当化は 3 段階で行われる。まず，テストが測定対象とする領域を明確にする。次にテスト得点の解釈と使用，それらの前提・推論を明確にすることで，解釈的論証 (interpretive arguments) を行う。最後に，複数の分析と実証をとおし，解釈的論証をさまざまな観点から評価することで，妥当性の論証 (validity arguments) を行う。

Kane (2006) の提案するプレイスメントテスト妥当化の手順について, 英語のテストに当てはめて考えてみる。例えば，作成したプレイスメントテストに基づき，学習者を英語の習熟度別にクラス分けしたいとする。テストで対象とする領域は，英語の運用力とする。作成されたプレイスメントテストの得点は，学習者の英語の習熟度を表すと解釈され，習熟度別クラス編成に使用されるとする。作成したテストの得点を基に習熟度別クラス分けをすることが妥当であるためには，得点の解釈に関する前提・推論が満たされている必要がある。Kane (2006) は，プレイスメントテストの解釈・使用における前提・推論について，採点 (scoring)，一般化 (generalization)，外挿 (extrapolation)，決定 (decision) の 4 種類を挙げている。

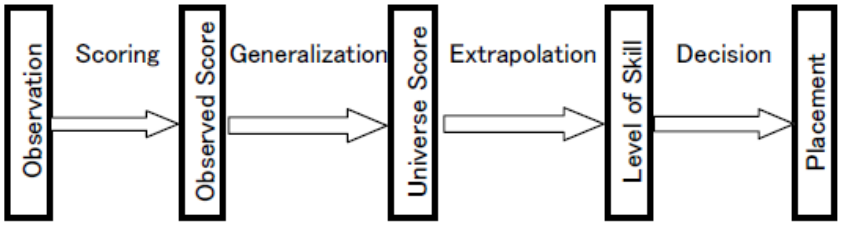

図５　妥当化の手順 (プレイスメントテストの例) (Kane, 2006 に基づき作成)

図 5 は，Kane (2006) に基づき筆者が作成した図である。例えば，得点が適切に解釈されるためには，「採点は適切に行われた」という採点の前提・推論があるだろう。採点の前提・推論が適切であれば，観察されたテスト (observation) からテスト得点 (observed score) を解釈することが妥当となる。また出題されたテスト項目は，測定したい領域の範囲内から偏りなく選ばれており，テスト領域の適切性や代表性を持つという前提・推論があるだろう。このような一般化の前提・推論が適切であれば，テスト得点からテスト領域全体の得点 (universe score) を解釈することが妥当となる。またこのようなテスト領域全体の得点は，学習者がコースの受講に必要な習熟度を表すという前提・推論があるだろう。このような外挿の前提・推論が適切であれば，テスト領域全体の得点から学習者の習熟度を解釈することが妥当となる。さらに，コースで必要とされるスキルのレベルが低い学生は，そのコースで成功しにくいため，習熟度によりクラス分けを行うことが適切であるという前提・推論があるだろう。この決定の前提・推論が適切であれば，学習者の習熟度によりクラス分けを行うことが妥当となる。このような前提・推論を明確にすることが解釈的論証となる。解釈的論証に基づき，複数の分析と実証をとおして解釈的論証をさまざまな観点から評価することで，妥当性の論証を行う。

表２　前提・推論と分析・実証方法の例 (プレイスメントテストの場合) (Kane (2006) と小泉 (2007) に基づき作成)

種類	前提・推論の例	分析･実証方法の例	Messick (1996)
採点	・採点の規則は適切である。 ・採点の規則は正確かつ一貫して適用される。 ・データは採点に用いた尺度モデルに適合する。	・専門家の判断 ・評価者間信頼性	・内容的 ・構造的
一般化	・観察されたパフォーマンスは，一般化したい領域の代表性を持つ。 ・観察されたパフォーマンスは，誤差をコントロールするための十分な大きさがある。	・信頼性	・内容的 ・一般化可能性
外挿	・テストのタスクは，コースで伸ばし，必要とされる能力と関連がある。 ・得点の解釈を深刻に歪めるような，他要因による得点のばらつきはない。	・専門家の判断・相関分析	・内容的 ・実質的 ・一般化可能性 ・外的
決定	・コースでの成績は，受講前に持っていた能力やスキルと関連がある。 ・コースで必要とされるスキルのレベルが低い学生は，そのコースで成功しにくい。 ・コースで伸ばす能力やスキルが既に高い学生は，そのコースを受講してもあまり利益がない。	・決定による肯定的・否定的な影響・結果の分析	・結果的

表 2 は，解釈的論証や妥当性の論証を支える前提・推論や分析・実証方法の例である。表中の「前提・推論の例」は，

Kane (2006, p. 24) の table 2.1 を筆者が日本語に訳したものであり，「分析・実証方法の例」は，Kane (2006) の本文中の解説から筆者がまとめたものである。「分析・実証方法の例」の「データは採点に用いた尺度モデルに適合」は Kane (2006, p. 24) の table 2.1 にはないが，Kane (2006) の本文中では触れられているため加えた。これはテスト採点後に項目応答理論のようなモデルを適用する場合，モデルとデータは適合するという前提があるため，モデルの適合度を調べる必要があるという例である。あらゆるテストについて，「前提・推論の例」や「分析･実証方法の例」であげた項目すべてを検討する必要があるというわけではなく，意図されるテスト得点の解釈・使用やそれらの前提・推論に応じて，そのテストの妥当化に必要な証拠の種類が決められる。Messick (1989, 1996) で提案された 6 つの基準との対応関係は，小泉 (2007) を参照した。

Messick (1989) は，妥当化は絶え間なく続けられる過程であると述べている。Cronbach (1971) も構成概念モデルに基づく妥当化について，果てしなく拡張する調査 (ever-extending inquiry) であると述べているのに対し，Kane (2006) は，妥当化がどこで始まり，どこで終わるかについて手順を示した。また，それぞれの妥当性の証拠がどのように関係しているかについても明確にした。Kane (2006) によると，最も実際的な価値を持つ妥当性の証拠は，解釈的論証において最も問題となる推論と前提を支持するものである。Kane (2006) は，もし論証の中の推論や前提が実証的に支持されない場合，解釈的論証は修正されるか放棄される必要があると述べている。

特性 (traits) を測るテストや理論に基づくテストの妥当化についても，プレイスメントテストの妥当化と基本的な考え方は同じである。小島 (2010) で提案した語彙の豊かさ指標 S は，特性を測るテストであると考えられるため，特性を測るテストの妥当化について，Kane (2006) のアプローチを言語的な特性を測るテストに当てはめて考えてみよう。

例えば，学習者の読み書き能力を測るために，多肢選択式の読解テストを作成したとする。テストで対象とする領域は，読解能力となる。作成されたテストの得点は，学習者の読み書き能力を表すと解釈され，研究目的に使用されるとする。作成したテストの得点を基に学習者の読み書き能力を評価することが妥当であるためには，得点の解釈に関する前提・推論が満たされている必要がある。Kane (2006) は，特性を測るテストの解釈・使用における前提・推論について，採点 (scoring)，一般化 (generalization)，外挿 (extrapolation)，示唆 (implication) の 4 種類を挙げている。図 6 は，Kane (2006, p. 33) の figure 2.2 を簡略化し，筆者が作成した図である。

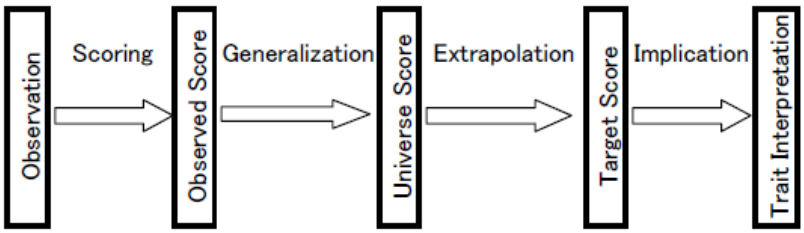

図 6　妥当化の手順 (特性を測るテストの例) (Kane (2006) に基づき作成)

例えば，得点が適切に解釈されるためには，「採点は適切に行われた」という採点の前提・推論があるだろう。採点の前提・推論が適切であれば，観察されたテスト (observation) からテスト得点 (observed score) を解釈することが妥当となる。また出題されたテスト項目は，多肢選択式読解問題という領域の範囲内から偏りなく選ばれており，テスト領域の適切性や代表性を持つという前提・推論があるだろう。このような一般化の前提・推論が適切であれば，テスト得点からテスト領域全体の得点 (universe score) を解釈することが妥当となる。またこのようなテスト領域全体の得点は，学習者の読解能力を表すという前提・推論があるだろう。このような外挿の前提・推論が適切であれば，テスト領域全体の得点から学習者の読解能力の得点 (target score) 解釈することが妥当となる。さらに，学習者の読解能力は書く能力とも密接に関係しているため，読解能力から読み書き能力を推測できるという前提・推論があるだろう。このような示唆の前提・推論が適切であれば，学習者の読解力から読み書き能力全体を解釈 (trait interpretation) することは妥当となる。

表 3 は，特性を測るテストにおける解釈的論証や妥当性の論証を支える前提・推論，分析・実証方法の例である。表中の「前提・推論の例」は，Kane (2006, p. 34) の table 2.2 を筆者が日本語に訳したものであり，「分析・実証方法の例」は，Kane (2006) の本文中の解説から筆者がまとめたものである。「分析・実証方法の例」の「データは採点に用いた尺度モデルに適合」とあるが，これは先にも述べたように，テスト採点後に項目応答理論のようなモデルを適用する場合に，モデ

ルとデータの適合度を調べることが適切であるという例である。あらゆるテストについて，「前提・推論の例」や「分析・実証方法の例」であげた項目すべてについて検討する必要があるというわけではなく，妥当化を行う研究者は，意図されるテスト得点の解釈・使用やそれらの前提・推論に応じて，そのテストの妥当化に必要な証拠の種類を決定する。

表３　前提・推論と分析・実証方法の例 (特性を測るテストの場合) (Kane (2006) に基づき作成)

推論の種類	前提・推論の例	分析・実証方法の例
採点	・採点の規則は適切である。 ・採点の規則は守られる。 ・データは採点に用いた尺度モデルに適合する。	・専門家の判断 ・評価者間信頼性 ・モデル適合度
一般化	・観察されたパフォーマンスは，一般化したい領域の代表性を持つ。 ・観察されたパフォーマンスは，誤差をコントロールするための十分な大きさがある。	・信頼性
外挿	・一般化したい領域の得点 (universe score) は，対象とする領域の得点 (target score) と関連している。 ・得点の解釈を深刻に歪めるような，他要因による得点のばらつきはない。	・質問紙 ・プロトコル分析 ・相関分析 ・因子分析
示唆	・得点から，評価したい特性に関する適切な示唆が得られる。 ・得点の特徴は，評価したい特性の特徴と一致する。	・特性が異なるグループ間で，異なる得点の傾向が見られるか。 ・特性が他の要因と関係があることが分かっていれば，そのような証拠を示す。

Kane (2006) は，さまざまな妥当性の基準がどのように関連しているかを明示し，妥当化の具体的な手続きを述べているため，小島 (2010) で提案した語彙の豊かさ指標 S の妥当化の枠組みとして，この Kane (2006) の論証に基づくアプローチを使用することは適切であると考えられる。S は被験者の特性を測るテストであると考えられるが, Kane (2006) は，プレイスメントテストにおける妥当化の手続きばかりでなく，特性 (traits) を測るテストや理論に基づくテストを例に，妥当化の具体的な手続きを述べているため，S の妥当化のアプローチとして大変参考になると考えられる。

5. S の妥当化

第 3 節，4 節でこれまでの妥当性研究について概観するとともに，テストの妥当性をどう捉え，妥当性検証をどのように行うべきかについて検討した。その結果，S の妥当化には，Kane (2006) の論証に基づくアプローチを使用することが適切であると考えられた。Kane (2006) は，構成概念妥当性を妥当性の統合的なモデルとして確立させた Messick (1989, 1996) の考え方を引き継ぎ，さらにそれぞれの妥当性の基準がどのように関連しているかを明確に示している。Kane (2006) は，特性 (traits) を測定するテストについても具体的な論証の手順を述べているため，S の妥当化の指針となる。

第 4 節で述べたように，Kane (2006) のアプローチでは，妥当化は 3 段階で行われる。まず，対象とする領域を決める。次にテスト得点の解釈と使用，それらの前提・推論を明確にすることで，解釈的論証を行う。最後に，複数の分析と実証をとおし，解釈的論証をさまざまな観点から評価することで，妥当性の論証を行う。

S が測定する対象領域は，「学習者の発表語彙における，一般的な頻度に基づく語彙の豊かさ」である。意図した S の

解釈は，「S は，学習者の発表語彙レベルを表し，S が高いほど，学習者の語彙は豊かである」となる。S の使用目的は，「言語の発達指標の 1 つとして，教育や研究目的での使用」である。

S のスコアがこのように解釈・使用できると考えるためには，さまざまな前提・推論が存在する。本研究では，Chapelle (1999) の提案した仮説検証による妥当化のアプローチを取り入れ，S の解釈・使用の前提・推論を仮説として提示することで，解釈的論証を行う。この解釈的論証は，今後の研究で S の妥当性の論証を行うための基盤となるものである。表 4 は，S の妥当化で用いるべき推論の種類・仮説・分析手法である。表 3 で示したように，Kane (2006, p. 34) では「データは採点に用いた尺度モデルに適合する (The data fit any scaling model employed in scoring)」が前提・推論の例としてあげられていたが，統計を専門とする文献では，「モデルがデータによくフィット」 (金, 2007, p. 140) や「データへの回帰直線の当てはめ」 (間瀬・神保・鎌倉・金藤, 2004, p. 108) という表現が使用されるため，本研究の仮説 2 は，「モデルとする関数式は，エッセイから得られた頻度別累積カバー率のデータに適合する」とした。

表 4　S の妥当化に用いる推論の種類・仮説・分析手法

推論	仮説	分析方法
採点	1. 語彙的エラーチェックの規則は適切であり，チェックにバイアスがかかっていないため，評価者間信頼性は高い。 2. 実際のデータから得られた頻度別累積カバー率は，モデルとする関数式に適合する。	・評価者間信頼性 ・モデル適合度
一般化	3. S はテクストの長さに依存しないため，1 つのエッセイを 2 分して求めた S と，全体での S に有意な差はなく，有意な相関がある。 4. S は信頼性 (内的一貫性) があるため，同一被験者によって書かれたエッセイの前半と後半で，S に有意な差はなく，有意な相関がある。 5. S は異なるタスク間でも信頼性があるため，同一被験者によって書かれたトピックの異なる 2 つのエッセイで，S に有意な差はなく，有意な相関がある。	・相関分析 ・分散分析
外挿	6. 語彙の多様性の指標 (TTR, Guiraud Index, D) と頻度表に基づく語彙の豊かさの指標 (S, LFP, P_Lex) で因子分析を行った場合，後者で 1 つの因子を取り出すことができる。	・因子分析
示唆	7. 学習者の語彙の豊かさは，メンタルレキシコンの語彙サイズと関係があると考えられるため，それぞれの被験者の S は，Productive Vocabulary Levels Test (Laufer & Nation, 1999) や Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001) と有意な相関がある。 8. 語彙の豊かさは，言語の一般的な発達指標の一つであると考えられるため，熟達度の異なる学習者グループや，上級学習者グループと母語話者グループで，S は異なる傾向を示す。	・相関分析 ・分散分析

表 4 の採点に関する仮説が支持されれば，被験者の書いたエッセイから S のスコアを解釈することは妥当であり，一

般化に関する仮説が支持されれば，観察された被験者の S から，その被験者の一般的な S を解釈することは妥当であると考えられる。また，外挿に関する仮説が支持されれば，S から語彙の頻度に基づく語彙の豊かさを解釈することは妥当であり，示唆に関する仮説が支持されれば，S から語彙発達や一般的な言語発達レベルを解釈することは妥当であると考えられる。

6. 結論

本研究は，これまでの妥当性研究について概観するとともに，言語発達指標の妥当性をどう捉え，妥当性検証をどのように行ったらよいかについて考察し，小島 (2010) で提案した語彙の豊かさ指標 S の妥当化のアプローチを提案した。1960～1970 年代では，基準関連妥当性・内容的妥当性・構成概念妥当性は妥当性の 3 つの種類とされ，3 つの妥当性の関係についての体系的な理論は存在しなかったのに対し，Messick (1989) は，包括的に構成概念モデルを定義し，構成概念妥当性を妥当性の統合的なモデルとして確立させた。Kane (2006)は Messick (1989) の妥当性の捉え方を引き継ぎ，具体的な妥当化の手順を示す論証に基づくアプローチを提案した。このアプローチでは，それぞれの妥当性の基準がどのように関連しているかについて，体系的な枠組みが提示されているため，語彙の豊かさ指標 S の妥当化のアプローチとして適切であると考えられる。

この Kane (2006) のアプローチに基づき，S が測定する対象領域や，意図した S の解釈・使用を明示し，それらの前提・推論を仮説として提示することで，S の解釈的論証を行った。今後は，複数の分析と実証をとおし，本研究で行った解釈的論証をさまざまな観点から評価することで，語彙の豊かさ指標「S」の妥当性の論証を行う必要がある。

注

本稿は，2010 年 3 月に名古屋大学大学院国際開発研究科に提出した筆者の博士学位論文『英語学習者の産出語彙における語彙の豊かさ指標 S の提案と論証による S の妥当化』の第４章「テストの妥当化」と第５章「語彙の豊かさ指標 S の提案と S の解釈的論証」を基に加筆・修正を加えたものである。

引用文献

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 14-28). Colchester, England: University of Essex.

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

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall. （ブラウン J. D. 和田稔（訳）(1999). 『言語テストの基礎知識：正しい問題作成・評価のために』東京：大修館書店）

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

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., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139-155.

石川慎一郎 (2008).「英語学習者の語彙知識と語彙産出:構造方程式モデリングを利用した英語エッセイコーパスの解析」『統計数理研究所共同研究リポート 215：学習者コーパスの解析に基づく客観的作文評価指標の検討』 15-28.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

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

Council on Education and Praeger.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18* (2), 5-17.

金明哲 (2007). 『R によるデータサイエンス：データ解析の基礎から最新手法まで』東京：森北出版

小泉利恵 (2005a).「日本人中高生における発表語彙知識の広さと深さの関係」 *STEP Bulletin*, *17*, 63-80.

Koizumi, R. (2005b). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET Journal*, *9*, 5-33.

小泉利恵 (2007 年 4 月). 「より適切なテスト得点の解釈と使用を目指して：妥当性と妥当性検証法」JACET 関東支部月例研究会発表資料（於：JACET 事務所）

小島ますみ (2010)「新しい lexical richness 指標 S の提案 : 学習者の産出語彙頻度レベルの推定」『英語コーパス研究』 第 17 号 1-15 頁

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16* (3), 307-322.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learner's written English*. Malmö, Sweden: CWK Gleerup.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, England: Multilingual Matters.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19* (1), 85-104.

間瀬茂・神保雅一・鎌倉稔成・金藤浩司 (2004). 『工学のためのデータサイエンス入門』東京：数理工学社

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16* (3), 5-19.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). New York: American Council on Education and Macmillan.（リン R. L. (編) 池田央・藤田恵璽・柳井晴夫・繁桝算男（訳）(1992). 『教育測定学原著第 3 版』上巻 横浜：みくに出版）

Messick, S. A. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-256.

水本 篤 (2008) 「自由英作文における語彙の統計指標と評定者の総合的評価の関係」『統計数理研究所共同研究リポート 215：学習者コーパスの解析に基づく客観的作文評価指標の検討 』15-28.

村山航 (2006). 「妥当性概念の展開」『テストの妥当性の概念および検証方法の新たな展開』日本テスト学会公開シンポジウム発表資料 <www.p.u-tokyo.ac.jp/~murakou/validity.ppt/>

日本教育心理学会（編）(2002). 『教育心理学ハンドブック』東京：有斐閣

Richards, B. J., & Malvern, D. D. (2000). Accommodation in oral interviews between foreign language learners and teachers who are not native speakers. *Studia Linguistica*, *54* (2), 260-271.

杉浦正利 (2008). 「英文ライティング能力の評価に寄与する言語的特徴について」成田真澄（代表） 『学習者コーパスに基づく英語ライティング能力の評価法に関する研究』(pp.33-58). 平成 17 年度～平成 19 年度科学研究費補助金（基盤研究（C））研究成果報告書 (課題番号 17520394)

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17* (1), 65-83.

（提出日 平成 25 年 1 月 11 日）