第二言語としての日本語の習得研究
第 18 号, 2015 年 12 月, pp.53-69.

〔特集論文〕

テスト分析に基づく「SPOT」と「J-CAT」の比較

李 在鎬（筑波大学），小林 典子（元筑波大学）
今井 新悟（筑波大学），酒井 たか子（筑波大学）
迫田 久美子（国立国語研究所）

要旨

「SPOT（Simple Performance-Oriented Test）」と「J-CAT（Japanese Computerized Adaptive Test）」の特徴を論じ、両テストの共通受験者の得点データを統計的な方法で分析し，両テストの相互関連性を明らかにした。分析の結果，1）2 つのテストの合計得点間には高い相関（r=. 86）が認められること，2）「SPOT90-2」の得点は，「J-CAT」の「Reading」（r=.74）と「Listening」（r=.76）の得点と相関していること，3)「SPOT」の得点によって評価される初級学習者，中級学習者，上級学習者の認定基準は，「J-CAT」の認定基準とほぼ対応していることが明らかになった。

キーワード：熟達度テスト，定量的分析，言語運用と言語知識，SPOT，J-CAT

1. はじめに

「SPOT（Simple Performance-Oriented Test）」WEB 版[1]（小林・フォード・山元, 1996；小林, 2015）と「J-CAT（Japanese Computerized Adaptive Test）」（今井・赤木・中園, 2012）は，どちらもコンピュータベースの言語テストであり，受験申込からテスト結果の確認までの全プロセスをウェブブラウザ上で行える。また，どちらのテストも熟達度テストであることから，幅広い日本語学習者の言語能力評価に活用できる。一方，両テストは異なる背景のもとで開発された。「SPOT」は比較的短時間で言語運用力を測る目的で開発されたが，「J-CAT」は言語運用力と言語知識の両方を効率的に測る目的で開発された。そのため，テスト形式，テストセットの構成，配点，得点解釈の部分では相違が見られるが，両者の関連について明らかにした研究は存在しない[2]。研究が進まなかった原因としては，次の 2 つが考えられる。1 つ目は，データ収集の困難さ，2 つ目は，分析手法の困難さである。1 つ目の課題として，異なるテストを比較するためには，言語能力の同一性を保証するために同時期に 2 つのテストを受けてもらう必要がある。しかも，結果の信頼性を確保するためには，相当量の受験者が必要になる。2 つ目に，構成概念も得点値も異なるテストなので，得点を直接比較することはでき

ない。この場合は，2 つのテストの共通軸を求めてから，個々のテスト得点を解釈する必要がある。

以上の課題を踏まえ，本研究では，次の方法で調査を行った。1 つ目の課題に対して，「SPOT」と「J-CAT」の共通受験者の 201 名分の得点データを用いて，両テストの関連性を明らかにする。2 つ目の課題に対して，多変量解析の方法で，2 つのテストの共通軸を構築し，それをもとに受験者の能力を再評価する。具体的には，中国，トルコ，ロシア，オーストリアの 4 か国における 201 名の共通受験者の得点データを多次元尺度構成法，相関分析，クラスター分析で解析し，両テストにどのような関連性が見られるかを明らかにする。本研究の分析を通じて，両テストの関連性を明らかにし，受験者に対する情報提供を行うと同時に，テスト結果を利用する側にも広く情報提供を行っていきたい。

2. 先行研究

2.1 「SPOT」と「J-CAT」

「SPOT」と「J-CAT」はいずれも科学研究費助成事業などの外部研究資金による補助を受け，複数の研究者および協力者によって開発がなされた。現在は，ウェブ上で無償公開されており，適切な速度のインターネット回線に接続されたコンピュータ環境があれば「誰でも，どこからでも，いつでも」利用できる。

両テストに共通する開発動機としては，90 年代以降，留学生の数が急増したことがあり，プレースメントテストを改善するという喫緊の課題があった。多くの日本語教育の現場では，履修希望者に対して日本語のレベル分けを目的とするプレースメントテストを行うことになるが，いわゆる紙を使ったテスト（Paper Based Test; 以後 PBT）の場合，作成と実施にかかる労力は相当なものである。またプレースメントテストの場合は短期間でクラス分けを行わなくてはならず，採点とレベル判定の負担も大きい。こうした課題を解決する手段としてコンピュータベースのテストが注目された。

「SPOT」WEB 版と「J-CAT」はいずれもコンピュータテストとして開発されたものであるが，コンピュータを使ったテスト（Computer Based Test；以下，CBT）の場合，以下のようなメリット・デメリットがある（李（編），2015；25）。

(1) メリット
① 得点の集計が自動的に行われる
② 多様なコンテンツが利用できる
③ 受験者の能力に合わせた問題提示ができる
④ 場所や時間の制約が少ない

(2) デメリット
① コンピュータ操作の慣れ具合が得点に影響を与える
② システム開発に巨大な費用がかかる
③ 受験環境が整ったところでしか実施できない
④ パフォーマンステストの場合，自動採点の誤差の幅が大きい

CBT には，メリットとデメリットがあり，PBT に対して絶対的に優位であるとは必ずしも言えないが，CBT が持つ強みについては，評価すべき部分が多いのも確かであると言えよう。こうしたことが，「SPOT」WEB 版と「J-CAT」を開発した背景であったことは間違いない。

2.2 「SPOT」の特徴

「SPOT」の特徴は，以下の 3 点としてまとめることができる。

(1) 言語運用能力を間接的・客観的に測定するテストである。
(2) 短時間で実施できるテストである。
(3) 能力差が比較的大きな集団を 2～4 段階程度の能力別グループに分けるテストである。

「SPOT」は自然な速度で読み上げられる 1 文ずつを聞きながら，1 カ所の空欄に平仮名 1 文字を挿入するという形式の問題で，それぞれの文は独立しており，互いに関係がない。これを数十問，数分で行うことで，おおよその日本語能力のレベルが判別できる。これは学習者の日ごろの聞き取りの失敗をヒントにして，「知識のない者は，たとえ音声を与えられても正確に書き取れない」という仮説から出発したテストである。無限にある語彙を対象とせず，言語使用において必然性の高く，音声的には弱形で読み上げられる文法項目部分に注目し，そのひらがな 1 文字（採点がしやすい）の書き取りを課したものである。自然な速度で即時的に処理していけるかどうかが，言語運用力と関連が深いと考え，このテストは Performance-Oriented だと主張してきている（小林他, 1996）。

1990 年代は紙と音声テープを利用していたが，現在は「Tsukuba Test Battery of Japanese（TTBJ: http://ttbj.jp/）」の中のテストの一種類としてあり，ウェブブラウザを使ったコンピュータテストになっている。そして，解答方法も記入ではなく平仮名 1 字の四肢選択となっている。なお，本研究が利用した「SPOT90」と称しているものは，紙版の「SPOT」と問題数及び構成で異なる[3]。「SPOT90」は 30 問ずつの「SPOT90-1」「SPOT90-2」「SPOT90-3」で構成されており，それぞれのテストセットの難易度は次のように設定されている。まず，「SPOT90-1」は初級向けで，日本語能力試験の N5-N4

レベルの学習者を対象にしている。「SPOT90-1」の音声は声優による明瞭なものを使用している。次に,「SPOT90-2」は初級後半から上級前半向けで,日本語能力試験のN4-N2 レベルの学習者を対象にしている。「SPOT90-2」の音声は日本語教師のやや不明瞭なものを使用している。最後に,「SPOT90-3」は上級向けで,日本語能力試験のN2-N1 レベルの学習者を対象にしている。「SPOT90-3」の音声は日本語教師の自然なものをパソコン上で音を圧縮して聞き取りにくくなるような加工を施している。

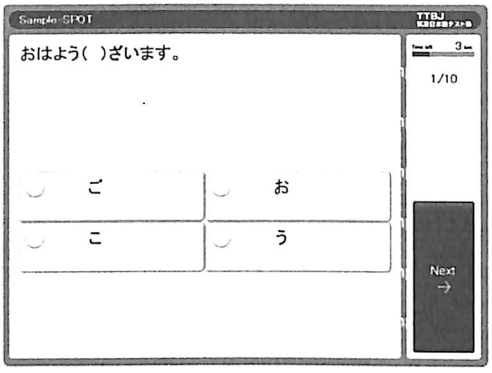

図 1 「SPOT90」の問題提示画面

受験者は,コンピュータ画面上で,図 1 のような問題提示画面を見ながら解答する。1 画面に 1 問が表示され,同時に読み上げ音声が流れ,四肢選択の 1 つをクリックして解答するが,解答時間は 4 秒に設定されており,それを過ぎると入力できなくなる[4]。4 秒以内であれば,選択し直すこともできるが,時間切れになった場合,最後に選択していたものが解答として格納される。ゆっくり考えている時間はなく,従来の文法や語彙の知識テストとは異なり,自然な読み上げ速度に合わせて読みながら聞きながら、選択するといった「ながら作業」ができるかどうかが得点を左右する。これは言語処理の自動化の程度と関係する。採点は 3 セクションごとに 1 問 1 点の正答数得点で画面に表示され,最後に総得点数も表示される。

総合得点に対しては,表 1 をもとに解釈を行う。

表 1 「SPOT」のレベルの目安

得点	レベル	説明
0～30	入門	日本語を学習したことがほとんどない。
31～55	初級	ゆっくりであれば日常生活の基本的な日本語を理解できる。
56～80	中級	自然な発話速度で日常的な場面の日本語がある程度理解できる。
81～90	上級	自然な発話速度で幅広い場面の日本語が理解できる。

2.3 「J-CAT」の特徴

「J-CAT」の特徴は，以下の3点としてまとめることができる。

(1) 項目反応理論によるテストである。
(2) 適応型テスト（adaptive test）である。
(3) 1 つのアイテムプールを用いて，すべてのレベルの日本語力を測るテストである。

「J-CAT」は，項目反応理論を使った適応型テストである。項目反応理論とは，問題項目への解答のパターンに基づいて、問題項目の特性（困難度・識別力）から受験者の特性（能力）を測定するためのテスト理論である。これを用いることのメリットとして次のことが挙げられる。項目反応理論によるテストでは，素点を用いる古典的テスト理論における点数の歪みを補正することができる。例えば，100 点満点のテストでの 50 点と 55 点の 5 点の差と，95 点と 100 点の 5 点の差の価値が異なるであろうことは，直感的に感じられるだろう。これが点数の歪みである。項目反応理論はこのような点数の歪みを補正する。また，異なる問題を解いても算出される得点の価値を同じにできる。これを得点の等化と呼ぶ。項目反応理論は例えば新しくなった日本語能力試験にも使われているが，紙に問題を印刷する日本語能力試験では，問題は固定される（日本語能力試験；https://www.jlpt.jp/reference/pdf/guidebook1.pdf）。しかし，「J-CAT」では，コンピュータが受験者の能力を随時推定しながら，問題の難易度を変化させる。これにより，受験者の能力レベルとかけ離れた，難しすぎる問題，あるいは易し過ぎる問題を出題せず，受験者のレベルを測るのに適した問題を出題することができる。これが適応型テストと呼ばれるものである。このようにして，アイテムプールと呼ばれるデータベースにストックされた問題のうち，各受験者に適した問題が自動的に選ばれて，出題されるので，日本語能力試験のようにレベルごとに異なる問題のセットをあらかじめ準備する必要がない。これにより，得点がワンスケールで与えられる。つまり，初級から上級・超級までが 0 点から 400 点の間の得点で示されるため，能力の比較が容易である。日本語能力試験のようにレベルごとに異なる問題セットで試験をした場合，例えば，N3 を受験した高得点者と N2 を受験した低得点者のレベルの比較は難しい。

「J-CAT」 は聴解，語彙，文法，読解の 4 セクションで構成されている。アカデミックジャパニーズのように特化した日本語能力ではなく，一般的な日本語能力を測る。語彙および文法は言語知識を測り，聴解，読解は受容能力を測る。「J-CAT」はコンピュータを用いるテストであるため，多様なコンテンツを利用したテストができる。例えば，図 2 のような動画を使った問題がある。

問題)　何をしていますか。

a. あくしゅをしています。　b. はくしゅをしています。
c. きょしゅをしています。　d. はしゅをしています。

図2　「J-CAT」の語彙テスト例

図2は動画を使ったテスト例であり，実際はカラーの動画が表示され，握手をしている様子が分かるようになっている。このような動作やオノマトペを問うには紙媒体のテストより，動画が適していると言える。

「SPOT」の受験者が用意されている90アイテムの問題を約10分で解いていくのに対して，「J-CAT」は前述の適応型テストの仕組みを取り入れているため，受験者によって受験時間も異なり，おおよそ45分～90分程度かかる。総合得点をもとにした能力評価は表2に示すとおりである。

表2　「J-CAT」のレベルの目安

得点	レベル	説明
0-100	初級	基本的な考えを述べることができる
101-150	中級前半	日常的な会話をこなすことができる
151-200	中級	
201-250	中級後半	
251-300	上級前半	学術的・専門的なコミュニケーションができる
301-350	上級	
351-400	日本語母語話者相当	

3.　データと分析方法

本研究では，同時期に「SPOT90」と「J-CAT」を受けた受験者の得点データを定量的な方法で分析し，両テストの比較を目指す。分析データは，日本語の学習者コーパ

ス構築のためのプロジェクトによって収集されたものである[5]。このプロジェクトは，海外 20 の地域から 12 の異なった言語を母語とする日本語学習者の発話・作文コーパスの構築を主目的としているが，コーパスを構築するにあたって客観的指標に基づく言語能力の測定を重要視している。というのは，コーパスのデザインの段階から複数の言語テストを用いて学習者（データ提供者）の言語能力を測定した上で，データ収集を行うことになっており，学習者コーパスの信頼性・妥当性を言語テストで確保することを目指しているからである。

テストデータの収集は，日本語母語話者による約 80 分の対面調査と約 60〜100 分のコンピュータによる日本語能力調査の 2 部構成の調査により、次の手順で行われた。

【対面調査】調査者（日本語母語話者）は日本語学習者と 1 対 1 で対面調査をして対話とライティングのデータを収集した[6]。
【パソコン調査】受験者は対面調査の終了後、複数台のパソコンが設置してある別室で、監督者のパソコン入力の説明を受けた後、以下の日本語能力テストを受けた。
① 「SPOT90」:「SPOT90-1」「SPOT90-2」「SPOT90-3」の合計 90 問
② 「J-CAT」: 聴解、語彙、文法、読解の 4 つのセクション

本稿で分析の対象としたものは中国（上海），ロシア，トルコ，オーストリアの 4 か国で収集された合計 201 名のデータである。受験者属性を表 3 に示す。

表 3　受験者属性

国	受験者数	平均年齢	日本語の平均学習歴
中国	54 名	22.4 歳	24.6 ヵ月
ロシア	52 名	21.8 歳	29.4 ヵ月
トルコ	52 名	23.1 歳	22.4 ヵ月
オーストリア	43 名	24.8 歳	36.6 ヵ月

全受験者の得点は，次の得点区分別に集計した。「SPOT90」については，「SPOT90-1」（30 点満点），「SPOT90-2」（30 点満点），「SPOT90-3」（30 点満点），「SPOT90」合計得点（90 点満点）である。「J-CAT」については「Listening」（100 点満点），「Vocabulary」（100 点満点），「Grammar」（100 点満点），「Reading」（100 点満点），「J-CAT」合計得点（400 点満点）である。これらのデータに対して，三つの分析を行った。なお，すべての統計分析は，「IBM SPSS Statistics（Ver. 22）」で行った。

(1) 分析 1 : テストセット間の類似性を検討すべく，多次元尺度構成法による分析を行った。

(2) 分析 2：得点間の相関を見るべく，ピアソンの積率相関係数を計算した。
(3) 分析 3:受験者のグループ分けを行うべく,階層的クラスター分析を行った。

分析 1 では，テスト間の関連性の強さを距離化し，相互類似性を検討するため，多変量解析の方法である「多次元尺度構成法（Multi-Dimensional Scaling, MDS)」による分析を行った。分析 2 では，「SPOT90」と「J-CAT」の合計点だけでなく，「SPOT90」のサブセクション，「J-CAT」のサブセクション間の相関係数を計算し，テストセット間の関連の強さを確認した。分析 3 では，受験者のセグメンテーションを行うべく，ウォード法による階層クラスター分析を行った。その結果をもとにテストセットの特徴や得点分布を確認した。

本研究で，上述の多変量解析を用いた一番の理由は，2 節で確認した通り，「SPOT」と「J-CAT」は得点の出し方や解釈にかなりの違いがあり，点数を直接比較することはできないため，それぞれの得点をどちらのテストとも独立した方法で標準化し，その上で，各受験者の能力を捉えるべきだと考えたからである。特に分析 3 では，クラスター分析を行って，「SPOT」と「J-CAT」の両方の得点をもとに，「上」「中」「下」のグループを作成した上で，「SPOT」と「J-CAT」の得点を再解釈する。こうすることで，2 つのテストの対応を中立的に分析できる。

4. 結果

具体的な分析を行う前に，全体の項目統計量の計算および信頼性統計量を確認した。

表 4　項目統計量

テスト名	サブテスト	最小値	最大値	平均値	標準偏差	分散
J-CAT	Listening	8	86	50.49	16.916	286.141
	Vocabulary	0	100	50.25	19.067	363.538
	Grammar	8	91	44.73	17.065	291.230
	Reading	3	82	43.50	13.173	173.531
	合計得点	48	354	188.97	56.143	3152.009
「SPOT」	「SPOT90-1」	11	30	28.05	3.213	10.322
	「SPOT90-2」	5	30	21.07	5.209	27.129
	「SPOT90-3」	5	30	16.30	5.324	28.340
	合計得点	26	90	65.43	12.130	147.126

表 4 では，「J-CAT」と「SPOT」の得点の項目統計量を示した。表 4 で注目すべき

点として「J-CAT」の場合，ワンスケールで全レベルの学習者の日本語力を測定するものであるため，分散も大きく，全体の平均値そのものは，それほど意味を持つものではない。一方，「SPOT」の場合，相対的に標準偏差が小さく，平均値の解釈において一定の傾向が観察される。とりわけ，「SPOT90-1」に注目した場合，30 点満点で 28.05 という平均値が出た。これは，ほとんどの受験者が満点に近い得点をとっていることを意味しており，「SPOT90」がサブセクションによって測定対象が異なることに照らし合わせて考えてみた場合，本研究の受験者は初級を終了した学習者がほとんどであることを示唆している。なお，信頼性統計量として，クロンバックのα係数を計算したところ，「0.843」となり，αの数値的には信頼できるデータであると判断できる。

以上の結果を踏まえ，具体的な分析を行った。まず，分析 1 として，受験者得点をもとに多次元尺度構成法[7]による分析を行った。距離行列の作成における測定方法はユークリッド平方距離を使い，値の標準化として 0 から 1 の範囲への尺度化を行った。

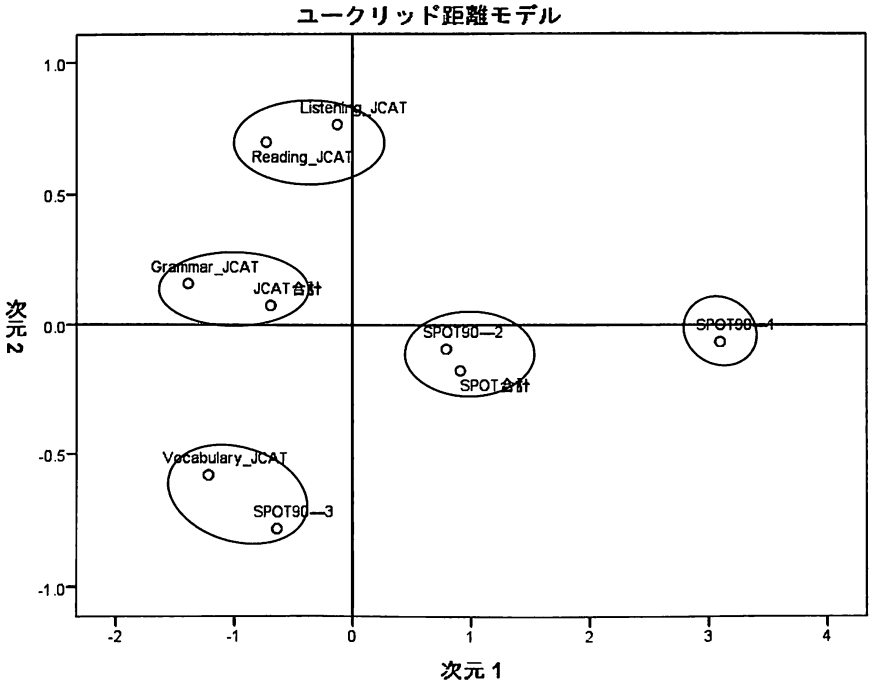

図3　多次元尺度構成法によるテストセットの布置

モデルの適合度を示す stress 値は「.037」，決定係数の RSQ 値は「.995」で妥当性の高いモデルであると判断できる[8]。図 3 から確認できる事実として，1）言語運用力を示す Listening （「J-CAT」）と Reading（「J-CAT」）の得点分布に類似性が見られること，2）Grammar（「J-CAT」）の得点と「J-CAT」合計の得点分布に類似性が見られること，

3)「SPOT90-3」は，Vocabulary（「J-CAT」）の得点と類似性が見られることが明らかになった。一方，4)「SPOT90-1」はいずれのテストとも類似していないことも明らかになった。さらに，「SPOT90」の合計得点は，「SPOT90-2」と得点分布において類似していることが明らかになった。

「SPOT90」はその開発趣旨において運用力を測ることを目的に作られたものであるが，「J-CAT」の Listening や Reading とはどこまで関連していると言えるだろうか。このことを検討すべく，ピアソンの積率相関係数による 2 変量間の相関分析を行った。結果を表 5 に示す。

表 5　相関分析の結果

	Listening*	Vocabulary*	Grammar*	Reading*	「J-CAT」合計	「SPOT 90-1」**	「SPOT 90-2」**	「SPOT 90-3」**	「SPOT90」合計
Listening*	1								
Vocabulary*	.52	1							
Grammar*	.60	.76	1						
Reading*	.66	.58	.61	1					
「J-CAT」合計	.82	.86	.89	.81	1				
「SPOT90-1」**	.59	.62	.63	.56	.71	1			
「SPOT90-2」**	.76	.69	.68	.74	.84	.75	1		
「SPOT90-3」**	.58	.67	.58	.56	.71	.49	.71	1	
「SPOT」合計	.74	.75	.71	.71	.86	.81	.94	.87	1

*：「J-CAT」/ **：「SPOT」

表 5 で注目すべき点はほとんどのテストセット間で 0.6～0.8 の高水準で相関がある点である。特に「J-CAT」と「SPOT」の合計点においては，r=.86 の高い相関が見られた。次に，個々のテストセット間の相関係数に注目した場合，1)「J-CAT」の Vocabulary と Grammar において高い相関（r=.76）が見られること，2)「SPOT90-2」は「J-CAT」の Listening（r=.76）と Reading（r=.74）の高い相関が見られること，3)「SPOT90-1」と「SPOT90-3」で相関が低かったこと（r=.50）が明らかになった。

最後に分析 3 として，階層的クラスター分析を行い，受験者集団のグループ分けを行った。入力データとしては，各テストセットの得点を使った。ケース間の距離は「ユークリッド距離」を使用し，クラスター法は「ウォード法」を使用した。最適なクラスター数の判断においては，李・井佐原（2006）の方法論を導入し，判別分析を使うことにした。これは，クラスター分析によって出力した所属クラスターを従属変数に，テストセットの得点を独立変数にし，正準判別分析を行うものである。

表6　判別分析による正答率

クラスター数	判別率	交差確認済み判別率
C=3（3つの集団）	97.5%	97.0%
C=4（4つの集団）	94.5%	92.5%
C=5（5つの集団）	94.5%	92.0%

表6は，クラスターの数に応じた判別分析の判別率（観測グループに対して予測グループの正答率）とleave-one-out法による交差検証を行った判別率を示したものである。3つの集団として分けた場合の判別率は，97.5%であり，もっとも高い精度で観測グループに対して予測グループが一致していることを表す。さらに交差確認済み判別率においても97%であり，3つのクラスターとしてとらえるのが最も妥当な分析であることが明らかになった[9]。

3つのグループの得点を散布図にしたものが図4である。3つのグループを便宜上「上グループ」「中グループ」「下グループ」としてラベリングした[10]。

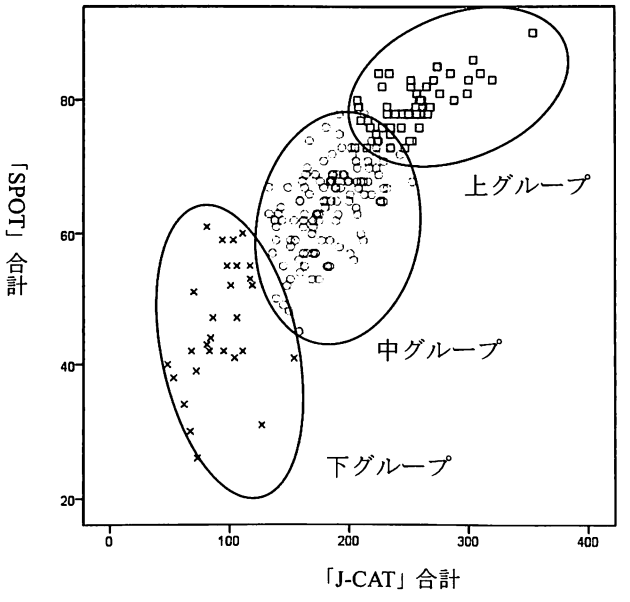

図4　グループの散布図

図4において「□」でマーキングされているのは，比較的能力が高い上グループ（52名）で，「○」でマーキングされているのは，中間に位置する中グループ（121名），「×」でマーキングされているのは，能力が低い下グループ（28名）である。

表 7　「SPOT90」と「J-CAT」の得点分布

「SPOT90」	ｸﾞﾙｰﾌﾟ	平均値	標準偏差
「SPOT90-1」(30 点満点)	上	29.81	0.45
	中	28.78	1.31
	下	21.68	4.24
	全体	28.05	3.21
「SPOT90-2」(30 点満点)	上	26.19	2.05
	中	20.83	3.42
	下	12.64	4.08
	全体	21.07	5.21
「SPOT90-3」(30 点満点)	上	23.06	2.75
	中	14.52	3.67
	下	11.43	3.01
	全体	16.30	5.32
「SPOT」合計(90 点満点)	上	79.06	3.80
	中	64.12	6.79
	下	45.75	9.52
	全体	65.43	12.13

「J-CAT」	ｸﾞﾙｰﾌﾟ	平均値	標準偏差
Listening (100 点満点)	上	65.71	13.42
	中	49.12	12.13
	下	28.18	12.41
	全体	50.49	16.92
Vocabulary (100 点満点)	上	68.83	11.09
	中	49.10	11.46
	下	20.71	17.37
	全体	50.25	19.07
Grammar (100 点満点)	上	62.85	11.70
	中	43.00	10.45
	下	18.54	7.50
	全体	44.73	17.07
Reading (100 点満点)	上	55.40	10.37
	中	42.64	8.96
	下	25.11	9.93
	全体	43.50	13.17
「J-CAT」合計 (400 点満点)	上	252.79	30.82
	中	183.86	27.51
	下	92.54	24.17
	全体	188.97	56.14

「SPOT90」と「J-CAT」の得点分布を示した表 7 で注目すべき点は，「SPOT90-1」の場合，上グループと中グループの平均値は，いずれも満点の 30 点に非常に近い値になっているだけでなく，標準偏差も小さい値になっていることである。この結果からは，上グループと中グループのほとんどが満点に近い得点をとっており，能力の識別には貢献していないことが見てとれる。「SPOT90-2」の場合，上・中・下がほぼ等間隔に分布しており，言語能力をうまく説明していることが分かる。「SPOT90-3」の場合，上グループの平均が 23.06 であるのに対して，中・下グループは，いずれも 50%未満の得点率になっており，平均得点も近い。次に，「J-CAT」の得点に注目した場合，いずれのセクションにおいても，上グループの平均得点は，60 点前後，中グループの平均得点は 40 点前後，下グループは 20 点前後で収まっており，ほぼ等間隔の得点差になっている。最後に「SPOT」の合計得点と「J-CAT」の合計得点のグループ間の平均を確認した。上グループの「SPOT」の平均得点は 79.06 点，「J-CAT」は 252.79 点，中グループの「SPOT」の平均得点は 64.12 点，「J-CAT」は 183.86 点，下グループの「SPOT」の平均得点は 45.75 点，「J-CAT」は 92.54 点であった。両者の対応に関する

より詳細な考察は5章で行う。

5. 考察

本研究の調査によって，異なる背景と目的によって開発された2つのテストにおける相互関連性が明らかになった。しかし，両テストとも日本語教育の分野では長年の運用実績があり，テストとしての信頼性が報告されている（小林, 2015; 今井・赤木・中園, 2012)。したがって，両テスト間の関連性自体は，十分に予測可能なことであり，特筆に値する発見とは言えない。しかし，「SPOT」をめぐる観察に関して，以下の2点については考察が必要である。

(1) 「SPOT90-2」が「J-CAT」のすべてのテストセクションと相関が高いのはなぜか。
(2) 同じテスト形式である「SPOT90-1」と「SPOT90-3」の相関が最も低い（r=.495）のはなぜか。

これら2点の現象の背景として，「SPOT」は，設計段階において，異なる時期に独立に作題されたということが関係する。というのは，「SPOT」は，当初は「SPOT90-2」に格納されている30問でもって全レベルを測定する目的で作題された。しかし，「SPOT90-2」は初級レベルの学習者にとって難しすぎたため，問題項目を初級項目に限定し，文体も「です・ます」体のみとして「SPOT90-1」を作題した。そして，「SPOT90」のWEB化をきっかけに，N1レベルの測定精度を持ったテストの必要性から，「SPOT90-3」を作題した。一方，「J-CAT」の場合，ワンスケールで全レベルを測定するように設計されている。これらの事実から「SPOT90-2」と「J-CAT」の得点に相関が高いのは，測定対象が大きく重なっていることに起因するものと言えよう。同時に，「SPOT90-1」と「SPOT90-3」の相関が低いのは測定対象が大きく異なっていることに起因する問題であると言える。

最後に，合計得点に対して，それぞれのテストのスコア解釈に当てはめて考えてみたい。「J-CAT」では，0〜100点が初級，100〜150点が中級後半，150〜200点が中級，200〜250点が中級後半，250〜300点が上級前半，300〜350点が上級，350〜400点が母語話者相当とされている。「SPOT90」の場合，0〜30点が未習者，30〜60点が初級，60〜80点が中級，80〜90点が上級とされている。これに合わせて表7の受験者の得点幅を検討する。

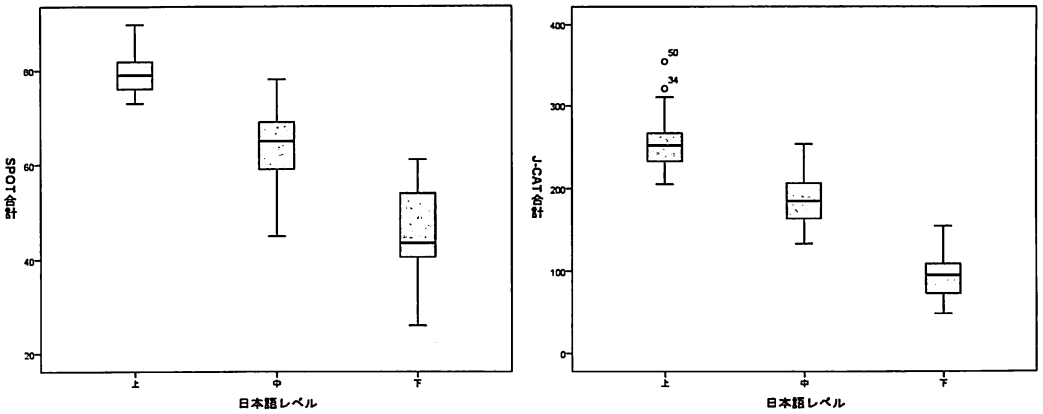

図5　SPOTの合計得点と「J-CAT」の合計得点の箱ひげ図

図5を見ると，左側の「SPOT」に関しては下グループはほぼ初級学習者と解釈できる。中グループは中級と初級が混ざった分布になっている。上グループは上級と中級が混ざった分布になっている。次に，右側の「J-CAT」に関しては下グループはほぼ初級学習者と解釈できるが，一部は中級前半の学習者が混ざっている。中グループは中級から中級後半にまたがった分布になっている。上グループは300点を超えた学習者が4名いたが，それ以外に関しては中級後半から上級前半にまたがった分布になっている。表7および図5の結果を総合して考えてみた場合，表8の対応が考えられる。

表8　「SPOT」と「J-CAT」による能力判定の比較

「SPOT」による能力判定	「J-CAT」による能力判定
初級学習者	初級学習者
	中級前半学習者
中級学習者	中級学習者
	中級後半学習者
上級学習者	
	上級前半学習者
?	上級学習者
	日本語母語話者相当学習者

表8のとおり，「SPOT」で初級と判定される学習者は，「J-CAT」で判定される初級

学習者と中級前半学習者とほぼ対応するものと考えられる。また,「SPOT」で中級と判定される学習者は,「J-CAT」で判定される中級から中級後半の学習者とほぼ対応する。そして,「SPOT」で上級と判定される学習者は,「J-CAT」の中級後半から上級前半の学習者とほぼ対応する。なお,表 8 で「?」にしたのは,「SPOT90」の得点としては上級学習者であるが,「J-CAT」の得点として上級学習者と判定されるケースが 3 名,日本語母語話者相当学習者と判定されるケースが 1 名いたが,データ量として信頼できる数ではなく,確定的なことは言えないためである。「J-CAT」の上級学習者および母語話者相当学習者の「SPOT90」との対応関係については,データ数を増やした上で再調査する必要があり,今後の課題としたい。

6. まとめと課題

本研究では,「SPOT」と「J-CAT」の同時受験者の得点データを定量的に分析することで,2 つの異なるテストの比較を行った。その結果として,両テストは,高いレベルで相関が見られることが明らかになり,得点による日本語レベルの測定においても対応があることが明らかになった。

今後の課題として,2 点述べる。1 点目は,今回の調査では海外の学習者 201 名に対して同時受験を実施し,分析を行ったが,より多様な受験者属性に対応した調査が必要と考える。具体的には調査地の追加や国内学習者との比較などで,今回の調査結果を再検討することも必要であると考える。2 点目は,テスト結果と実際のインタビューのやり取りでのコミュニケーション能力との関係を分析する必要がある。具体的には,本稿で示した「上,中,下」の学習者で,どのような言語能力の特徴が見られるか明らかにしたい。

注

1. 「SPOT」は開発の過程で紙版のものなど,複数のバージョンが存在するが,本研究では「SPOT90」WEB 版を使用した。
2. SPOT と ACTFL OPI との対応については,岩崎 (2002) を参照してほしい。J-CAT とプレースメントテストの関連性については,今井・赤木・中園 (2012) を参照してほしい。
3. 紙版「SPOT」の問題構成,及び,構成概念については小林他 (1996), Ford-Niwa & Kobayashi (1999) 等参照。「SPOT90-1」は紙版「SPOT」の ver.B と、「SPOT90-2」は ver.2 と同等レベルである。
4. 解答時間の設定は変更可能で、本稿のデータとしたものは 4 秒であった。
5. 本プロジェクトは、科学研究費助成事業基盤研究 (A)「海外連携による日本語学習者コーパスの構築ー研究と構築の有機的な繋がりに基づいて」(課題番号 : 24251010) (研究代表者 : 迫田久美子) である。
6. 対面調査は,以下の順でタスクを行った。

①ストーリーテリング2種：4〜5コマの絵を見て自由に物語を作って話す課題
②対話：調査者のウォーミングアップ的な会話から始まり、雑談のような雰囲気の中で、ある程度構成された話題や質問によって進められる30分の対話
③ロールプレイ：調査者と学習者による「依頼」と「断り」の機能の2種のロールプレイ
④ストーリーライティング：①のストーリーテリングを再度、作文し、パソコンに入力する

7. 多次元尺度構成法とは入力データからボトムアップ的に類似関係を表示する方法である。とりわけ，2次元ないしは3次元の図上で類似している者同士は近くに，類似していないものは遠くに配置し，サンプル間の親疎関係を直感的に把握するための分析手法である。
8. 多次元尺度構成法の精度を判断するベースラインとしてstress値は，0.05以下，RSQ値は0.6以上が妥当とされている。
9. 念のため一元配置の分散分析でクラスター間の得点差に統計的な有意味性があるかを確認した。分析では所属クラスターを因子にしてすべての得点の平均値に有意差があるか調べてみた。その結果，すべての変数において，p=.000 レベルで有意差が確認された。
10. 上，中，下というラベルは便宜上の名称であり，上＝上級，中＝中級，下＝初級の意味ではない。既述の表4の「SPOT90-1」の得点が示すように，ほとんどの調査対象者はいわゆる初級以上であることが多く，上，中，下は，母集団の中での便宜上のラベルであることに注意してほしい。

参考文献

今井新悟・赤木彌生・中園博美 (2012).『J-CAT オフィシャルガイド: コンピュータによる自動採点日本語テスト』ココ出版
岩崎典子 (2002).「日本語能力簡易試験（SPOT）の得点と ACTFL 口頭能力測定（OPI）のレベルの関係について」『日本語教育』114, 100-105.
小林典子・フォード順子・山元啓史 (1996).「日本語能力の新しい測定法『SPOT』」『世界の日本語教育』6, 201-236
小林典子 (2015).「SPOT」『日本語教育のための言語テストガイドブック』(pp.110-126).くろしお出版.
李在鎬・井佐原均 (2006).「第二言語獲得における助詞「に」の習得過程の定量的分析」『計量国語学』25(4), 163-180（http://jhlee.sakura.ne.jp/geo-backup/paper/keiryo2006.pdf 2015年4月1日閲覧）
李在鎬 (編著) (2015).『日本語教育のための言語テストガイドブック』くろしお出版.
Ford-Niwa, Junko & Kobayashi, Noriko (1999). SPOT: A Test Measuring "Control" Exercised by Learners of Japanese, In K. Kanno (Ed). *The Acquisition of Japanese as a Second Language* (pp.53-69). John Benjamins Publishing Co.