Journal of International and Advanced Japanese Studies 国際日本研究

Vol. 4, March 2012, pp. 1–16
Doctoral Program in International and Advanced Japanese Studies
Graduate School of Humanities and Social Sciences, University of Tsukuba

Article

言語テスト「SPOT」の難易度に影響を与える要因
―選択肢の効果について―

楊　元・酒井　たか子・小林　典子
筑波大学

本稿では、即時的処理を要求し運用力を反映すると言われるSPOTの解答方法が記入式から四肢選択に変化した場合に、言語知識の有無の測定を目的とした通常の言語テストと同じように、選択肢が難易度コントロールに有効かどうかを検証し、考察を行った。

SPOTにおける選択肢の影響について、以下のような結論が得られた。

⑴受験者の日本語能力に対して、SPOTの項目が易しい場合は、選択肢は正答率にほとんど影響しない。問題文・文法項目が易しかった項目を、選択肢の調整で項目難易度を上げることが難しいことが明らかになった。

⑵受験者の日本語能力に対して、SPOTの項目が難しい場合は、選択肢が正答率に影響している。問題文・文法項目が難しかった項目を、選択肢の調整で項目難易度を下げることが可能であることが明らかになった。

⑶項目識別力に問題がないWEB版SPOTに対して、正解以外の選択肢をランダムにした場合、テスト全体の項目識別力の平均に影響を与えていないことがわかった。このことは、WEB版SPOTの問題を作成する際に、選択肢をランダムで作成可能であることを示唆している。

キーワード：SPOT、即時的処理、難易度、項目識別力、選択肢

1．はじめに

SPOT（Simple Performance-Oriented Test）は自然な発話速度で次々と読み上げられる文を聞きながら、解答用紙の同じ文中の空欄にひらがな１文字を書き取るというテストである。基本的には次の例１、例２のように１問は１文で、文法項目部分１文字の１箇所の空欄となっている。

例１　そこ（　）何をしているんですか。
例２　明日はちょっと大事（　）用があって行けないんです。

小林（2005）は、自然な速度で即時解答という即時的処理能力を要求するテストで、そのため手続き的知識[1]の自動化[2]を間接的に推計していると述べている。短時間で実施でき（60問で、10分程度）、採点も簡便で、信頼性が高いコストパフォーマンスの高いテストと言われている（小林他1992，1996，フォード丹羽他1995）。

現在では、SPOTは信頼性、妥当性のある統合的な日本語能力が推定できる間接テストとし

1　長期記憶に貯蔵されている知識には２種類あると見る。一つは、宣言的知識（declarative knowledge）で、内容を記述できるような“knowing what”の知識である。もう一つは、手続き的知識（procedural knowing）で、無意識に何かができるスキル的な“knowing how”の知識である（小柳2004：68）。

て広く認知され、多くの教育機関でプレイスメント・テストの一つとして、また、日本語教育研究においては簡易な日本語レベル判定の手段として、広く使用されるに至っている。しかし、SPOTにおいて、どのように難易度をコントロールすればいいのか、それぞれの能力レベルに一番適切な問題はどのようなものかなど、科学的な検証が十分行われていないことが問題である。

SPOTは音声テープと用紙と鉛筆を使用するテスト方法(以下、用紙版SPOTと呼ぶ)から、小林(2005)にあるようにインターネット利用のテストへと開発が進められている。用紙版SPOTの解答方法が、ひらがな1文字の記入式であるのに対して、WEB版SPOTでは、ひらがな4文字の中から選ばせる多肢選択方式となっている。先行研究ではWEB版も用紙版と同じように測定できているとしているが、コンピュータ利用による解答方法の違いが難易度に影響を与えるのか否か、与えるとすれば、どのように影響を及ぼすのかについての検証はまだ十分になされていない。

2．研究目的

現在、WEB版SPOTの実施により、様々な変化が起こっている。空欄にひらがな1文字を入れるテスト形式から、四肢選択形式に変わったが、そのため、用紙版SPOTと異なり、WEB版SPOTでは、正解を含む音声を聞きながら、目で4つの選択肢を確認し、正解を選ぶというように、受験者の解答行動が変化した。用紙版SPOTはいわゆる再生問題であり、WEB版SPOTは再認問題である。Heim & Watts(1986)は事前に示された可能性のある答えの中から、正しいか最善の答えを選択させるというような課題を受験者に与えることは不自然だとし、また、概して再認問題は、同じ内容の再生問題よりも幾分簡単であると述べている。しかし、近年ではRodriguez(2003)が「再認問題と再生問題の困難度の差異は尺度化の問題であり、もし分割点(cut score)がテスト得点に作られるならば、標準を設定する過程で補うことができるだろう。幹が同等の(内容的に等しい)再生問題と再認問題のメタ分析は、2つの形式の間に相関がほとんど1に近いことを示している(筆者訳)」とし、この2つの解答形式の論争に休止符を打っている。小林・酒井・フォード(2007)は、同じSPOT問題の用紙版とWEB版を受験者にそれぞれ受験させ、用紙版とWEB版の相関が、SPOT-FとSPOT-G(易しめのSPOT)の場合0.87、SPOT-DとSPOT-E(F、Gより難しめのSPOT)の場合0.94であったと述べている。この調査結果はRodriguez(2003)の指摘と一致している。また、偶然の当て推量が選択式問題の主要な弱点だと言われていることに対しては、「受験者たちは誤答を取り除くために部分的な知識を使うであろう(Ebel& Frisbie,1991)(筆者訳)」という指摘もある。以上の先行研究の知見から見れば、四肢選択形式になったWEB版SPOTの問題形式は妥当性に問題がないと言えるだろう。

WEB版SPOTは解答形式が四肢選択で、この点が用紙版ともっとも異なるが、フォード(2007)及び小林・酒井・フォード(2007)は、この違いがテスト結果にどのように影響するかを検討している。フォード(2007)は同じ受験者に対して用紙版とWEB版を実施し、次のように述べている。

「易しいSPOTでは用紙版得点が有意に高く、難しいSPOTではWEB版の得点が僅かだが高かった。(中略)用紙版の正解がWEB版では不正解というのはSPOTの難易にかかわらず同様に見られたが、用紙版の不正解がWEB版では正解というのは、難しいSPOTでは易しい

2 学習者が迅速にかつ容易に、そのタスクを行うことができるようになること。Automaticityの基準は研究者によって異なるが、DeKeyser(2001)によると、過去の研究者のautomaticityの基準は、主に「速さ」「同時並行的に行うことが可能」「努力の不要」「認知資源の容量を無制限にすることが可能」「非意図的」「一定の練習の結果」「干渉がほとんどない」「無意識」「いつでも記憶からの検索が可能」の中から成り立っているという。さらにEllis(2003)は、自動化(automatization) は単なる言語処理の強化や速度の向上という観点だけではなく、再構築(restructuring)も関わっているし、知識の新たな形式への再組織化も含まれていると主張している。

SPOTの２～４倍であった。WEB版の解答形式が、用紙版に比べより多く複雑な作業を要求することが、WEB版の得点を低くする一方で、問題が難しい場合には選択肢の存在が正解を助けると言える。」

上記の先行研究は、同じ受験者に実施したWEB版と用紙版SPOTの結果に基づくものだが、受験者人数が24名と少なく、四肢選択が、どのようにWEB版SPOTの難易度に影響を与えるかについての検証はまだ不十分だと思われる。

これまでのWEB版SPOTの選択肢は、用紙版で多く見られた誤答を基に作られているというが主として文法形式を重視しているように見える。しかし、WEB版では正解を含む選択肢を目で確認できることを考慮すると、学習者にとって混乱しやすい類似の音声の選択肢の設定も必要になるのではないかと考えられる。マクナマラ(2004)は、テストの方法がどの程度適切か或いは不適切か(すなわち妥当性)を問う場合には、テスト方法が得点に及ぼす影響について検討する必要があると述べている。そこで、本稿では、選択肢の設定以外の要素を変えず、元の選択肢から新しく作成した選択肢へ変更した場合、項目の難易度・識別力などにどのように影響を及ぼすかを明らかにする。

WEB版SPOTは１文の問題文にあった空欄箇所に、ひらがな１文字を選択するというテストであるため、通常の文法テストのように使い方の類似する文法知識を選択肢にするのは難しい。例えば、「わけ」の使い方を問う項目には、「はず」などの使い方の類似する選択肢を入れると、項目の難易度が上がることが考えられる。しかし、「＿け」を使って意味のある日本語にするためには「だけ」しか有効な錯乱肢が考えられない。そこで、新しい選択肢としては、文法形式や音声の類似性を考慮した選択肢を作成するとともに、ひらがな１文字をランダムに選んだ選択肢についても、検討することにした。もし、選択肢がSPOTの項目難易度・識別力に影響しないのであれば、正解以外の選択肢をランダム[3]で作成できるのではないかと思われる。このように選択肢をランダムで作成した場合、WEB版SPOTの項目難易度・識別力などにどのように影響を及ぼすかについても検証する。

３．研究方法

本研究は、以下の３つのステップを踏み、選択肢がどのように問題の難易度に影響を与えるかを明らかにするものである。

⑴　調査１

過去に実施したWEB版SPOT-D(30問)の項目分析を行い、テストの信頼性、項目難易度・識別力、実質選択肢数などを検討する。

⑵　調査２

仮説１「より錯乱効果のありそうな選択肢に変更することで、項目難易度・識別力・実質選択肢数は上がる」

この検証のために、WEB版SPOT-D(30問)における正答率が80％以上の14問に対して、ほとんど機能していない選択肢を取り除き、これらの選択肢を、文法形式と音声の類似性の二つの側面を考慮したより受験者を迷わす効果の高そうな選択肢へ変更し、項目難易度などに影響を及ぼすかを検討する。

⑶　調査３

仮説２「錯乱効果を考慮しない選択肢に変更することで、項目難易度・識別力・実質選択肢数は下がる」

この検証のために、WEB版SPOT-Dの正解以外の選択肢をランダムで作成したSPOT-Dランダム選択肢を検討する。

3　本稿における「ランダムに作成する」とは、錯乱効果を意図せず、無作為に選択肢を作成することである。

4．WEB版SPOT-Dの項目分析（調査１）

4.1　概要

SPOT-Dのテスト項目（30項目）は初級及び中級文法項目で構成されており、話しことばスタイルである。ここでは、2011年４月にテストを受けた222名の受験者のデータを利用し、項目分析を行う。受験者の日本語能力は、初級レベルから上級レベルまで幅広い。

4.2　実施方法

調査１は2011年４月、大学のコンピュータ室で一斉にテストを実施した[4]。受験者はヘッドフォンを付け、パソコン画面の問題文と選択肢を見て、マウスで正答をクリックするという形態を取った。WEB版SPOTは、１問ずつ、問題文が提示される。問題文と選択肢がパソコン画面に提示されると同時に、問題文の音声が流れ、音声が終わると4秒間のカウントダウンが始まる。音声が流れている間、及び音声の終わった後の４秒間は選んだ答えが選び直せるように設定されている。４秒間が過ぎると、問題文の画面が消え、次の問題に進む。以下の図4－1は実際のWEB版SPOT実施時のパソコンの画面例題の部分の表示である。

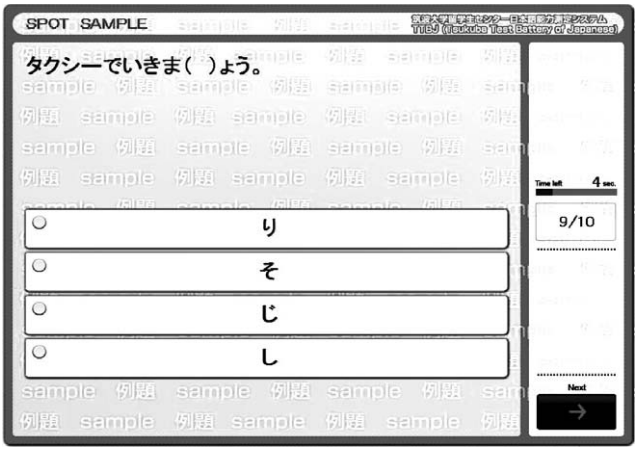


図4－1　WEB版SPOTの表示画面

4.3　調査１の結果

本稿の分析に使用する項目難易度・項目識別力・実質選択肢数はTest Data Analysis Program（TDAP）Ver. 2.0（『テストで言語能力は測れるか』（大友賢二監修／中村洋一著）添付プログラム）によって、出力されたものである。

SPOT-Dの各項目の問題文[5]、選択肢、項目難易度、項目識別力、実質選択肢数は表3－1で示している。SPOT-Dテストの得点分布は、平均値約76％、標準偏差20％であり、散らばりがやや大きいと言える。項目難易度と項目識別力を図4－2に散布図で表す。この図からも、難易度0.80以上の項目が13問あり、受験者にとって易しい問題となっている。項目識別力は30問のうちに、28問が0.40以上あり、その他の２問が0.30を超えている。項目数が30と少ないにもかかわらず、α係数は0.90あり、十分に高いと言えよう。

4　この実験では、筑波大学留学生センター日本語能力測定システム（Tsukuba Test Battery of Japanese）（略称TTBJ）を使用した。

5　SPOT-Dの問題文は「例１そこ（　）何をしているんですか」のようになっている。SPOT-Dのテスト問題は非公開のため、表4－1では、原問題文の空欄前後の一部分だけ示す。

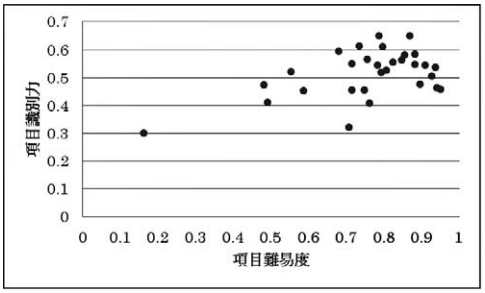


図4－2　SPOT-Dの項目難易度と項目識別力

表4－1　SPOT-Dの詳細データ（2011年４月実施）

項目	問題文	選択肢				DIFF	DISC	AENO
		A	B	C	D			
1	や(　)と言われても	れ(88)	と(4)	り(3)	る(3)	0.88	0.55	1.53
2	見つかる(　)いいですね	と(79)	て(5)	で(7)	が(5)	0.79	0.65	1.96
3	何ともい(　)ない	え(85)	い(9)	か(2)	な(2)	0.85	0.56	1.67
4	寝ちゃ(　)んです	う(73)	な(4)	る(1)	ん(12)	0.73	0.61	1.93
5	しな(　)ちゃいけません	く(78)	い(4)	け(10)	し(5)	0.78	0.55	2.00
6	聞か(　)ていただけませんか	せ(82)	い(4)	さ(7)	し(5)	0.82	0.56	1.84
7	(　)かがう	う(68)	い(3)	お(10)	か(17)	0.68	0.60	2.44
8	てもら(　)ませんか	え(55)	い(36)	う(2)	こ(4)	0.55	0.52	2.46
9	飲(　)すぎ	み(95)	き(1)	と(1)	む(2)	0.95	0.46	1.22
10	どちら(　)と言えば	か(48)	の(5)	へ(2)	が(44)	0.48	0.48	2.55
11	(　)うに言われた	よ(87)	い(2)	に(8)	ん(2)	0.87	0.65	1.61
12	帰(　)そうです	れ(16)	り(4)	る(64)	ろ(10)	0.16	0.30	2.55
13	多く(　)は	て(75)	た(5)	で(15)	と(5)	0.75	0.46	2.20
14	(　)ことで	の(94)	を(2)	が(0)	で(3)	0.94	0.46	1.29
15	と(　)えば	い(88)	お(4)	て(2)	は(3)	0.88	0.59	1.51
16	せ(　)に	ず(86)	つ(4)	て(2)	ん(5)	0.86	0.58	1.63
17	起き(　)うにも起きられない	よ(76)	も(10)	ら(5)	る(4)	0.76	0.57	2.06
18	やめた(　)んて	な(76)	い(6)	し(5)	の(6)	0.76	0.41	2.06
19	旅行に(　)も行きたい	で(59)	い(13)	て(15)	も(5)	0.59	0.45	2.75
20	行かない(　)り	よ(79)	い(4)	た(2)	ゆ(12)	0.79	0.52	1.90
21	わけには(　)かない	い(93)	あ(4)	か(1)	け(1)	0.93	0.51	1.34
22	(　)いでに	つ(70)	い(1)	す(5)	づ(19)	0.71	0.32	2.17
23	忘れた(　)	い(72)	の(8)	は(6)	ん(8)	0.72	0.55	2.24
24	ない(　)ちに	う(80)	い(4)	こ(7)	ち(5)	0.80	0.61	1.88
25	(　)うにする	よ(94)	こ(2)	す(2)	そ(1)	0.94	0.54	1.30
26	早く帰ろうと思っ(　)って	た(72)	だ(10)	て(10)	と(7)	0.72	0.46	2.38
27	そうに(　)って	な(90)	た(4)	だ(2)	か(2)	0.90	0.48	1.46
28	車を買うな(　)	ら(80)	い(5)	な(5)	を(8)	0.81	0.53	1.96
29	勉強する(　)ころか、	ど(49)	か(1)	と(4)	の(44)	0.49	0.41	2.38
30	べ(　)だ	き(91)	て(1)	で(1)	る(3)	0.91	0.55	1.28

注：１）DIFF = Item difficulty index　項目難易度：通過率、正答率

２）DISC = Discrimination power index
項目弁別力指数（点双列相関係数による弁別力指数）

３）AENO = Actual equivalent number of option
実質選択肢数：「実質的に機能している」選択肢数

４）表を見やすくするため、選択肢Aをすべて正解にした。実際のテストの時には、各選択肢の表示順位がランダムで出るように設定している。

５）各選択肢の欄の「(数字)」は、スペースの関係で、「％」を省略したものである。たとえば、「(5)」というのは「５％」を意味する。無回答の割合を省略しているため４つの選択肢の合計が100％にならないものもある。

4.4　調査１に関する考察

小林(2005)は、コンピュータにて実施する場合は、記入式の用紙版より、選択肢から選ぶほうが易しくなっていると述べている。今回の調査の結果から、WEB版SPOT-Dの正答率平均値は高く、プレイスメント・テストとしてはやや易しいことが分かった。そのため、WEB版SPOTの難度を高めるに、現有の問題の錯乱肢を更に吟味し、殆ど選ばれていない選択肢を取り除き、十分吟味した上に、能力の低い受験者を迷わせるような選択肢を作成して実質選択肢数を上げることが必要であると考えた。

５．SPOT-D選択肢変更の調査(調査２)

5.1　テストセット

今回の実験計画の資料としたものは、2009年度に実施したものである。そのテストで、実質選択肢数(AENO)が2.00以下のものが全部で14項目あり、その14項目を選択肢変更の対象とした。実質選択肢数とは、提供された選択肢が実質的にいくつの選択肢として作用していたかを示す数値であり、今回のテストのように選択肢が４つの場合、0.00から4.00までの数値を取る。2.00となった場合には、選択肢が４つであるにも関わらず、実質的には２つしかないのと同じ意味だということになる。

選択肢の変更を行う際には、項目分析の結果に基づいて、よい選択肢を残し、ほとんど機能していない選択肢に対して、選択選択肢の変更を行った。新しい選択肢を作成する際には、文法面を考慮した選択肢と音声類似の選択肢の両側面から考えた。また、選択肢によっては、文法的な要素と音声的な要素を両方備えている。表4－1には、「____は文法的な要素」、「~~~~は音声的な要素」、「_____は文法と音声両方考えたもの」を意味する。

以下の表5－1は選択肢変更項目のまとめである。SPOT-D選択肢変更バージョンは、新しい選択肢の14項目と共通の16項目によって構成されている。以下、SPOT-D元バージョンをフォーム１、SPOT-D選択肢変更バージョンをフォーム２とする。

表5－1　選択肢変更項目の詳細

項目	問題文	選択肢				DIFF	DISC	AENO
		A	B	C	D			
1	や(　)と言われても	れ	**る→て**	**り→ね**	と	0.91	0.43	1.35
2	見つかる(　)いいですね	と	**が→ど**	で	て	0.84	0.49	1.61
5	しな(　)ちゃいけません	く	**し→ぐ**	け	い	0.71	0.57	2.14
6	聞か(　)ていただけませんか	せ	**い→ぜ**	さ	し	0.79	0.66	1.91
9	飲(　)すぎ	み	**き→ひ**	**と→い**	む	0.92	0.50	1.33
11	(　)うに言われた	よ	**ん→そ**	**に→ゆ**	い	0.84	0.63	1.61
14	(　)ことで	の	**を→と**	**が→ど**	で	0.89	0.64	1.52
15	と(　)えば	い	**お→し**	**て→ち**	は	0.86	0.58	1.54
16	せ(　)に	ず	**て→す**	つ	ん	0.81	0.66	1.87
21	わけには(　)かない	い	**あ→に**	**か→し**	**け→ひ**	0.87	0.53	1.47
25	(　)うにする	よ	**こ→ふ**	**す→ゆ**	そ	0.93	0.47	1.26
27	そうに(　)って	な	**か→ら**	**だ→あ**	た	0.81	0.61	1.78
28	車を買うな(　)	ら	**い→か**	**を→と**	な	0.89	0.40	1.54
30	べ(　)だ	き	**て→し**	**で→い**	**る→ぎ**	0.93	0.45	1.23

注：表を見やすくするため、選択肢Ａをすべて正解にし、選択肢を変更した項目を左寄りで並ぶようにした。実際のテストの時には、各選択肢の表示順位がランダムで出るように設定している。

5.2　調査２の概要

調査２は2011年４月及び９月に、大学のコンピュータ室で一斉にテストを実施した。受験者は、筑波大学で学ぶ留学生362人である。フォーム１を受験した者が222人、フォーム２を受験した者が140人であった。受験者の日本語能力は、初級レベルから上級レベルまで幅広い。実施方法は調査１と同様である。

5.3　調査結果の概要

表5－2にSPOT-DとSPOT-D選択肢変更における基本統計値と信頼性係数としてクロンバックの α 係数を示す。表5－3は、各フォームの共通項目、それぞれの正答率の平均値と標準偏差を示したものである。

表5－2　フォーム１とフォーム２の基本統計量

	フォーム１	フォーム２
受験者数	222	140
項目数	30	30
最低点	2	4
最高点	30	30
平均点	22.79	22.31
標準偏差	5.93	6.46
α 係数	0.90	0.91

表5－3　各フォーム問題形式別正答率結果

フォーム	問題の種類	項目数	受験者数	平均値	標準偏差
1	元の選択肢	14	222	0.88	0.19
1	共通項目	16	222	0.66	0.22
2	選択肢変更	14	140	0.86	0.21
2	共通項目	16	140	0.65	0.25

表5－4　各フォーム共通項目基本統計量

フォーム	項目数	受験者数	平均値	標準偏差
1	16	222	10.53	3.62
2	16	140	10.32	4.01

次に、フォーム１とフォーム２、それぞれのテストの受験者の能力に差があるかどうかを確認した。表5－4は各フォームにおける共通項目合計点の平均値と標準偏差を示したものである。フォーム1とフォーム２の結果の等分散性の検定を行ったところ、等分散性が確認できた（$F(221,139)=1.744$, $p>.10$）。そこで、共通項目におけるフォーム１受験者とフォーム２受験者の平均の差をt検定にかけた結果、両者の平均値には有意な差がなかった（$t(360)=0.877$, ns）。つまり、フォーム１受験者とフォーム２受験者の間に日本語能力の差がないと言える。

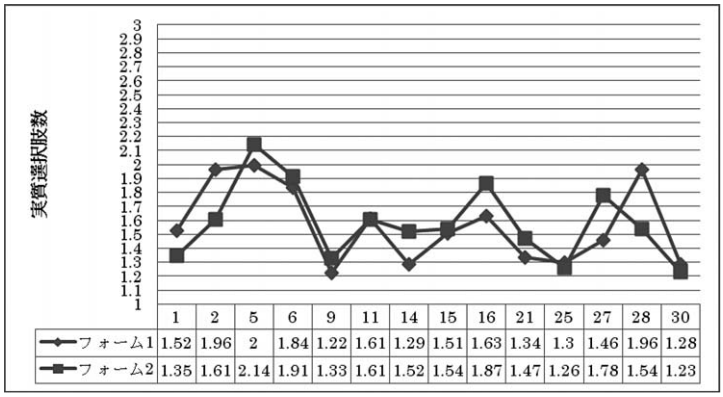
5.4　選択肢変更による違い―実質選択肢数について

5.4.1　結果の概要

問題項目の錯乱肢が効果的に機能していたかを実質選択肢数によって分析を行う。実質選択肢数の数値は、0.00から選択肢の数までの数値となる。多肢選択形式の問題においてどのような選択肢を作成するかは、テスト作成の際に非常に重要な問題となる。特に錯乱肢は、中村（2004）が述べているように「能力の低い受験者に正答だと思わせる魅力があり」、かつ、「能

力の高い受験者には誤答であると分かるものでなくてはならない」。受験者の誰もが選ばないような選択肢がテスト項目の中に含まれている場合には、改善の必要がある。

以下の図5－1は、フォーム１とフォーム２選択肢変更部分の実質選択肢数の分布を示したものである。



	1	2	5	6	9	11	14	15	16	21	25	27	28	30
フォーム1	1.52	1.96	2	1.84	1.22	1.61	1.29	1.51	1.63	1.34	1.3	1.46	1.96	1.28
フォーム2	1.35	1.61	2.14	1.91	1.33	1.61	1.52	1.54	1.87	1.47	1.26	1.78	1.54	1.23

図5－1　フォーム１とフォーム２選択肢変更部分の実質選択肢数の分布

表5－5　選択肢変更部分の実質選択肢数

	元の選択肢	選択肢変更
項目	14	14
実質選択肢数平均値	1.57	1.58
実質選択肢数標準偏差	0.28	0.27
t 値（有意確率）	2.160（0.769）	

同じ原問題項目で、選択肢を変更した場合と元の選択肢の場合、両条件の実質選択肢数の平均の差が有意なものであるかを検討した。対応のあるｔ検定の結果、両条件の実質選択肢数の平均の差に有意な差は認められなかった（$t(13)=2.160, p>.10$）。すなわち、全体的に見ると、フォーム２の選択肢変更は、実質選択肢数の向上に効果がないことが分かった。

5.4.2　項目ごとの分析

14項目全体では実質選択肢数は変わらなかったが、項目別で見ると、実質選択肢数の上がったのは８項目あり（項目27は0.32、２項目が0.2以上、その他の５項目は約0.1）、変わらなかったのは1項目で、逆に下ったのは５項目ある。

以下では、実質選択肢数が0.2以上上がった３項目及び逆に0.2以上下がった項目を細かく分析する。

5.4.2.1　実質選択肢数が0.2以上上がった項目

項目14　問題文（一部）　入学試験（　）ことで　（実質選択肢数1.29 → 1.54）
　　選択肢　の　を→と　が→ど　で　（正答率　0.94 → 0.89）

項目16　問題文（一部）　朝から食事もせ（　）に　（実質選択肢数1.63 → 1.87）
　　選択肢　ず　て→す　つ　ん　（正答率　0.86 → 0.81）

項目27　問題文(一部)　階段から落ちそうに(　)って　(実質選択肢数1.46 → 1.78)
　　選択肢　な　か→ら　だ→あ　た　(正答率　0.90 → 0.81)

項目14は正解ではない「と」と「ど」のペアで能力の低い受験者を迷わせた可能性があるため、実質選択肢数が上がった。正解「の」以外の選択肢に、「清音」と「濁音」のペアで選択肢にあった場合は、受験者の受験ストラテジーから見れば、「正解はこの二つの選択肢にある」と理解されることがある。そのため、テスト結果は日本語能力以外の受験ストラテジーの働きを受けやすいことになる。このような選択肢の設定方法が妥当なのかをさらに検討する必要がある。

項目16は錯乱肢「て」を正解である「ず」の清音である「す」に変更し、実質選択肢数が大きく上がった。「～せずに」は旧日本語能力試験2級文法であり、日本語能力の低い受験者にとってはまだ習っていない文法項目であるため、音声のみに頼って「ず」「す」の間で迷っていることがうかがえる。

項目27は、この項目の音声を聞くと、「落ちそうになって」の「な」の子音「ｎ」の発音が弱く、「ら」或いは「あ」に聞く可能性があるため、選択肢をそのように変更した。これらの選択肢の変更は、括弧のところの音声のみに頼って、答えを選ぶ学習者には有効な変更だと考えられる。

5.4.2.2 実質選択肢数が0.2以上下がった項目

項目２　問題文(一部)　見つかる(　)いいですね　(実質選択肢数1.96 → 1.61)
　　選択肢　と　が→ど　て　で　(正答率　0.79 → 0.84)

項目28　問題文(一部)車を買うな(　)よく車のこと　(実質選択肢数1.96 → 1.54)
　　選択肢　ら　い→か　を→と　な　(正答率　0.81 → 0.89)

項目２の正解である「と」の濁音「ど」を選択肢に変更してみたが、錯乱肢としてうまく働いた結果にはなっていない。元の選択肢にあった「が」の方が、「～がいいですね」のという形は日本語能力の低い受験者にとってなじみがあり、よりよい錯乱肢になっている。項目２と項目14を例に、「濁音」と「清音」のような音声が類似するペアで選択肢を作っても、必ずしもよい錯乱肢になるとは限らないことがわかった。

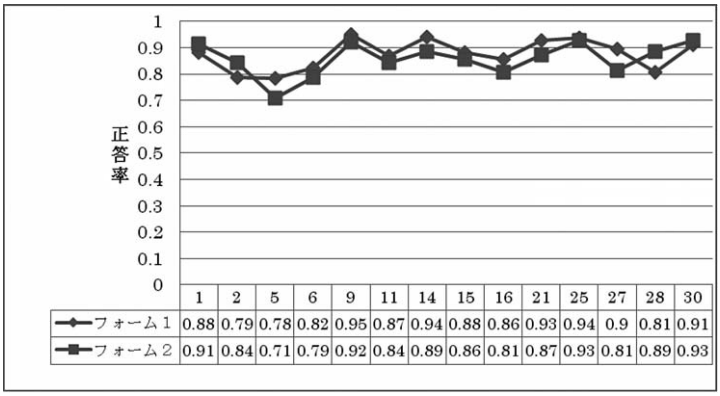
項目28は、元の選択肢にあった「い」のほうが、「ない」という初級で習う文法項目になるため、知らないものを選ぶより、見たことがある文法項目を大いに選ぶ傾向があることがうかがえる。したがって、選択肢を作成する際には、空欄箇所の前後の文字列で、最も使用頻度の高い文法項目で選択肢を作成するのがよいと考えられる。

5.5　選択肢変更による違い―正答率

フォーム１及びフォーム２において、選択肢変更の項目、及び元の選択肢の項目、それぞれの正答率分布を図5－2で示した。同じ問題文の項目で、選択肢を変更した場合と元の選択肢の場合、両条件の正答率の平均の差が有意なものであるかを検討した。対応のあるｔ検定の結果、両条件の正答率の平均の差に有意な差は認められなかった($t(13)=2.160, p>.10$)。

表5－6　選択肢変更部分の正答率

	元の選択肢	選択肢変更
項目	14	14
正答率平均値	0.88	0.86
正答率標準偏差	0.06	0.06
t値(有意確率)	2.160(0.174)	

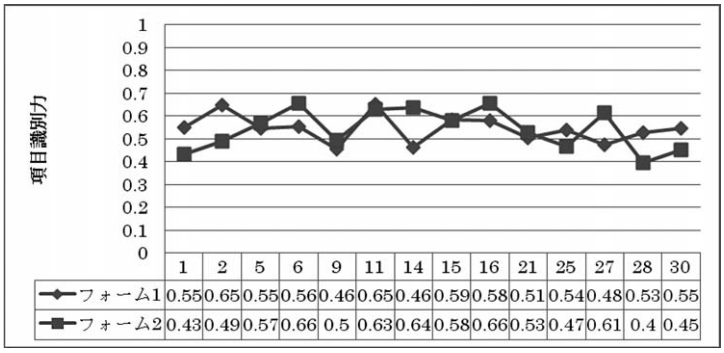


	1	2	5	6	9	11	14	15	16	21	25	27	28	30
フォーム1	0.88	0.79	0.78	0.82	0.95	0.87	0.94	0.88	0.86	0.93	0.94	0.9	0.81	0.91
フォーム2	0.91	0.84	0.71	0.79	0.92	0.84	0.89	0.86	0.81	0.87	0.93	0.81	0.89	0.93

図5－2　フォーム１とフォーム２選択肢変更部分の正答率の分布

5.6　選択肢変更による違い―識別力

フォーム1及びフォーム２において、選択肢変更の項目、元の選択肢の項目、それぞれの識別力を図5－3で示した。ここで示した識別力は、各フォーム内の共通16項目と選択肢変更14項目を合わせた合計点との点双列相関のことである。同じ原問題項目で、選択肢を変更した場合と元の選択肢の場合、両条件の項目識別力の平均の差が有意なものであるかを検討した。対応のある *t* 検定の結果、両条件の識別力の平均の差に有意な差は認められなかった（$t(13)=2.160$, $p>.10$）。



	1	2	5	6	9	11	14	15	16	21	25	27	28	30
フォーム1	0.55	0.65	0.55	0.56	0.46	0.65	0.46	0.59	0.58	0.51	0.54	0.48	0.53	0.55
フォーム2	0.43	0.49	0.57	0.66	0.5	0.63	0.64	0.58	0.66	0.53	0.47	0.61	0.4	0.45

図5－3　フォーム１とフォーム２選択肢変更部分の項目識別力の分布

表5－7　選択肢変更部分の識別力

	元の選択肢	選択肢変更
項目	14	14
識別力平均値	0.55	0.54
識別力標準偏差	0.06	0.09
t 値（有意確率）	2.160（0.935）	

5.7　調査２に関する考察

調査２では、フォーム２の選択肢変更の14項目において、ほとんど機能していない選択肢を、文法性と音声の類似性を考慮した新たな選択肢にし、選択肢変更後の実質選択肢数が上がるのではないかと仮説をたてた。しかし、実質選択肢数の結果を見ると、実質選択肢数の上が

ったのは８項目あり、変わらなかったのは１項目で、逆に下ったのは５項目ある。この結果は仮説に反し、今回の選択肢の修正は有効な項目と有効ではない項目が混合していることが明らかになった。選択肢変更を対象にした14項目はいずれも正答率が高く(85％位)、天井効果になっていることが選択肢変更に効果が現れなかった原因の一つだと考えられる。つまり、受験者レベルに対して、問題文及び測ろうとしている文法項目そのものが易しかったため、問題の難易度に選択肢の関与する影響が少なかったと考えられる。

上記で分析したように、今回の選択肢の変更自身が全て有効なものではなく、有効でない選択肢の変更も混在している。また、易しい問題のみを対象にし、選択肢変更を行ったことの妥当性も考える必要がある。したがって、仮説１「より錯乱効果のありそうな選択肢に変更することで、項目難易度・識別力・実質選択肢数は上がる」については更に検討する必要がある。

そこで、次節の調査３では、SPOT-Dの30項目の全問題に対して、正解以外の選択肢をランダムで作成し、項目難易度・実質選択肢数などにどのように影響を及ぼすかを検討する。

６．SPOT-Dランダム選択肢の調査(調査3)

6.1　テストセット

SPOT-D(30項目)各項目の正解以外の選択肢を、ランダムで作成した。選択肢の候補は、50音図の中から「ゐ」と「ゑ」を除いて、日常使われる46文字と濁音などを加えた70項目とした。この70項目を、ウェブサイトA.k Officeの「抽選王」[6]というソフトを利用し、選択肢を無作為に決定した。

6.2　調査３の概要

調査３のために、2011年12月に留学生143人を対象に、SPOT-Dフォーム３(ランダム選択肢)を実施した。受験者の日本語能力は、初級レベルから上級レベルまで幅広い。実施方法は調査１及び調査２と同様である。

6.3　受験者の能力

2011年４月及び12月の調査では、SPOT以外に、旧日本語能力試験２級、３級に相当する通常の文法テスト36項目も受験者に受験させた。表6－1は、各フォームの共通36問の文法テストの正答率の平均値と標準偏差を示したものである。

表6－1　共通文法テストの正答率の平均値と標準偏差

フォーム	項目数	受験者数	平均値	標準偏差
1	36	222	24.03	8.74
3	36	143	23.73	8.95

次に、フォーム１とフォーム３、それぞれのテストの受験者の能力に差があるかどうかを確認した。表5－1は各フォームにおける共通項目合計点の平均値と標準偏差を示したものである。フォーム１とフォーム３の結果の等分散性の検定を行ったところ、等分散性が確認できた($F(221,142)=.246$, $p>.10$)。そこで、共通文法項目におけるフォーム１受験者とフォーム３受験者の平均の差をt検定にかけた結果、両者の平均値には有意な差がなかった。($t(363)=.314$, *ns*)。つまり、フォーム１受験者とフォーム３受験者の間に日本語能力の差はないと言える。

6.4　調査結果の概要

表6－2はオリジナルSPOT-D(フォーム１)とSPOT-Dランダム選択肢(フォーム３)における

6　A.k Officeの「抽選王」はhttp://www.ak-office.jp/software/tyusenk.htmlで参照可能。

基本統計値と信頼性係数クロンバックのα係数を示したものである。表6－3は、フォーム１及びフォーム３の各項目の項目難易度、項目識別力及び実質選択肢数をまとめたものである。

表6－2　フォーム１及びフォーム３の基本統計量

	フォーム１	フォーム3
受験者数	222	143
項目数	30	30
最低点	2	8
最高点	30	30
平均点	22.79	23.83
標準偏差	5.93	4.88
α係数	0.90	0.89

表6－3　フォーム１及びフォーム３の各項目の正答率、項目識別力及び実質選択肢数

項目	問題文	選択肢				フォーム1	フォーム3	フォーム1	フォーム3	フォーム1	フォーム3
		A	B	C	D	DIFF	DIFF	DISC	DISC	AENO	AENO
12	帰（　）そうです	れ	こ	が	び	0.16	0.67	0.30	0.41	2.55	1.71
10	どちら（　）と言えば	か	な	ね	し	0.48	0.91	0.48	0.50	2.55	1.42
29	勉強する（　）ころか	ど	ぜ	ぎ	む	0.49	0.92	0.41	0.35	2.38	1.28
8	てもら（　）ませんか	え	れ	ら	わ	0.55	0.85	0.52	0.77	2.46	1.56
19	旅行に（　）も行きたい	で	げ	だ	な	0.59	0.77	0.45	0.33	2.75	1.88
7	（　）かがう	う	た	し	ね	0.68	0.75	0.60	0.62	2.44	1.96
22	（　）いでに	つ	ど	ゆ	ぬ	0.71	0.89	0.32	0.66	2.17	1.50
23	忘れた（　）	い	ぷ	ぐ	ほ	0.72	0.78	0.55	0.57	2.24	1.85
26	早く帰ろうと思っ（　）って	た	は	そ	こ	0.72	0.88	0.46	0.36	2.38	1.47
4	寝ちゃ（　）んです	う	ぶ	み	せ	0.73	0.87	0.61	0.44	1.93	1.35
13	多く（　）は	て	わ	め	り	0.75	0.93	0.46	0.69	2.20	1.31
17	起き（　）うにも	よ	た	ぱ	お	0.76	0.66	0.57	0.26	2.06	2.18
18	やめた（　）んて	な	ひ	で	め	0.76	0.78	0.41	0.34	2.06	1.73
5	しな（　）ちゃいけません	く	そ	み	う	0.78	0.87	0.55	0.61	2.00	1.49
2	見つかる（　）いいですね	と	ぜ	ほ	そ	0.79	0.88	0.65	0.47	1.96	1.47
20	行かない（　）り	よ	う	ぶ	ど	0.79	0.90	0.52	0.48	1.90	1.47
24	ない（　）ちに	う	ば	き	れ	0.80	0.81	0.61	0.49	1.88	1.81
28	車を買うな（　）	ら	へ	げ	で	0.81	0.94	0.53	0.37	1.96	1.22
6	聞か（　）ていただけませんか	せ	ど	よ	ほ	0.82	0.94	0.56	0.46	1.84	1.20
3	何ともい（　）ない	え	ご	が	ね	0.85	0.84	0.56	0.61	1.67	1.63
16	せ（　）に	ず	ゆ	む	か	0.86	0.90	0.58	0.73	1.63	1.44
11	（　）うに言われた	よ	げ	み	じ	0.87	0.92	0.65	0.58	1.61	1.36
1	や（　）と言われても	れ	そ	は	つ	0.88	0.94	0.55	0.62	1.53	1.25
15	と（　）えば	い	が	く	り	0.88	0.90	0.59	0.56	1.51	1.50
27	そうに（　）って	な	ぱ	り	こ	0.90	0.85	0.48	0.31	1.46	1.59
30	べ（　）だ	き	ご	に	な	0.91	0.90	0.55	0.46	1.28	1.41
21	わけには（　）かない	い	び	た	め	0.93	0.90	0.51	0.62	1.34	1.37
25	（　）うにする	よ	で	ご	お	0.94	0.94	0.54	0.71	1.30	1.26
14	（　）ことで	の	い	へ	ぷ	0.94	0.97	0.46	0.47	1.29	1.17
9	飲（　）すぎ	み	は	で	む	0.95	0.93	0.46	0.59	1.22	1.36

注：１）表を見やすくするため、選択肢Ａをすべて正解にした。実際のテストの時には、各選択肢の表示順位がランダムで出るように設定している。

２）正解以外の選択肢がランダムで選ばれたものである。

３）項目の順番は、フォーム１の正答率の昇順で並べたものである。

6.5　ランダム選択肢による違い ― 実質選択肢数について

フォーム１及びフォーム３において、ランダム選択肢の項目及び元の選択肢項目、それぞれの実質選択肢数の分布を表6－3で示した。同じ原問題項目で、選択肢をランダムで作成した場合と元の選択肢の場合、両条件の実質選択肢数の平均の差が有意なものであるかを検討した。対応のある *t* 検定の結果、両条件の実質選択肢数の平均に有意な差を認めた（$t(29)=5.726, p<.01$）。ランダム選択肢にした場合、実質選択肢数の平均値が0.41下がっている。

表6－4　フォーム１及びフォーム３の実質選択肢数の *t* 検定結果

	元の選択肢	ランダム選択肢
項目	30	30
実質選択肢数平均値	1.92	1.51
実質選択肢数標準偏差	0.44	0.24
t 値（有意確率）	5.726（0.000）	

表6－3によると、実質選択肢数が1.00以上下がったのは、項目10と項目29の２項目である。

項目10「どちら（　）と言えば」にとっては、フォーム１に元々あった「が」という選択肢のほうが効果的な錯乱肢だったと思われる。「が」という選択肢がランダム設定の結果なくなり、実質選択肢数が1.00以上下がったことが分かった。

項目29「勉強する（　）ころか」において、フォーム１に元々あった「と」と「の」が効果的な錯乱肢であり、この２つの選択肢がなくなり、ランダム設定で「ぜ」「ぎ」「む」となった結果、実質選択肢数が1.00以上下がったことが分かった。

6.6　ランダム選択肢による違い ― 正答率

同じ問題文の項目で、選択肢をランダムで作成した場合と元の選択肢の場合、両条件の正答率の平均の差が有意なものであるかを検討した（表6－3）。対応のある t 検定の結果、両条件の正答率の平均に有意な差が認められた（$t(29)=-4.069, p<.01$）。すなわち、フォーム３のランダム選択肢は、項目の正答率に影響を与えている。ランダム選択肢にした場合、正答率の平均値が0.11上がった。

表6－5　フォーム１及びフォーム３の正答率の *t* 検定結果

	元の選択肢	ランダム選択肢
項目	30	30
正答率平均値	0.76	0.87
正答率標準偏差	0.17	0.08
t 値（有意確率）	-4.069（0.000）	

表6－3によると、正答率が0.40以上上がったのは、項目12、項目10及び項目29の３項目である。項目10と項目29の正答率が大きく上がったことについては、前節6.5で述べたとおりである。項目12「帰（　）そうです」において、フォーム１に元々あった「り」という選択肢が一見正しいと見えるが、問題文の文脈と音声で判断すれば「れ」が正解である。効果的だと思われる選択肢「り」がなくなり、ランダム設定の選択肢になった結果、正答率が0.40以上上がったことが分かった。

また、正答率0.10以上下がったのは、項目17の1項目のみであった。項目17「起き（　）うにも」において、ランダム設定の選択肢に「た」という錯乱肢があった。「起きたうにも」という文字列が非言語的にもかかわらず、日本語能力の低い受験者は「起きた」だけを見て、「た」を選んだと考えられる。項目17の正答率が0.10下がったのはランダム設定の選択肢にたまたま

効果的な錯乱肢が混ざったためである。

6.7　ランダム選択肢による違い ― 識別力

フォーム１及びフォーム３において、選択肢変更の項目及び元の選択肢の項目、それぞれの識別力を表6－3で示した。ここで示した識別力は、各フォーム内の30項目の合計点との点双列相関のことである。同じ原問題項目で、選択肢をランダムで作成した場合と元の選択肢の場合、両条件の項目識別力の平均の差が有意なものであるかを検討した。対応のある *t* 検定の結果、両条件の識別力の平均に有意な差を認められなかった（$t(29)=.035$, $p>.10$）。すなわち、ランダムにした選択肢は項目識別力の平均値に影響を与えていないと言えるであろう。

表6－6　フォーム１及びフォーム３の項目識別力の*t*検定結果

	元の選択肢	ランダム選択肢
項目	30	30
識別力平均値	0.52	0.51
識別力標準偏差	0.08	0.14
t 値（有意確率）	0.035（0.972）	

6.8　調査３に関する考察

WEB版SPOT-Dの選択肢は用紙版SPOT-Dを実施した時に、受験者の間違いの多い解答を参考に作成したものである。調査３では、正解以外の選択肢を全てランダムで作成した。実施した結果、項目実質肢数が0.41下がり、正答率が0.11上がった。この結果から、選択肢はSPOTの難易度に影響していると言えるであろう。表6－3を見ると、元の選択肢で正答率が0.85以上の問題では、ランダムの選択肢で実施した結果は元の選択肢で実施した結果とほとんど変わっていないことが分かる。逆に、元の選択肢で正答率が0.60以下の問題では、ランダムの選択肢で実施した結果元の選択肢で実施した結果より、正答率が0.2以上上がっていることがわかる。つまり、受験者の日本語能力に対して、易しい問題であれば、どのような選択肢にした場合でも、正答率に影響を与えない。逆に、受験者の日本語能力に対して、難しい問題の場合に、選択肢が問題の正答率に影響していることが明らかになった。この結果は、フォード（2007）及び小林・酒井・フォード（2007）の結論とある程度一致している。調査３の結果から見ると、受験者が選択肢からヒントを得たり、排除法を取ったりして解答していることがうかがえる。受験者が即時解答を要求されるWEB版SPOTにおいても、通常の四肢選択形式テストと同じ解答行動を行っていることが明らかになった。

項目識別力に関して言えば、選択肢をランダムで作成した場合と元の選択肢の場合、両条件の項目識別力の平均の差は有意ではなかった。つまり、ランダムにした選択肢は項目識別力の平均値に影響を与えていないと言えるであろう。この結論は、以下の２点から説明できる。まず、フォーム１の各項目の識別力に関しては、30問のうち、28問が0.40以上あり、その他の２問が0.30を超えている。選択肢をランダムにした場合のフォーム３の項目識別力に関しては、30問のうち、１問が0.26、6問が0.30以上、その他の23問が0.40を超えている。つまり、SPOT-Dの問題自身に良好な項目識別力があり、選択肢を替えても、識別力に影響を与えないことを示唆している。次に、フォーム１及びフォーム３を受験した受験者の能力は、初級から上級まで様々で散らばりがある。そのため、項目識別力が高く出るのが通常である。日本語レベルのばらつきが小さいグループに、フォーム１とフォーム３を受験させた場合、項目識別力がどのように変わるかを今後の課題としたい。

7．まとめ

本稿では、即時的処理を要求し運用力を反映すると言われるSPOTの解答方法が記入式から四肢選択に変化した場合に、言語知識の有無の測定を目的とした通常の言語テストと同じように、選択肢が難易度コントロールに有効かどうか、仮説１と仮説２を検証することで、考察を行った。

仮説１「より錯乱効果のありそうな選択肢に変更することで、項目難易度・識別力・実質選択肢数は上がる」

仮説２「錯乱効果を考慮しない選択肢に変更することで、項目難易度・識別力・実質選択肢数は下がる」

二つの仮説の検証のために、調査１～調査３を行った。

上記の調査を通して、SPOT形式テストにおける選択肢の影響について、以下のような結論が得られた。

⑴　受験者の日本語能力に対して、SPOTの項目が易しい場合は、受験者が「文字と音声情報を同時に利用し、即時的に選択肢から答えを指定する」という解答行動を行っているのではないかと思われる。そのため、受験者が正答以外の解答にあまり目が向かず、選択肢は正答率にほとんど影響しない。問題文・文法項目が易しかった項目を、選択肢の調整で項目難易度を上げることが難しいことが明らかになった。

⑵　受験者の日本語能力に対して、SPOTの項目が難しい場合は、選択肢が正答率に影響している。問題文・文法項目が難しかった項目を、選択肢の調整で項目難易度を下げることが可能であることが明らかになった。

⑶　項目識別力に問題がないWEB版SPOTに対して、正解以外の選択肢をランダムにした場合、テスト全体の項目識別力の平均に影響を与えていないことがわかった。このことは、WEB版SPOTの問題を作成する際に、選択肢をランダムで作成可能であることを示唆している。

SPOTのテストの難易度に影響を与える要因は、対象の受験者にとっての文法項目の難しさ、問題文の語彙の難しさ、また音声の速度・ポーズなどの要素が複合的に関係している。今回の調査では、選択肢のみ注目したが、今後は選択肢と上記の要素が絡んだ場合の項目難易度に与える影響を考えたい。

参考文献

伊東祐郎(2008)『留学生の日本語能力測定のための項目プールの構築』平成16年度～平成19年度科学研究費補助金　基盤研究(A)研究成果報告書　東京外国語大学

今井新悟(2010)『J-CAT Japanese Computerized Adaptive Test－日本語能力をコンピュータで測る－』山口大学出版

小林典子・フォード丹羽順子(1992)「文法項目の音声聴取に関する実証的研究」『日本語教育』78号、日本語教育学会

小林典子・フォード丹羽順子・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号

小林典子・酒井たか子・フォード丹羽順子(2007)「即時要求型言語テストのWEB化-SPOT-WEBの場合-」CASTEL-J in Hawaii

小林典子(2005)「言語テストSPOTについて－用紙形式からWEB形式へ」『筑波大学留学生センター日本語教育紀要』20号

小柳かおる(2004)『日本語教師のための新しい言語習得概論』スリーエーネットワーク

中村洋一(2002)『テストで言語能力は測れるか』桐原書店

フォード丹羽順子(2007)「言語テストSPOTのWEB版・用用紙版の比較」『佐賀大学留学センター紀要』６号

フォード丹羽順子・小林典子・山元啓史(1995)「日本語能力簡易試験(SPOT)は何を測定しているか－音声テープ要因の解析－」『日本語教育』86号, 日本語教育学会
ブラウン J.D.(著)・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』大修館書店
マクナマラ, T.(2004)(伊東祐郎・三枝令子・島田めぐみ・野口裕之監訳)『言語テスティング概論』スリーエーネットワーク
宮内俊慈ほか(2006)「プレースメント用リスニング・テスト改善報告」『関西外国語大学留学生別科日本語教育論集』関西外国語大学留学生別科16号
DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson, (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge University Press.
Ebel, R. L,&Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ:Prentice Hall
Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.
Heim, A. W., & Watts, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of educational Psychology*,37,339–346
Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement,*40(2), 163–184

筑波大学大学院人文社会科学研究科国際日本研究専攻
楊　元　(博士課程)
酒井　たか子(教授)
小林　典子　(元教授)