非漢字系日本語学習者向け漢字語彙テストの開発と評価
－ウズベク人日本語学習者によるデータを基に－

大和　祐子　　ベリディクロヴァ・ニギナ

1. はじめに

本稿では，研究利用を目的とした非漢字系日本語学習者向けの多肢選択式の漢字語彙テストを開発し，項目分析を通してその有用性を評価した結果を報告する。

日本語のテキスト(文章)には和語・漢語問わず多くの漢字表記の語が含まれ，日本語のテキスト理解には漢字の知識が欠かせない。しかし，単に多くの漢字を知っていれば日本語のテキストが理解できるとは限らない。日本語の語には多くの二字漢字語があり，それらの「語」としての知識があってはじめて，日本語のテキスト理解が可能になるからである。

このように漢字語彙の理解は日本語を理解する上で重要な部分を占めるだけに，日本語学習者にとっても漢字語彙の習得は重要である。とりわけ，それは非漢字系日本語学習者にとっては日本語学習を成功させるうえで大きな課題となる(Gamage 2003)ことが知られており，非漢字系日本語学習者の漢字語彙習得やその学習ストラテジーについて数多く研究されている。これらの研究では調査協力者となる日本語学習者の現時点での漢字語彙知識を正確に測ることが必要になる。そこで，本稿では非漢字系日本語学習者の漢字語彙習得の研究で使用可能な漢字語彙テストを開発し，その有用性を検討することとした。

2. 非漢字系日本語学習者向け漢字語彙テストの概要

本稿で分析結果を報告する漢字語彙テストは，母語の書字形態に漢字を持たない非漢字系日本語学習者の漢字語彙知識を測る熟達度テスト(proficiency test)であり，このテストは非漢字系日本語学習者を対象とした漢字語彙習得研究における調査協力者のグループ分けなどに利用することを目的としたものである。本研究で開発したテスト同様，研究への利用を目的とした語彙テストには，中国語母語話者(漢字系日本語学習者)を対象とした宮岡・玉岡・酒井(2011)，非漢字系日本語学習者を対象とした大和・玉岡・茅本(2016)がある。両者は，漢字語彙に限らず日本語の語彙知識を幅広く問うているものである。それに対し，本テストは漢字語彙の中でも二字漢字語をターゲットとした，範囲を限定した語彙テストであるといえる。このように二字漢字語の語彙知識を問うことに特化したテストを開発したのには，2点の理由がある。まず1点目として，漢字語彙習得研究における調査協力者のグループ分けに，日本語の語彙知識を幅広く測定したテストでは研究目的と合致したグループ分けが難しいと考えるからである。例えば，日本語学習者の漢字(語彙)学習ストラテジーを調査するために上位群と下位群の学習者を分けるために使用するテストは，学習者の漢字(語彙)知識

を測るものであることが望ましい。しかしながら，このような学習者の能力を分けるようなテストは見あたらず，漢字(語彙)学習ストラテジーを日本語学習が進んだ者とそうでない者で比較する場合，先行研究では漢字知識以外の知識を問うテストによってグループ分けされたり，漢字知識を問う到達度テストによってグループ分けされたりすることが少なくなかった(大和 2018)。そこで漢字語彙知識を問うテストを開発する必要があると考えた。2 点目として，本テストが漢字語彙の中でも特に二字漢字語をターゲットとしたのは，二字漢字語が漢字語彙の大部分を占めると判断したからである。Yokosawa and Umeda (1988)によると，国語辞典に収録されている語の約 70%が漢字二字からなる漢字二字熟語だという。このことを踏まえると，二字漢字語の知識が漢字語彙の知識のすべてを網羅しているとは言えないものの大部分を占めており，二字漢字語の知識を測定することで，概ね漢字語彙の知識を測れるのではないかと考えた。

また，本研究で開発した漢字語彙テストで測定できると考えている知識は，漢字の書き取りテスト・読み取りテストで測定している知識とは異なるという点についても言及しておきたい。漢字の知識を問うテストとしてよく用いられる漢字の書き取りテストや読み取りテストでは，主に漢字(語彙)の持つ音韻情報と書字情報をつなげることができるかを漢字の書字を再生できる(書き取れる)か，漢字の読みをひらがなで再生できるかを通して問うている。それに対して，本テストは問題文の文脈に合う二字漢字語を選択肢から選択するものであり，主に漢字語彙の意味を問うているといえる。もちろん，選択肢構成によって，錯乱肢に音韻的に類似しているもの，書字的に類似しているものが含まれれば，ターゲットとする漢字語彙の意味だけを問うているとは言えない。また，漢字の書き取りテストや読み取りテストのような再生テストと比較して，本テストのような多肢選択形式の再認テストは難易度が低くなる(易しくなる)ことが広く知られており(Shohamy 1984; In’nami and Koizumi 2009 など)，「偶然の正解」も起こりうるテストである。しかしながら，本研究では宮岡・玉岡・酒井(2011)でも指摘されているテストの「実施容易性」も考慮し，ある程度限られた時間で実施可能で，かつ評価者による影響を受けにくい，主に漢字語彙の意味を問う多肢選択式のテストを開発することにした。

3. 漢字語彙テストの作成基準

本研究では，まず非漢字系日本語学習者を対象とした四者択一の漢字語彙テストを作成した(表 1)。作成にはターゲット語について，錯乱肢について，そして問題文についてそれぞれの作成基準を設けた。各基準は以下の通りである。

本テストの各項目の選択肢は，正答となる語であるターゲット語，錯乱肢ともにすべて漢字二字で構成されるものとした。ただし，語彙テストにおけるターゲット語の語種が正答率に与える影響は非常に限定的であると考えられること(大和・玉岡・茅本 2016)，さらに小室

リー(2019)が指摘するように，日本語学習者にとって学ぶべき漢字語彙には漢語だけでなく漢字表記の和語も含む必要があることから，本テストでもターゲット語の語種を漢語に限定せず漢字表記する漢字二字からなる語の知識を測定することとした。本語彙テストが様々な日本語能力の学習者が受験することを想定した熟達度テストであることから，ターゲット語の難易度のバランスをとるために，ターゲットとなる漢字語彙 36 語のうち，12 語が旧日本語能力試験語彙級 2 級レベル，12 語が 3 級レベル，12 語が 4 級レベルとなるように選択した。

さらにターゲット語の難易度だけでなく錯乱肢の難易度と構成も問題の難易度を左右する(Drum, Calfee, and Cook 1981)ため，各問の選択肢構成にも配慮した。まずターゲット語と錯乱肢の語としての難易度が同等になるように統制した。また，錯乱肢は「もっともらしく(plausible)見えるもの」である必要があるとされている(Haladyna, Downing, and Rodriguez 2002)ことを踏まえ，正答あるいは錯乱肢の 1 つに意味的・音韻的・書字的に類似しているものをおくよう調整した。本テストでは錯乱肢構成の条件を設定するにあたって，各レベル 12 語のうち 4 語の選択肢構成は錯乱肢に意味的にずれた語(あるいは非単語)を含むもの(「意味の類似」)，4 語の選択肢構成は錯乱肢に正答と語の読みが類似している語(あるいは非単語)を含むもの(「音韻の類似」)，4 語の選択肢構成は錯乱肢に正答の二字漢字語を構成する二字の漢字と書字的に類似している語(あるいは非単語)を含むもの(「字形の類似」)とした。さらに，36 問のうち 30 問では，錯乱肢の 1 つを非単語とした。ただし，錯乱肢の 1 つとする非単語は，例えばターゲット語「大人」(2 級レベル・意味の類似)と漢字が持つ意味が比較的類似している「高人*」とするなど，選択肢構成と関わりがあるものとした。なお，非単語も含め錯乱肢で使用する漢字は，旧日本語能力試験の漢字級でターゲット語の提出級と同じ，もしくは易しいものとした。また，選択肢構成「字形の類似」に分類される項目の錯乱肢には，例えば「事返*」のように，他の選択肢相当の難易度の実在語「返事」(3 級)を構成する二字の漢字の字順を交替させた非単語も含んでいる。これは，非漢字系日本語学習者に対して実施した漢字二字語の正誤判断課題の非単語の処理において，字順が交替している非単語の処理が非漢字系日本語学習者にとって他の非単語と比較して処理時間に遅延が生じ，さらに誤答率も高い(大和 2019)ことを参考にし，誤りを誘発しやすい錯乱肢になると考えたためである。

最後に問題文の作成においても，選択肢同様，問題文に含まれる語の難易度を調整し，ターゲット語の語彙級と同じかそれより低い(易しい)語と文型を使用し，問題文の難しさが解答に影響しないよう配慮した。また，問題文の中の語の漢字級がターゲット語の漢字級より高く難しい場合は，ひらがなで表記した。なお，問題文に含まれる文型の難易度は，『日本語能力試験出題基準(改訂版)』(2007)の文法級を参照した。

表1　本研究で使用した漢字語彙テストの問題文・ターゲット語・錯乱肢一覧

JLPT 語彙級	選択肢構成	問題番号	問題文	ターゲット（正答）	錯乱肢		
					①	②	③
2級	意味の類以	32	ドイツの小学校で（　　）を受けました。	教育	学習	入門	学導*
		28	みんなで来年の（　　）を立てます。	予定	未来	将来	予来*
		10	この作家の（　　）は映画になりました。	小説	書類	雑誌	説紙*
		2	息子は「（　　）になったら弁護士になりたい」と言いました。	大人	少年	姉妹	高人*
	音韻の類似	11	一人で（　　）な道を通りました。	危険	記憶	経験	計画
		13	試合で勝つためにみんな（　　）しました。	努力	労働	期待	協会
		3	山田さんは（　　）な先生として知られています。	熱心	出身	収入	禁止
		19	「環境」の（　　）は人気があります。	講義	準備	順番	途中
	字形の類似（字順交替）	12	交通安全のために（　　）を守りましょう。	規則	財布	現在	件用*
		5	事故の（　　）についておしえてください。	状況	順位	税金	険保*
		17	日本語の間違いを（　　）してもらいました。	修正	証明	検討	討検*
		35	この地方は気候が（　　）で、住みやすいです。	温暖	愛情	反応	想理*
3級	意味の類以	31	富士山の（　　）を撮りました。	写真	音楽	絵本	真楽*
		20	外国語の（　　）が大変です。	勉強	見学	通学	強科*
		1	私の国のおまつりについて（　　）します。	説明	注意	会話	話言*
		21	大学の（　　）はロシア語を勉強しています。	友達	家族	両親	友族*
	音韻の類似	23	毎週水曜日は（　　）です。	休日	関係	合格	急行
		24	去年（　　）で火事がありました。	工場	子供	卒業	子場*
		29	今日は先生がいないから（　　）します。	自習	近所	数字	授業
		26	部屋の（　　）をきれいにしました。	空気	意見	銀行	空計*
	字形の類似（字順交替）	8	一緒に（　　）へ行きませんか。	花見	例文	作文	社神*
		15	山田さんはいろいろな（　　）に興味を持っています。	文化	注文	休校	車発*
		25	（　　）に答えましたか。	質問	来週	新聞	事返*
		7	このふうとうは（　　）を使わなくてもいいです。	切手	親切	大切	字数*
4級	意味の類以	6	毎日（　　）まで勉強します。	夕方	一日	毎週	方夜*
		9	（　　）に5年間すみました。	外国	帰国	国語	帰外*
		16	山田さんは今（　　）です。	学生	毎日	百円	毎生*
		18	今日の夜、（　　）がありますか。	時間	毎月	五日	毎間*
	音韻の類似	22	（　　）一緒にテレビを見ましょう。	今晩	先月	先週	先今*
		27	朝、（　　）へ行きます。	学校	数学	科学	校数*
		14	今日はいい（　　）ですね。	天気	元気	電気	気日*
		33	（　　）、国へかえります。	今年	昨日	去年	去日*
	字形の類似	30	来年からこの（　　）ではたらきます。	会社	今年	社会	長今*
		34	（　　）はたんじょうびです。	今日	半日	日本	半長*
		36	（　　）でタシケントへ行きます。	電車	電力	電話	電日*
		4	このケーキはおおきいですから、（　　）だけたべます。	半分	自分	六分	分電*

注　⋆の語は，選択肢構成を考慮し作成した非単語である。

4. 漢字語彙テストの実施

本研究では，漢字語彙テストの有用性を検討するために，上述の基準で作成した漢字語彙テストをウズベク人日本語学習者 55 名に対して実施した。55 名の日本語学習者は全員ウズベク語を母語としており，ウズベキスタン国内で日本語を学んでいる。このうち 49 名は大学で，3 名はリッセイ[1]という中等教育機関で，3 名はカレッジ[2]という中等教育機関で日本語教育を受けている。学習者の日本語学習期間はさまざまであったが，全員日本語教育の中で漢字教育を受けた経験を持つ者であった。テストは，各教育機関の教師の監督の下，実施された。

5. 漢字語彙テスト結果の分析

5.1 基礎統計量と信頼性係数（クロンバックのα係数）

まず，テストの結果の概要を述べる。本テストの 1 問 1 点，36 点満点で計算したところ，平均点は 14.36 点（SD＝7.99 点）で，平均正答率は 39.89%であった。55 名の調査協力者で最高点は 36 点（満点），最低点は 5 点であった。尖度は 0.421，歪度は 1.313 であった。通常，尖度は-1.0 から+2.0，歪度は-1.0 から+1.0 の範囲であれば正規分布しているとみなされる(Huck 2000)。この基準に照らすと，本テストにおける調査協力者の得点の分布は正規分布しているとは言えず，やや低い得点層に多くの受験者が集まっている分布となっていることがわかる。以上のことから，個人差はあるものの，本研究における調査協力者にとって本テストは難しい項目が多かったと考えられる。

次に，信頼性係数として一般的に使用されるクロンバックのα係数を用いて，本テストの信頼性を確認する。クロンバックのα係数は 0 から 1 の間の数値で表され，α係数が 0.8 以上であれば，信頼性が確保されていると判断できる(中村 2002)。本テストのクロンバックのα係数を算出したところ，α＝0.895 であった。一般にクロンバックのα係数は項目数(問題数)が多い場合，またテストの受験者数が多い場合に高い値が出やすいと言われているが，本テストの調査協力者が 55 名で，項目数も 36 問といずれも決して多くないことを考えると，本テストは高い信頼性を確保しているテストであると判断することができる。

以上のテスト結果の概要と信頼性係数から，本研究の漢字語彙テストは正答率が低く，また得点が低い調査対象者が多かったことから難しいテストであったが，信頼性は確保されており，調査対象者の漢字語彙能力を識別することができるテストであったといえる。

5.2 TDAP の指標による各項目の適合度

各項目(問題)が有用なものであるか検討するために，本研究では「Test Data Analysis Program(TDAP)ver.2 」(大友・中村・秋山 2002) を用いて各項目の適合度を調べた[3]。なお，TDAP では各項目が「いい項目であるか」を示す指標として，項目困難度(DIFF)，項目弁別

力(DISC)，実質選択肢数(AENO)の 3 つの値が項目ごとに算出される。そして，以上の 3 つの値の総合得点ともいえる値(SATOT)をもって，項目の良し悪しが示される。

5.2.1 項目困難度(DIFF)

項目困難度は，正答率あるいは通過率とも呼ばれるもので，集団における各項目の難しさを表した指標で，項目ごとに算出される。項目困難度は 0.000 から 1.000 の間の値をとり，1 に近いほど易しく，0 に近いほど難しい項目であると判断することができる。本テストでは，項目困難度が最も高い，すなわち最も易しいターゲット語が「質問」(3 級・字形の類似; DIFF=0.564)であった。本テストは学習者の二字漢字語の知識を問う熟達度テストであったことから，到達度テストほど高い項目困難度とはならないことが想定されていた。しかしながら，36 問の項目の中で最も易しいものであっても半数程度の調査協力者しか正解できなかったということは明らかに項目が難しかったことを示しており，難易度の面では本研究の調査協力者に合ったものではなかったと考えられる。

一方，項目困難度が最も低い，すなわち最も難しいターゲット語は「教育」(2 級・意味の類似; DIFF=0.109)であった。本テストが四者択一の選択問題であるということを考えると，項目困難度が 0.109 であるということは，ランダムに選択した場合(0.25)より低いということになる。これは，単純にこの項目が難しい項目だったというだけではなく，錯乱肢の中に過剰に選択されていたものがあった可能性もあり，この項目の各選択肢の選択率の詳細をみる必要がある。

表 2 に示したのは，本テストの 36 項目を項目困難度の高い順に並べた上位 10 項目と下位 10 項目である。表 2 を参照すると，項目困難度が高い項目には必ずしも 4 級レベルの語があがっているわけではないことがわかる。一方，項目困難度が低い項目には下位 10 項目中 5 項目が 2 級レベルの語であるなど，比較的 2 級レベルの語が多くあがっているが，その中に 4 級レベルのターゲット語「今晩」(4 級・音韻の類似)も含まれている。この項目についても，問題文と選択肢構成になんらかの問題があった可能性がある。

表2　項目困難度上位10項目と下位10項目

DIFFによる順位	問題番号	ターゲット語	語彙レベル	選択肢構成	DIFF
1	25	質問	3級	字形の類似	0.564
2	1	説明	3級	意味の類以	0.545
3	24	工場	3級	音韻の類似	0.545
4	8	花見	3級	字形の類似	0.527
5	6	夕方	4級	意味の類以	0.527
6	34	今日	4級	字形の類似	0.509
7	20	勉強	3級	意味の類以	0.509
8	21	友達	3級	意味の類以	0.491
9	19	講義	2級	音韻の類似	0.491
10	36	電車	4級	字形の類似	0.473
⋮	⋮	⋮	⋮	⋮	⋮
27	10	小説	2級	意味の類以	0.345
28	23	休日	3級	音韻の類似	0.327
29	29	自習	3級	音韻の類似	0.309
30	22	今晩	4級	音韻の類似	0.309
31	12	規則	2級	字形の類似	0.309
32	5	状況	2級	字形の類似	0.273
33	2	大人	2級	意味の類以	0.273
34	7	切手	3級	字形の類似	0.255
35	15	文化	3級	字形の類似	0.255
36	32	教育	2級	意味の類以	0.109

5.2.2 項目弁別力(DISC)

項目弁別力は，識別力とも呼ばれるもので，その項目が成績がよい受験者とそうではない受験者を弁別することができるかを示した指標である。この指標には，各項目の正誤とテスト全体の正答数の相関(点双列相関係数)が使用されることが多い(野口・大隅 2012)。項目弁別力は-1.000 から+1.000 まで間の値を取り，値が +1.000 に近い項目ほど，受験者をより良く弁別することができる項目であり，項目弁別力指数の最適値は 0.3 以上であると言われている(中村 2002)。ただし，この値が極端に高い場合は，難しすぎる項目であったり，設問に不備があったりする可能性も考えられる。また値がマイナスになる場合には，本来正答すると考えられる受験者が不正解で，反対に正答できないと考えられる受験者が正解している現象がみられる，ということを意味している。つまり，そのような項目は，テストで測定しようとしているものを測れていない可能性があり，項目の差し替えが必要となると考えられる。本テストの項目弁別力の値を見てみると，最も高いもので 0.659(ターゲット語「自習」・3 級・音韻の類似)，最も低いもので 0.160(ターゲット語「説明」・3 級・意味の類似)と，項目弁別力の値がマイナスとなる，受験者の能力の弁別ができない項目は見当たらなかった[4]。しかしながら，項目弁別力の値 0.6 を超えるものが 8 項目あるなど，全体的にやや項目弁別力の値が高すぎるものが多い傾向にある。これらの 8 項目について項目困難度もあわせてみてみると，このテスト全体の項目困難度平均の 0.399 を 8 項目すべてが下回っていることがわかった。上述のように，項目弁別力が極端に高い項目については様々な問題がある可能性があるが，本テストの項目弁別力で高い項目がみられる原因の 1 つには，これら 8 項目が受験者にとって本テストの中でもとりわけ難しい項目であったことが考えられる。

5.2.3 実質選択肢数(AENO)

実質選択肢数とは，多肢選択式の項目において準備した選択肢が偏りなく選択されたかを検討する値である(中村 2002)。本研究で実施した漢字語彙テストは四者択一の選択式のテストであったため，表面上は 4 つの選択肢がある。しかしながら，それらの選択肢の中には受験者にとって簡単に正答の候補から除外されてしまう選択肢も存在することがある。実質選択肢数は，表面的に見える選択肢数ではなく，受験者の解答行動から実質的に機能していた選択肢の数を示すものである。したがって，本テストのように 4 つの選択肢がある場合は，実質選択肢数は 0.000 から 4.000 までの間の値を取ることになる。この値は選択肢数によって異なるため，どの程度の値が最適かは示されていないが，例えば実質選択肢数が実際の選択肢数の半分以下である場合などは，機能していない選択肢が多数あるということで，選択肢の改良が必要になると考えられる。本テストの各項目の実質選択肢数を見てみると，最も高かったのが「切手」(3 級・字形の類似; AENO=3.942)であった。一方，実質選択肢数が本テスト 36 問の中で最も低かったのが「大人」(2 級・意味の類似; AENO=2.652)であった。四者択一問題であった本テストの中で実質選択肢数が 3.000 未満であった項目は，実

表3 漢字語彙テストの項目別指標

SATOTによる順位	SATOT	問題番号	ターゲット語	語彙レベル	選択肢構成	DIFF	DISC	AENO
1	1.810	36	電車	4級	字形の類似	0.473	0.654	3.503
2	1.699	18	時間	4級	意味の類以	0.345	0.656	3.870
3	1.667	25	質問	3級	字形の類似	0.564	0.480	3.084
4	1.649	14	天気	4級	音韻の類似	0.345	0.637	3.830
5	1.639	21	友達	3級	意味の類以	0.491	0.470	3.464
6	1.618	34	今日	4級	字形の類似	0.509	0.513	3.259
7	1.604	30	会社	4級	字形の類似	0.455	0.575	3.405
8	1.597	29	自習	3級	音韻の類似	0.309	0.659	3.761
9	1.580	22	今晩	4級	音韻の類似	0.309	0.605	3.913
10	1.564	8	花見	3級	字形の類似	0.527	0.248	3.293
11	1.555	16	学生	4級	意味の類以	0.436	0.561	3.426
12	1.547	23	休日	3級	音韻の類似	0.327	0.579	3.842
13	1.546	9	外国	4級	意味の類以	0.473	0.508	3.322
14	1.544	26	空気	3級	音韻の類似	0.382	0.489	3.796
15	1.536	1	説明	3級	意味の類以	0.545	0.160	3.168
16	1.535	10	小説	2級	意味の類以	0.345	0.565	3.780
17	1.533	27	学校	4級	音韻の類似	0.400	0.520	3.634
18	1.525	7	切手	3級	字形の類似	0.255	0.605	3.942
19	1.520	11	危険	2級	音韻の類似	0.382	0.517	3.696
20	1.519	31	写真	3級	意味の類以	0.418	0.441	3.637
21	1.497	20	勉強	3級	意味の類以	0.509	0.245	3.271
22	1.480	15	文化	3級	字形の類似	0.255	0.616	3.806
23	1.476	24	工場	3級	音韻の類似	0.545	0.508	2.814
24	1.475	35	温暖	2級	字形の類似	0.436	0.299	3.586
25	1.473	19	講義	2級	音韻の類似	0.491	0.319	3.283
26	1.458	4	半分	4級	字形の類似	0.436	0.254	3.578
27	1.456	3	熱心	2級	音韻の類似	0.418	0.575	3.271
28	1.453	28	予定	2級	意味の類以	0.400	0.460	3.558
29	1.440	6	夕方	4級	意味の類以	0.527	0.312	3.031
30	1.434	33	今年	4級	音韻の類似	0.364	0.448	3.690
31	1.419	5	状況	2級	字形の類似	0.273	0.565	3.759
32	1.416	17	修正	2級	字形の類似	0.364	0.240	3.836
33	1.392	13	努力	2級	音韻の類似	0.364	0.216	3.793
34	1.288	12	規則	2級	字形の類似	0.309	0.299	3.716
35	1.070	2	大人	2級	意味の類以	0.273	0.657	2.652
36	0.985	32	教育	2級	意味の類以	0.109	0.291	3.307

注 DIFFは項目困難度，DISCは項目弁別力，AENOは実質選択肢数，
SATOTは，これら3つの標準適合度の総合評価である。

質選択肢数が最も低かった「大人」の他に「工場」(3 級・意味の類似; AENO=2.814)のみであった。このことから，本テストで提示した選択肢の多くは本研究の調査協力者にとって有効に機能していたと考えることができよう。

5.2.4 項目困難度，項目弁別力，実質選択肢数の標準適合度(SATOT)の合計

TDAP では，ここまで結果を述べてきた項目困難度，項目弁別力，実質選択肢数の数値を踏まえ各項目の良し悪しを総合的に評価する SATOT が算出される。表 3 は，この SATOT の高い順に本テストのターゲット語と各指標の値を並べたものである。表 3 によると，最も SATOT が高かった，すなわち 36 問中最も「よい項目」であるとされたのが「電車」(4 級・字形の類似; SATOT=1.810)であった。一方，最も SATOT の値が低かった項目が「教育」(2 級・意味の類似; SATOT＝0.985)であった。「教育」は項目困難度が 36 問の項目の中で最も低かった項目でもあり，この点が SATOT の値に影響したと考えられる。

そのほかに，表 3 から以下の 2 点がわかる。第 1 に，項目の良し悪しにターゲット語の難易度が影響しているということである。SATOT による順位が高い「よい項目」とされるものに，語彙レベルの低い 4 級レベルの語が多く，順位が下がっていくにつれて，語彙レベルの高い 2 級レベルの語が多くなっている。つまり，比較的ターゲット語の難易度が低い(易しい)語のほうが「よい項目」と判定されやすかった。これは，今回の調査協力者の日本語能力に本テストのターゲット語 3 級レベルおよび 4 級レベルの語が適合していたことを示していると考えられる。とはいえ，例えば「夕方」(4 級・意味の類似; SATOT=1.440)は項目困難度 0.527 と本テストの中では項目困難度が高い項目であっても，SATOT の順位では 36 項目中 29 位と，あまり「よい項目」とは評価されていないように，易しいと思われる語が必ずしも「よい項目」になるとは限らないということもわかる。これは，項目困難度のみならず，他の指標や受験者の解答行動の影響を受けたためであると考えられる。

第 2 に，項目の良し悪しに本テストの作成にあたって考慮した選択肢構成による影響はほとんどないと考えられることである。SATOT の高い順に並べた表 3 を見る限り，特定の選択肢構成の項目が上位あるいは下位に並んでいるということはないことがわかる。つまり，問題の質に影響するのは選択肢構成の種類(「意味の類似」「音韻の類似」「字形の類似」)より，個々の項目が持つそれ以外の要因であろうと考えられる。この点を明らかにするためには，各項目の錯乱肢も含めた選択肢の選択率，さらに受験者の中の上位群・中位群・下位群がどのような解答行動をとったかを詳しくみる必要がある。

5.3 各選択肢の選択率からの分析

TDAP で算出された各項目の項目困難度，項目弁別力，実質選択肢数からいくつかの項目について，錯乱肢が過剰に機能している可能性や選択肢構成に問題がある可能性を指摘した。この点について明らかにするために，表 4 に各項目の選択肢の選択率を一覧にしてまと

表4 各項目の正答･錯乱肢と選択率一覧

JLPT 語彙級	選択肢構成	問題番号	正答と選択率(%)		錯乱肢と選択率(%)					
2級	意味の類以	32	教育	10.9	**学習**	**50.9**	入門	12.7	学導*	25.5
		28	**予定**	**40.0**	未来	30.9	将来	9.1	予来*	20.0
		10	**小説**	**34.5**	書類	29.1	雑誌	12.7	説紙*	23.6
		2	大人	27.3	**少年**	**60.0**	姉妹	10.9	高人*	1.8
	音韻の類似	11	**危険**	**38.2**	記憶	16.4	経験	30.9	計画	14.5
		13	**努力**	**36.4**	労働	29.1	期待	18.2	協会	16.4
		3	**熱心**	**41.8**	出身	12.7	収入	38.2	禁止	7.3
		19	**講義**	**49.1**	準備	12.7	順番	29.1	途中	9.1
	字形の類似（字順交替）	12	規則	30.9	財布	10.9	**現在**	**34.5**	件用*	23.6
		5	状況	27.3	順位	20.0	**税金**	**38.2**	険保*	14.5
		17	**修正**	**36.4**	証明	25.5	検討	21.8	討検*	16.4
		35	**温暖**	**43.6**	愛情	14.5	反応	27.3	想理*	14.5
3級	意味の類以	31	**写真**	**41.8**	音楽	12.7	絵本	27.3	真楽*	18.2
		20	**勉強**	**50.9**	見学	18.2	通学	23.6	強科*	7.3
		1	**説明**	**54.5**	注意	20.0	会話	18.2	話言*	7.3
		21	**友達**	**49.1**	家族	14.5	両親	21.8	友族*	14.5
	音韻の類似	23	**休日**	**32.7**	関係	29.1	合格	23.6	急行	14.5
		24	**工場**	**54.5**	子供	9.1	卒業	32.7	子場*	14.5
		29	自習	30.9	**近所**	**34.5**	数字	21.8	授業	12.7
		26	**空気**	**38.2**	意見	25.5	銀行	16.4	空計*	20.0
	字形の類似（字順交替）	8	**花見**	**52.7**	例文	21.8	作文	14.5	社神*	10.9
		15	文化	25.5	注文	12.7	**休校**	**30.9**	**車発***	**30.9**
		25	**質問**	**56.4**	来週	14.5	新聞	21.8	事返*	7.3
		7	切手	25.5	親切	18.21	**大切**	**29.1**	字数*	27.3
4級	意味の類以	6	**夕方**	**52.7**	一日	9.1	毎週	30.9	方夜*	7.3
		9	**外国**	**47.3**	帰国	10.9	国語	30.9	帰外*	10.9
		16	**学生**	**43.6**	毎日	12.7	百円	32.7	毎生*	10.9
		18	**時間**	**34.5**	毎月	27.3	五日	20.0	毎間*	18.2
	音韻の類似	22	**今晩**	**30.9**	先月	21.8	先週	18.2	先今*	29.1
		27	**学校**	**40.0**	数学	14.5	科学	30.9	校数*	14.5
		14	**天気**	**34.5**	元気	27.3	電気	23.6	気日*	14.5
		33	**今年**	**36.4**	昨日	18.2	去年	32.7	去日*	12.7
	字形の類似	30	**会社**	**45.5**	今年	10.9	社会	30.9	長今*	12.7
		34	**今日**	**50.9**	半日	10.9	日本	27.3	半長*	10.9
		36	**電車**	**47.3**	電力	23.6	電話	16.4	電日*	12.7
		4	**半分**	**43.6**	自分	12.7	六分	27.3	分電*	16.4

注1 太字で記した選択肢および選択率は，各項目でもっとも選択率が高かった選択肢を示す。

注2 ⋆の語は，選択肢構成を考慮し作成した非単語である。

めた。各項目の選択肢で太字になっている選択肢および選択率は，その項目において最も選択率が高かったものを表している。表 4 をみると，本テストの 36 項目のうち 29 項目は正答であるターゲット語が最も選択率の高い選択肢となっていたが，7 項目では正答より錯乱肢がより選択されていた。このような項目は 2 級レベルの語をターゲットとした項目 4 項目， 3 級レベルの語をターゲットとした項目 3 項目でみられたが， 4 級レベルの語ではみられなかった。4 級レベルの語はそもそも本テストの受験者である調査協力者にとっても比較的難易度が低い易しい語が多く，迷わず正答を選択できたと考えられるが，2 級レベルの語と 3 級レベルの語で正答以外の語(錯乱肢)の選択率が高かった項目については，単純に難しかっただけではなく，他に特定の選択肢が機能しすぎて正答を選択しにくかった可能性がある。ここでは，特定の錯乱肢の選択率が突出して高く特徴的であったターゲット語「教育」(問題番号 32・2 級・意味の類似)と「大人」(問題番号 2・2 級・意味の類似)を例に，問題文とともに，なぜその錯乱肢が受験者にとって魅力的な選択肢であったかをみていく。

まず，ターゲット語「教育」の問題文，選択肢，そして各選択肢の選択率を示す。※は正答を表し，各選択肢の後の(　)は選択率を表す(以下，同じ)。

32.　ドイツの小学校で(　　)を受けました。
　a)　入門(12.7%)　　b)　学習(50.9%)　　c)　学導(25.5%)　　d)　教育※(10.9%)

この項目の正答は d)であったが，4 つの選択肢の中で正答が最も選択率が低かった。このように，正答が最も選択されなかった項目は，本テスト 36 項目の中でこの項目だけであった。多くの受験者は問題文に含まれる「小学校」という語から「勉強」「学習」「教育」に関わる語が(　)内に入ると考えたと思われるが，選択肢として提示された語の多くは，どれも受験者にとっては「もっともらしく見える」語であったと考えられる。実際，TDAP の算出したこの項目の実質選択肢数(AENO)は 3.307 で，どの選択肢もある程度受験者にとって，正答となりうる選択肢に見えていたことが示されている。その中でも選択率が 50.9%であった「学習」は受験者であった調査協力者には親密度が高い語であったと考えられる。また，選択肢 c)の「学導」は実在しない非単語で，ターゲット語である「教育」に意味的に類似している漢字二字を組み合わせて作成したものであった。しかし選択率を見ると，非単語である c)の「学導」が最も受験者に選択された b)の「学習」の次に多くの受験者に選択されているという結果となった。一方，正答の「教育」は他の選択肢と比較して特に使用頻度が低い語とは考えにくい。しかし，この項目がターゲットとなっている語の理解のみではなく，設問文にある「受ける」と共起する語を選択できるかという点まで問うていたため，このように意味的に類似した選択肢が提示されたとき，受験者の間に混乱が起こったものと考えられる。

次に，ターゲット語「教育」の問題文，選択肢，そして各選択肢の選択率を示す。

2.　息子は「(　　)になったら弁護士になりたい」と言いました。

a)　高人(1.8%)　　b)　少年(60.0%)　　c)　姉妹(10.9%)　　d)　大人※(27.3%)

この項目の正答は d)であったが，錯乱肢である b)の「少年」が 60.0%もの選択率であった。この項目が前掲の問題番号 32 のターゲット語「教育」の項目と明らかに異なるのは，実質選択肢数(AENO)が 2.652 と低いということである。これは，具体的には，最も選択された錯乱肢 b)「少年」と正答である d)「大人」の 2 つが受験者にとって実質的に機能した選択肢であったということを示している。この項目の実質選択指数が低かった原因はいくつか考えられるが，その 1 つが問題文内の漢字で表記された「息子」(語彙級 3 級，漢字級 2 級)になじみがなく，語を構成する「子」の部分に着目し，「子」と最も近い意味を持つと考えられる錯乱肢 b)「少年」を選択してしまう受験者が多かったという可能性である。さらに，問題文には確定条件の「たら」が使われており，本テストのように漢字語彙の意味の理解を問う項目としては受験者に負担のかかる文型を含む問題文であったことが，正答率に影響している可能性もある。

ここでは特定の錯乱肢の選択率が突出して高く特徴的であった項目としてターゲット語「教育」(問題番号 32・2 級・意味の類似)と「大人」(問題番号 2・2 級・意味の類似)の選択率を見てきたが，これらの項目は項目困難度が低かったというだけではなく，項目の質の総合評価を表す SATOT の値も両項目とも低く，SATOT を高い順に並べた順位では「教育」が 36 項目中 36 位，「大人」が 35 位であった。この 2 つの項目の各選択肢の選択率を見たところ，どちらの項目もターゲットとなっている漢字語彙の意味の他に，ターゲットとなった名詞と動詞との共起関係の知識，問題文の中の文型の理解がないと正答が選べないようになっていたため，項目困難度も低くなり，かつ多くの受験者に混乱が起こったものと考えられる。この結果からもわかるように，これらの 2 項目は，他の項目で問う以上の知識を問う項目になっていたという点で本テスト全体の質を下げる恐れがある項目であると言えるため，問題文や選択肢の差し替えが必要になると考えられる。

5.4　G-P 分析(Good-Poor Analysis)

G-P 分析とは，テスト受験者をテストの合計点で上位群・中位群・下位群にグループ分けし，各グループの受験者の解答行動を詳しく分析するものである[5]。この分析方法で，項目困難度，項目弁別力だけでは分かりにくいグループ間の各選択肢の選択率を比較することによって，どの選択肢が受験者を識別するのに有効な選択肢であったか知ることができる。G-P 分析では，正答である選択肢の選択率(正答率)は，上位群の方が下位群より高いことが望ましいが，そうなっていない場合は，選択肢や問題文を検討し，修正する必要があるとされている(服部 1999)。G-P 分析の上位群・中位群・下位群のグループ分けには，(1)全調査

表5　項目弁別力(DISC)0.25未満の項目のG-P分析結果

問題番号	問題文		選択肢と各グループの選択率							
8	一緒に(　　)へ行きませんか。		a) 例文	21.8%	b) 花見※	52.7%	c) 作文	14.5%	d) 社神*	10.9%
	(3級・字形の類似)		選択者数	選択率	選択者数	選択率	選択者数	選択率	選択者数	選択率
	DIFF=0.527	上位群(N=15)	2	13.3%	10	66.7%	0	0.0%	3	20.0%
	DISC=0.248	中位群(N=25)	7	28.0%	11	44.0%	7	28.0%	0	0.0%
	AENO=3.293	下位群(N=15)	3	20.0%	8	53.3%	1	6.7%	3	20.0%
20	外国語の(　　)が大変です。		a) 見学	18.2%	b) 通学	23.6%	c) 勉強※	50.9%	d) 強科*	7.3%
	(3級・意味の類似)		選択者数	選択率	選択者数	選択率	選択者数	選択率	選択者数	選択率
	DIFF=0.509	上位群(N=15)	2	13.3%	2	13.3%	11	73.3%	0	0.0%
	DISC=0.245	中位群(N=25)	5	20.0%	8	32.0%	10	40.0%	2	8.0%
	AENO=3.271	下位群(N=15)	3	20.0%	3	20.0%	7	46.7%	2	13.3%
17	日本語の間違いを(　　)してもらいました。		a) 証明	25.5%	b) 検討	21.8%	c) 修正※	36.4%	d) 討検*	16.4%
	(2級・字形の類似)		選択者数	選択率	選択者数	選択率	選択者数	選択率	選択者数	選択率
	DIFF=0.364	上位群(N=15)	5	33.3%	2	13.3%	8	53.3%	0	0.0%
	DISC=0.240	中位群(N=25)	3	12.0%	7	28.0%	9	36.0%	6	24.0%
	AENO=3.836	下位群(N=15)	6	40.0%	3	20.0%	3	20.0%	3	20.0%
13	試合で勝つためにみんな(　　)しました。		a) 協会	16.4%	b) 労働	29.1%	c) 努力※	36.4%	d) 期待	18.2%
	(2級・音韻の類似)		選択者数	選択率	選択者数	選択率	選択者数	選択率	選択者数	選択率
	DIFF=0.364	上位群(N=15)	1	6.7%	4	26.7%	7	46.7%	3	20.0%
	DISC=0.216	中位群(N=25)	5	20.0%	10	40.0%	5	20.0%	5	20.0%
	AENO=3.793	下位群(N=15)	3	20.0%	2	13.3%	8	53.3%	2	13.3%
1	私の国のおまつりについて(　　)します。		a) 説明※	54.5%	b) 注意	20.0%	c) 会話	18.2%	d) 話言*	7.3%
	(3級・意味の類似)		選択者数	選択率	選択者数	選択率	選択者数	選択率	選択者数	選択率
	DIFF=0.545	上位群(N=15)	10	66.7%	0	0.0%	4	26.7%	1	6.7%
	DISC=0.160	中位群(N=25)	12	48.0%	6	24.0%	5	20.0%	2	8.0%
	AENO=3.168	下位群(N=15)	8	53.3%	5	33.3%	1	6.7%	1	6.7%

注1　DIFFは項目困難度，DISCは項目弁別力，AENOは実質選択肢数を示す。

注2　太字で記した選択率は調査協力者55名全員の選択率で，網掛けの選択肢は各項目で最も選択率が高かったものを示す。

注3　※の語は正答，*の語は選択肢構成を考慮し作成した非単語である。

協力者を合計点に基づいて 3 分割する(各群 33%に分ける)方法(ブラウン 1999)，(2)全調査協力者の合計点に基づいて上位と下位 27%をそれぞれ上位群・下位群とし，どちらにもあてはまらなかった 46%を中位群とする方法，の 2 つの方法のどちらかを採用することが一般的である。本研究では後者の方法で，55 名の調査協力者を本テストの得点(36 点満点)に基づいて，上位群 15 名，中位群 25 名，下位群 15 名に分け，選択率を比較した。

ここでは，本テストの改善点を明らかにするために，TDAP で示された項目弁別力の値を参考にいくつかの項目を取り上げ，各グループの選択率に着目することにする。項目弁別力は既に述べた通り，その項目に正答すべき上位群は正答を選択することができ，そうではない下位群は錯乱肢を選んでいる場合に，高い値が出る。そうであれば，この項目弁別力の値が低い項目については上位群が特定の錯乱肢を選んでいたり，上位群と下位群が同じ程度ある選択肢を選択していたりする(つまり，その項目では上位群と下位群を識別できていない)と予想される。そのため，TDAP の項目弁別力の値によって，何らかの問題がある可能性がある項目に着目し，その項目について G-P 分析の結果をみていくことは，ある程度項目改善に有益であると考えられる。そこで，ここでは項目弁別力の値が 0.25 を下回る 5 項目について各グループの選択率をみていくことにする(表 5)。

まず，表 5 にあげた項目弁別力 0.25 未満の 5 項目に共通しているのは，項目困難度(DIFF)と実質選択肢数(AENO)の値には大きな問題がみられないということである。本語彙テストは全体的に調査協力者にとって難しい項目が多かった中で，本テストの各項目の項目困難度の平均である DIFF=0.399 を大幅に下回る項目は表 5 の 5 項目にはなく，むしろ本テストの中では項目困難度が高い方に分類される DIFF=0.5 を超える項目もあった。実質選択肢数については，本テストのほとんどの項目において良好であると考えられる値が出ていたが，表 5 の 5 項目にも実質選択肢数に問題があると認められる項目はなかった。

次に，表 5 にあげた 5 項目の各グループの選択率を順にみていく。

問題番号 8 のターゲット語「花見」(3 級・字形の類似)の各グループの選択率を見てみると，すべてのグループで 4 つの選択肢の中で正答の選択率が最も高いことがわかる。この項目の項目困難度も DIFF=0.527 と本テストの中では高く，易しい項目であったことがうかがえる。ただし，下位群の学習者は中位群の学習者より正答の選択率が高く，中位群と下位群の間において「正答すべき人が正答し，そうではない人が正答できない」という原則とは逆のことが起こっていると考えられる。この項目に関して，中位群の学習者は正答の他に a)「例文」や c)「作文」を選択した者もおり，その割合は下位群の学習者より高かった。錯乱肢である a)も c)も問題文には意味的に合致しないことが明らかな語であったが，中位群の学習者には意味は分からないが見覚えのある語であったため，選択した学習者もいたかもしれない。逆に下位群の学習者には未習の語であったため，錯乱肢 a)や c)はあまり積極的に選択されなかった可能性がある。また，ターゲット語である「花見」については，日本語そのものの知識以外を問う項目になっていた可能性も否定できない。「花見」は国際交流基

金・日本国際教育協会(2007)の日本語能力試験(旧試験)の語彙リストにも含まれるが，この語は日本事情をよく知る者には日本語の語彙知識そのものが高くなくても正答が選べる項目であったとも考えられる。そうであれば，このターゲット語の出題は日本語の漢字語彙知識を測る項目としてはふさわしくないといえる。

問題番号 20 のターゲット語「勉強」(3 級・意味の類似)は，上位群の 73.3%が正答を選ぶことができたという点においては，正答すべき受験者が正答できている項目であると考えていいだろう。しかし，中位群と下位群の正答の選択率を比較したところ，わずかではあるが下位群の方が正答の選択率が高い。この理由として考えられることの 1 つは，中位群にとって錯乱肢である b)「通学」が機能しすぎていることである。中位群の選択肢 b)の選択率は 32.0%であるが，下位群にとっては b)は他の錯乱肢と同程度にしか機能していない。問題文の空欄に「通学」があてはまるわけではないため，問題に不備があったとは言えない。しかし，「通学が大変だ」という表現自体は存在し，「通学」という語が既習であればこそ，中位群には b)が「もっともらしい」選択肢に映ったのかもしれない。

問題番号 17 のターゲット語「修正」(2 級・字形の類似)の各選択肢の選択率は，グループによって異なる特徴がみられた。上位群の正答である c)「修正」の選択率は 53.3%で，次に錯乱肢である a)「証明」(上位群の選択率 33.3%)が選択されていた。上位群のほとんどはこの 2 つの選択肢のいずれかを選択していた。一方, 中位群は, 正答である c)の選択率は 36.0%で 4 つの選択肢の中で最も選択されていたが，それ以外の 3 つの錯乱肢もまんべんなく選択されていた点が上位群とは異なった。選択肢の中には錯乱肢 b)「検討」(中位群の選択率 28.0%)の字順を交替させた非単語である錯乱肢 d)「討検」も含まれていたが，この選択率も 24.0%であった。上位群ほど明確に正答を選択できなかった中位群の学習者の中には，同じ項目の中によく似た選択肢があったことで, テスト・テイキング・ストラテジー(Cohen 1998)を働かせ, b)か d)のどちらかが正答であると予想した学習者もいた可能性がある。下位群は，そもそも最も選択されていた選択肢が上位群や中位群とは異なり，a)の選択率が 40.0%と，下位群には最も選択されていた。それ以外の(正答を含めた)3 つの選択肢はそれぞれ 20.0%の選択率であった。この項目のターゲット語「修正」は 2 級語彙であり，下位群には未習語であったことが，このような解答行動につながっているのかもしれない。また，下位群の間では最も選択率が高かった a)「証明」は，上位群にも正答の次によく選ばれていたが，3 級語彙の「説明」と意味を混同している可能性も否定できない。

問題番号 13 のターゲット語「努力」(2 級・音韻の類似)では，上位群と下位群は正答を選択した学習者が最も多かったが，中位群では錯乱肢である b)「労働」の選択率が最も高かった。その結果，正答の選択率は中位群が下位群より低く，グループ間の正答率は上位群＞下位群＞中位群の順になっている。この点が，この項目の項目弁別力の値が低かった要因であろう。中位群の学習者が多く選択した錯乱肢 b)は上位群にも正答の次に選択されており，よく機能した選択肢であったといえる。錯乱肢 b)は問題文にあてはまる語ではないが，正

答同様，語を構成する漢字の構成要素に「力」が含まれていて，学習者によっては語の意味は分からないが，文の意味にふさわしい語だと推測したのかもしれない。一方で，下位群の学習者はこの項目の解答に上位群と同程度の割合で正答を選ぶことができた。この項目のターゲット語「努力」と錯乱肢は 2 級語彙で，ターゲット語も含め下位群の学習者にはすべての選択肢が未習であった可能性も高い。その中でターゲット語の「努力」を構成する漢字「力」(漢字級 3 級)から意味を推測して正答を導いた学習者もいるかもしれない。本テストが多肢選択式のテストである以上，受験者の推測により「偶然に」正解することは想定しておかなければならないが，錯乱肢の中に易しい漢字級を含む語や意味的に近い二字漢字語(あるいは非単語)をおく工夫をすることで，少なくとも下位群にとって正答が選びやすくなることは避けられると考えられる。

問題番号 1 のターゲット語「説明」(3 級・意味の類似)では，上位群・中位群・下位群すべてのグループで正答である a)「説明」が最も選択されていた。しかしながら，その選択率は上位群が 66.7%，中位群が 48.0%，下位群が 53.3%と中位群の選択率が最も低く，下位群は上位群とそれほど差がない選択率であった。つまり，中位群と下位群の間に正答率が逆転しているような現象がみられた。この点が，この項目の項目弁別力が低かった要因の 1 つになっていると考えられる。受験者の能力が識別できる項目の特徴として，本テストのような多肢選択式のテストであれば，正答の選択肢は上位群ほど選択率が高く下位群ほど選択率が低くなることがあげられる。そして，錯乱肢の選択率は上位群ほど低く下位群ほど高い選択肢が受験者の能力を識別するためによく機能しているといえる。これに照らしてこの項目の錯乱肢の選択率を見てみると，選択肢 b)「注意」は上位群の選択率は 0%であるが下位群の選択率は 33.3%と，正解できないと予想される受験者ほど選択する傾向にある，受験者の能力の識別に効果的であった錯乱肢であった。しかし，選択肢 c)「会話」の選択率をみてみると，上位群の選択率が 26.7%，中位群の選択率が 20.0%，下位群の選択率が 6.7%と上位群ほどこの錯乱肢を選択していることがわかる。その理由を探るために，この項目の問題文とあわせて選択肢 c)をみてみると，「私の国のおまつりについて会話します」となる。この文の「会話します」を「話します」と同義であると考えたり，問題文には書かれていないが「〇〇さんと」が省略されていると考えたりすると，選択肢 c)を(　)に入れた場合でも正しい文として許容できると考える人もいるかもしれない。本来，このような多肢選択式の語彙テストでは正答は各項目に 1 つであることが普通であるため，複数の選択肢が空欄にあてはまると考えた受験者には混乱を招く恐れがある。以上のように考えると，この項目の質を上げるために，問題文を改変し正答しかあてはまらないように文脈を規定するか選択肢 c)を別の語と入れ替えるなどの修正をし，上位群に錯乱肢が選択され過ぎないような工夫をする必要があると考えられる。

6. まとめ

本稿では，非漢字系日本語学習者に向けの漢字語彙の意味を問う多肢選択式のテストを開発し，ウズベキスタンで学ぶ日本語学習者に対してこのテストを実施し，その有用性を検討した。その結果，本テストには以下のような特徴がみられることがわかった。

まず，本テストは日本語の二字漢字語の意味を問うテストとして，十分な信頼性を確保していることが確認された(α =0.895)。この結果は，本テストの項目数と調査協力者(受験者)数から考えると十分に高い値であった。このように本テストが高い信頼性を確保できた理由として，本テストの計 36 項目の項目弁別力平均が 0.465 と多くの項目が良問であると判断される値(0.3 以上)を超えており，受験者の能力を識別できる項目が多かったこと，四者択一の本テストの各項目の実質選択肢数もほとんどの項目で 3.000 を超えており，受験者にとって機能する選択肢が多かったことが考えられる。以上のような点から，テスト全体の質はおおむね良好であると判断できる。

ただし，本テストは，非漢字系日本語学習者にとって難しい項目が多かったこともわかった。本テストでターゲットとした語は日本語能力試験(旧試験)の基準で 2 級レベルのものから 4 級レベルのものまでさまざまであった。しかし，項目困難度および項目弁別力の指標を見る限り，項目困難度の平均は 0.399 と全体的に低く，項目弁別力の低い項目には 2 級レベルの語，3 級レベルの語が多かった。これはターゲット語が難しい項目であるために項目困難度が低くなり，本来正答すべき受験者まで正答を選べなかったなどで，項目が受験者の能力を識別できなかったことを意味している。一方で，項目弁別力に関しては，その値が高すぎる項目も望ましくないとされているが，本テストでは項目弁別力が 0.6 を超えるような項目もあった。項目弁別力が異常に高くなるのもまた，その項目が受験者にとって難しすぎる項目であるためだといわれており，本テストで項目弁別力が高すぎる項目にも同じ傾向がみられた。本稿で報告した項目困難度および項目弁別力の値は，受験者がかわることで多少変動するため，非漢字系日本語学習者全般に難しい項目が多かったとは言い切れない。しかしながら，本テストの難しい項目はテストの質を左右しかねないものであることから，今後このテストを別の受験者に対して実施する場合には，ターゲット語の難易度が低い項目を増やし，特に難しいと判断された項目との入れ替えることも必要になってこよう。

本テストでは，TDAP がテストの適合度を示す指標を参照するだけではなく，各項目の各選択肢の選択率からの分析や G-P 分析などを通して，受験者の解答行動の特徴から問題の良し悪しを考察することも試みた。その結果，ターゲットとなる二字漢字語そのものの意味のみではなく，共起語の知識や問題文に含まれる文型の知識までも問う形になっていた項目もあり，それらの項目では，特定の錯乱肢の選択率が突出して高いという現象が見られることもわかった。もちろん，そのような項目は「いい項目」とは言えない。また，上位群・中位群・下位群で異なる解答行動を見せる項目もあり，上位群に機能しすぎる錯乱肢があったために，上位群の正答率が中位群や下位群と比較しても高いといえない項目もみられた。

また，中位群が下位群より正答率が低い項目もあり，その各グループの解答行動には選択肢構成や選択肢の難易度・親密度が影響していることがうかがえた。さらに，本テストの項目の中には項目の不備とまでは言えないまでも，問題文の改訂や選択肢の入れ替えが必要だと考えられる項目もあった。このように項目の選択肢の選択率を詳しくみていくことにより，本テストの改善すべき点も確認できた。

最後に，今後の課題を述べる。本稿ではウズベク人日本語学習者 55 名を対象に実施したテストの有用性について検討してきたが，調査協力者の人数は決して多いとは言えず，この結果をもって非漢字系日本語学習者の解答行動の特徴を一般化することはできない。今後は，調査対象者を増やし，本テストの有用性について，引き続き検討していく必要がある。また，本テストでは，テストの中にいくつかの選択肢構成を設定したり，錯乱肢の中に「それらしい」非単語を選択肢としておいたりと，選択肢に工夫をした。しかしながら，実際のところ選択肢構成がどの程度，項目の質に影響したのか，また非単語がどの受験者グループにどの程度機能しうるか，などは十分に検討することができなかった。これらの分析については，稿を改めたい。

【参考文献】

Cohen, D.(eds) (1998). *Strategies in Learning and Using a Second Language*. London and New York : Longman.

Drum, P.A., Calfee, R.C., & Cook, L.K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16, 486-514.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd Ed.). Englewood Cliffs, NJ: PrenticeHall.

Gamage, G. H. (2003). Perceptions of kanji learning strategies: Do they differ among character and alphabetic background learners? *Australian Review of Applied Linguistics*, 26(2), 17-31.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15, 309-333.

Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Longman.

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

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

Yokosawa, K. & Umeda, M. (1988). Processing in human kanji-word recognition. *Proceedings of the 1988 IEEE international conference on systems, man, and cybernetics*, 378-380.

大友賢二(1996)『言語テスト・データの新しい分析法　項目応答理論入門』大修館書店.
国際交流基金・日本国際教育協会(2007)『日本語能力試験出題基準【改訂版】』凡人社.
小室リー郁子(2019)『中国語母語話者のための漢字語彙研究　母語知識と活かした教育をめざして』くろしお出版.
J.D.ブラウン (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.
中村洋一 (2002)『テストで言語能力は測れるか～言語テストデータ分析入門～』桐原書店.
野口裕之・大隅敦子(2012)「テスティング・評価」近藤安月子・小森和子編『研究社日本語教育事典』337-360.研究社.
服部環(1999)「G-P 分析 good-poor analysis」中島義明・安藤清志・子安増生・坂野雄二・繁桝数男・立花政夫・箱田裕司・編集『心理学辞典』357-358.有斐閣.
宮岡弥生・玉岡賀津雄・酒井弘(2011)「日本語語彙テストの開発と信頼性－中国語を母語とする日本語学習者のデータによるテスト評価－」『広島経済大学研究論集』34(1), 1-18.
大和祐子(2018)「非漢字圏日本語学習者の漢字学習ストラテジー」『グローバル化時代における日本語教育と日本研究』318-329.
大和祐子(2019)「日本語学習者による二字漢字語の書字的認知処理－非単語の処理に着目して－」『日本語・日本文化研究』29, 17-32.
大和祐子・玉岡賀津雄・茅本百合子(2016)「フィリピン人日本語学習者のデータを基にした非漢字圏学習者向け語彙テストの開発と評価」『ことばの科学』30, 39-58.

【分析使用ツール】
大友賢二・中村洋一・秋山實(2002) Test Data Analysis Program (TDAP) Ver.2.0 [Windows 版]（中村洋一 (2002) 『テストで言語能力は測れるか～言語テストデータ分析入門』大友賢二監修, 桐原書店, に添付された解析ソフト）

付記：本稿は，科学研究費補助金・基盤研究(C)「日本語学習者による漢字語彙の認知処理の特徴：学習者は漢字をどう捉えているか」(課題番号：19K00737，研究代表者，大和祐子)の助成を受けた研究成果の一部である。

[1] リッセイはウズベキスタンにおける義務教育機関であり，各リッセイでは人文科学，外国語，経済，農業など特定分野に重点をおいた教育が行われている。本研究の調査協力者は外国語教育に重点を置いたリッセイで学んでおり，そこで日本語を学んでいた。リッセイの学生は卒業後，大学で日本語を専攻する場合もある。
[2] カレッジもリッセイ同様，ウズベキスタンにおける義務教育を担っている。ウズベキスタンにおけるカレッジは，一般的な科目のほかに，美容，観光，料理などの専門的なスキルを

身につけ資格取得を目指す学生を対象とした教育機関である。本研究の調査協力者は，観光ガイドを目指す者を養成するカレッジで日本語を学んでいた。カレッジの学生もリッセイの学生同様，卒業後大学へ進学し，より日本語を学ぶことも可能である。

[3] TDAP では，本稿で扱った項目困難度・項目弁別力・実質選択肢数といった古典的テスト理論(Classical Test Theory; CTT)に基づいた数値のみではなく，項目応答理論(Item Response Theory; IRT)に基づいた値も算出することができる。ただし，それにはある程度多くのサンプル数が必要である。本稿では，サンプル数が 100 未満と多いとは言えないことから，古典的テスト理論に基づいた数値をもとに議論をすることとする。

[4] 多くの先行研究によると，項目弁別力がマイナスの値を取る場合，項目の削除あるいは改善が必要であるという見解で一致している。ただし，一部の研究では 0.19 以下の項目を不良項目として削除あるいは改訂が必要なものとしている(Ebel 1979)。この基準に照らせば，本テストでは「説明」(3 級・意味の類似)の 1 項目は不良項目となる。

[5] G-P 分析は，項目弁別力の分析と同義として論じている先行研究もあるが，本稿では正答率のみならず，項目の錯乱肢を含む各選択肢の選択率に着目することから，TDAP で指標の 1 つとして算出される項目弁別力とは別のものとして扱う。また，G-P 分析では上位群と下位群の選択率について *t* 検定を用いて統計的に比較する方法も取られることがあるが，本研究の調査対象者の上位群と下位群は両グループ 15 名とサンプル数が少ないため，統計的な処理は行わなかった。