

国立国語研究所学術情報リポジトリ

簡便な日本語能力測定を目的とした新たなテストの
開発：

「かんたん日本語テスト」の予備的实施結果の報告

メタデータ	言語: 出版者: 国立国語研究所 公開日: 2024-01-26 キーワード (Ja): キーワード (En): 作成者: 吉田, 暁, 市江, 愛 メールアドレス: 所属:
URL	https://doi.org/10.15084/0002000154

簡便な日本語能力測定を目的とした新たなテストの開発

——「かんたん日本語テスト」の予備的实施結果の報告——

吉田 暁^a 市江 愛^b

^a 城西大学／国立国語研究所 共同研究員

^b 東京都立大学／国立国語研究所 共同研究員

要旨

本研究では、研究者が自由に利用・拡張可能な low-stakes test である「かんたん日本語テスト」の新たな版を作成し、予備的实施による性能検証を行なった。本テストは、語彙・文法の2つに分かれており、大学4年間に亘る実施を想定してそれぞれ4版からなる。予備調査に参加した113名の解答を用いて信頼性と妥当性の検証を行なった結果、全体的に良好な数値が得られた。さらに、漢字圏・非漢字圏参加者の別で行なった解析結果も概ね良好だった。本テストが日本語教育研究に十分な性能を備えていることが示唆され、今後の活用を期待させる結果が得られた*。

キーワード：日本語能力テスト、語彙テスト、文法テスト、信頼性、妥当性

1. はじめに

言語能力テストは、言語に関わる知識・技能・能力を推定するための方法である (Green 2013)。近年、我が国では、多様化が進む日本語学習者のために様々な日本語能力テストが開発されてきた。このうち、学習者の公的な認定・評価に用いられる大規模試験 (high-stakes test) としては、日本語能力試験・日本留学試験・BJT ビジネス日本語能力テストなど、数多くが挙げられる。しかしながら、ある教材の効果を検証する、学習者の特性ごとに学習プロセスの違いを明らかにするといった教育研究にふさわしい利便性を備えたテスト (low-stakes test) はそれほど多くない。

日本語教育研究における low-stakes test に求められる要件は、研究課題によって様々であるが、例えば、(a) 実施日時・実施地の制約を受けにくいこと、(b) 受験費用が発生しないこと、(c) 研究者が自由にテスト項目を取捨選択し、解析できることが挙げられる。まず、ほとんどの high-stakes test が年数回程度、指定会場で実施されるため、研究利用には限界があるだけでなく、学習者の負担も大きい。したがって、時間的・空間的な制約を受けにくい low-stakes test の需要が想定される (a)。同様に、high-stakes test にかかる受験費用を考慮すると、特に日本語教育研

* 本研究は、JSPS 科研費 JP21H04417 および国立国語研究所の共同研究プロジェクト「多様な言語資源に基づく日本語非母語話者の言語運用の応用的研究」(プロジェクトリーダー：石黒圭) の研究成果である。本稿は 2022 年度日本語教育学会秋季大会での発表 (吉田・市江・石黒 2022) をもとに加筆修正したものである。本研究にご協力を頂きました大阪大学、鳥根大学、一橋大学、埼玉大学、ドンア大学 (ベトナム)、フエ大学外国語大学 (ベトナム)、韓国外国語大学校、東呉大学 (台湾)、マカオ大学の関係者の皆様に心より感謝申し上げます。

究に携わる学生は、経済的理由から調査に用いることが難しい。よって、研究利用のためのテストは無償で実施できることが望ましい (b)。さらに、公的な認定試験では一部の例外を除いて問題項目が非公開であり¹、研究者が自身の関心に即して項目を取捨選択する、あるいは一部を改変するといった工夫は不可能であった。例えば、心理学・社会学領域においては、それぞれの研究者が関心のある構成概念の測定尺度を（短縮版・超短縮版尺度も含めて）開発し、相互に利用し合う研究文化がある（小塩 2015）。自由に項目を閲覧・改変・解析できるような日本語テストがあれば、日本語教育研究においても、より多様な研究が可能になるであろう (c)。

上記の (a)～(c) の問題のうち、少なくとも (a) と (b) を克服した数少ない例としては、SPOT (Simple Performance-Oriented Test) が挙げられる（小林・山元 1996, 小林 2005）。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 文字を挿入する形式のテストであり、クラスプレースメント等で幅広く利用されている（李・小林・今井 2015）。もっとも、SPOT は日本語の即時的処理、すなわち言語運用能力を測定する Performance-Oriented のテストであるため（小林・山元 1996）、例えば漢字や語彙の知識量に関心がある研究者にとっては、自身の研究で検討したい日本語能力（知識量）と SPOT で測定される日本語能力（運用能力）との間に、構成概念上のギャップが生じることがある。さらにいえば、SPOT はすでに受験用の Web システムが構築されているため、先述の (c) の観点から、応用的な研究利用が難しい場合がある。そのほかに、研究者が利用可能なテストとしては、J-CAT (Japanese Computerized Adaptive Test) が挙げられる（今井他 2009）。J-CAT は項目反応理論に基づいた適応型テストであるため、受験者の能力を随時推定しながら最適な項目を出題でき、さらに、ワンスケールの等化済得点による幅広い能力評価が可能である。しかしながら、J-CAT は有料の非公開テストであるため、少なくとも (b) や (c) の観点から、研究利用上の制約が生じる。このように、資格取得や公的認定に関わらない low-stakes test として SPOT や J-CAT などの候補はあるが、(a)～(c) のような研究利用目的の要望を全て満たす汎用性の高いテストは未だ開発されていない。

そこで、市江・吉田・石黒 (2023) では、日本語教育研究のための利便性を備えた学習者レベル判定用簡易日本語テストとして「かんたん日本語テスト（英語名：Simple Japanese Test）」を開発した。それをもとに、本研究では新たな版を作成し、その信頼性と妥当性を検証する。なお、本研究は、4 年間に亘るテスト開発・性能検証調査における、予備的实施結果（初年度分）の報告であることを付記しておく。

2. テスト作成

本研究では、市江・吉田・石黒 (2023) と同様に、時間的・空間的な制約を受けにくい low-stakes test の要件を満たすため、computer-based test (CBT) を作成した。CBT は、従来の紙筆テストとの同等性が十分に検証されており、その実用性の高さから近年急速に普及したテスト形式である（Choi et al. 2003, Ockey 2009）。

¹ 日本語能力試験については、実際に出題した問題の中から『日本語能力試験公式問題集』が発刊されている（国際交流基金・日本国際教育支援協会編 2012; 2018）。

テストは、語彙・文法からなる四肢択一式の CBT（各制限時間 30 分）を作成した。幅広い日本語能力の測定を目指し、初級から上級の水準の項目が網羅できるよう、次のように作問を行なった。まず、語彙テストは、「日本語教育語彙表」（<https://jreadability.net/jev/>）をもとに、初級前半・後半、中級前半、中級後半、上級の 4 レベル（各 15 項目）からなる 60 項目を作成した。語種（和語・漢語・混種語・外来語）と品詞（名詞・動詞・形容詞・副詞／オノマトペ）の項目数に偏りがないうように配慮して作題した。文法テストは、国際交流基金・日本国際教育支援協会（編著）（2002）『日本語能力試験 出題基準 [改訂版]』をもとに、旧日本語能力試験の 4 級、3 級、2 級、1 級の 4 レベル（各 15 項目）からなる 60 項目を作成した。また、日本語記述文法研究会編（2003-2010）の区分を参考に、活用・文型、助詞、文末形式、テンス・アスペクト、複文（時間・状況、原因・理由、条件・目的、等位・並列、名詞修飾・引用）に分類して作題した。語彙、文法のいずれにおいても、出題順はレベル順とし、終盤に進むほど難易度が高くなるように項目を配置した。なお、これら問題構成は、市江・吉田・石黒（2023）と同様である。

3. 版の作成

本テストは、4 年間に亘る継続的な実施を想定し、市江・吉田・石黒（2023）で作成した第 1 版に加え、新たに 3 つの版を作成した（図 1）。まず、予め作成した第 1 版のうち、項目番号が奇数の項目を A、偶数の項目を B に分けた。そして、第 1 版（AB）を元に、A の各項目と同程度の難易度を有する新項目 A' を作題し、第 2 版（A'B）とした。すなわち、第 2 版は第 1 版から半数の項目が再出題され、残り半数が新規作題項目となっていた。一般的に、再出題項目を用いたテストでは、能力と関連しない学習効果（馴れ・記憶による正答）が生じる懸念があるが、このように再出題項目と新規項目を併用・比較することで、測定の信頼性を保ちながら作題コストを低減できると考えた。また、新規作題項目は、オリジナル項目と難易度を同等にするため、語彙項目では、ターゲット語のコーパス出現頻度がオリジナル項目のそれと同程度になるよう調整し、文法項目では、オリジナル項目の選択肢を改変して作題した。同様の手順で、第 1 版（AB）の B を改変した B' を作成し、第 3 版（A'B'）とし、最後に、新規項目のみで構成された版を第 4 版（A'B'）とした。

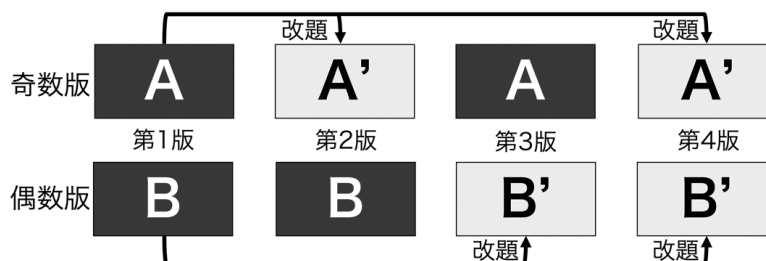


図 1 かんたん日本語テストの版作成プロセス

4. 予備的实施

4.1 実施方法

妥当性と信頼性の検証のため、2022年5-6月、本テストとSPOT90を日本・中国・台湾・韓国・ベトナムの大学機関で実施した。参加者は、4版のうち1版をランダムに割り当てられ、学習管理システム Moodle を利用して受験した。

4.2 予備的实施結果の検証方法

本テストの性能を検証するために、1) 内的一貫法と折半法を用いた信頼性検証、2) SPOT90や学習歴を用いた妥当性検証、3) 4版の正答率の同等性検証を行った。これに加えて、本研究では漢字圏学習者と非漢字圏学習者とが異なる母集団である場合を想定し、1) ~ 3) を漢字圏・非漢字圏の別でも実施した。全ての解析には R (Version 4.1.2) を用いた。

4.2.1 信頼性の検証

本テストの信頼性（同一個人に実施された際の得点の一貫性）を検証するため、岡田（2015）と高本・服部（2015）を参考に、内的一貫法と折半法の2つの方法で係数を算出した。まず、内的一貫法によって Cronbach（1951）の信頼性係数 α を以下の式（1）によって推定した。 n は項目数を、 σ_x^2 はテスト得点の分散を、 $\sigma_{x_i}^2$ は各項目の分散を、 \bar{r} は項目間の平均相関係数を示す。

$$\alpha = \left(\frac{n}{n-1} \right) \left(\frac{\sigma_x^2 - \sum \sigma_{x_i}^2}{\sigma_x^2} \right) = \frac{n\bar{r}}{1 + (n-1)\bar{r}} \quad (1)$$

次に、Spearman-Brown の式（2）を用いて、折半法による信頼性係数 ρ を推定した。折半法とは、1つのテストを無作為に分割した2つの部分テスト同士の相関係数 $\rho_{x_1x_2}$ を用いて信頼性係数を推定する方法である（式（1）で、 $n=2$ とした場合に相当）。本研究では、奇数版と偶数版（図1）を部分テストとした。ただし、本テストでは奇数・偶数版に交互に改題を行っており、奇数と偶数の別が必ずしも無作為の分割を意味しないため、折半法の本来的な用途とは異なる。本研究では、改題による信頼性係数への影響を観察する目的で折半法を用いた²。

$$\rho_{SB} = \frac{2\rho_{x_1x_2}}{1 + \rho_{x_1x_2}} \quad (2)$$

4.2.2 基準関連妥当性および学習歴との関連の検証

本テストが日本語能力を安定的に測定できること、すなわち妥当性を検証するため、SPOT90を基準とした基準関連妥当性を算出した。先述のように、SPOTは日本語の運用能力を測定するテストであり、Web版のSPOT90は30問ずつの「SPOT90-1（日本語能力試験 N5-N4 相当）」

² 例えば、図1の第1版（AB）と第2版（A'B）の信頼性係数 ρ を比較し、後者に著しい係数の低下が観察された場合、A から A' への改題の影響が推察される。反対に、両版の信頼性係数が同程度であれば、改題による信頼性への影響が小さいことが確認できると考えた。

「SPOT90-2 (N4-N2 相当)」「SPOT90-3 (N2-N1 相当)」で構成される (李・小林・今井 2015)。本テストの目標レベルと概ね合致しているため、外部基準として適切と考えて検証に用いた。

そのほかに、参考のため、参加者の学習歴ごとに本テストの平均得点が増加するかを分析した。本研究は一時点の横断研究であるため、参加者の学習が進むにつれて本テスト得点がどのように推移するかを直接検討することができない。そこで、参加者の学習歴が長いほど (学習が進むほど) 本テスト正答率の正答率が増加するかを確認するための相関分析を行った。なお、正規分布に近似させるため、正答率はロジット変換を行なったうえで解析に投入した (島田・野口 2021)。

4.2.3 4 版の同等性の検証

最後に、作成した 4 版の正答率に偏りがなかったことを検証した。古典的テスト理論に基づいたテストでは、実施回ごとにテストの困難度が変化するため、回間の得点比較が困難である。他方、現代的な項目反応理論に基づくテストではテスト間の得点等化が可能であるが、共通項目法なし共通受験者法を用いた等化係数の推定作業が必要になる (豊田 2002)。古典的テスト理論に基づく本テストでは、部分的な改変によって複数版を作成しており、簡易的な方法ではあるが、版ごとの困難度を概ね一定に統制できることが期待できる。そこで、一元配置分散分析および Tukey の多重比較法によって版ごとの平均正答率を比較し、同等性の検証を行なった。

5. 結果と考察

本調査には大学生 131 名が参加し、欠損データを除いた最終的な分析対象は 113 名だった。分析対象者の実施地は 46.9% が台湾、15.0% がベトナム、15.9% が韓国、14.2% が日本国内、8.0% が中国だった。第 1-4 版はそれぞれ 29 名、28 名、25 名、31 名が受験した。参加者の母語割合は、漢字圏が 66.4% (中国語 72 名、広東語 3 名) であり、非漢字圏が 33.6% (ベトナム語 17 名、韓国語 16 名、ドイツ語 2 名、英語・スペイン語・ポルトガル語それぞれ 1 名) だった。参加者に関する詳細を表 1 に示す。また、参加者全体の平均日本語学習歴は 37.6 ヶ月だった。漢字圏受験者と非漢字圏受験者の平均学習歴を Welch の t 検定によって比較したところ、有意差は認められなかったため、母語による学習歴の偏りがなかったものと考えられる。

表 1 予備的实施参加者の内訳

母語	版	人数	割合 (%)	平均学習歴 (月)
漢字圏	第 1 版 (A'B)	19	25.3	40.0
	第 2 版 (A'B)	19	25.3	34.1
	第 3 版 (A'B')	19	25.3	26.3
	第 4 版 (A'B')	18	24.0	36.8
非漢字圏	第 1 版 (A'B)	10	26.3	37.0
	第 2 版 (A'B)	9	23.7	34.7
	第 3 版 (A'B')	6	15.8	71.3
	第 4 版 (A'B')	13	34.2	44.1

5.1 信頼性の検証結果

まず、全般的な信頼性係数 α を推定した結果、表2の結果が得られた。参加者全体では、語彙 ($\alpha =$ 最小 .91–最大 .95)、文法 (.90–.96) の両方で α が.90以上と極めて高い値を示した。さらに、漢字圏受験者(語彙 (.83–.95)、文法 (.88–.95))と非漢字圏受験者(語彙 (.94–.97)、文法 (.93–.97))のいずれにおいても α が.80を上回った。これらのことから、本テストが一定の信頼性を備えていることが示唆された。

表2 Cronbach の α 係数一覧

版	語彙			文法		
	全体	漢字圏	非漢字圏	全体	漢字圏	非漢字圏
第1版 (AB)	.93	.83	.96	.92	.89	.93
第2版 (A'B)	.93	.91	.94	.90	.88	.94
第3版 (AB')	.95	.95	.96	.96	.95	.97
第4版 (A'B')	.91	.83	.97	.93	.89	.97

次に、奇数項目 (A・A') と偶数項目 (B・B') との折半法によって信頼性係数 ρ を推定した結果、図2の結果が得られた。参加者全体では、語彙 ($\rho =$ 最小 .94–最大 .95)、文法 (.94–.97) の両方で ρ が.90を超えた。さらに、漢字圏受験者(語彙 (.81–.95)、文法 (.91–.96))と非漢字圏受験者(語彙 (.94–.98)、文法 (.92–.99))のいずれでも.80以上の高い値が示された。語彙・文法のいずれにおいても、4版の信頼性係数 ρ がほとんど同じ値を示したことから、改題による信頼性への影響はそれほど大きくなかったと考えられる。

5.2 基準関連妥当性および学習歴との関連の検証結果

本テストの基準関連妥当性を検証するため、各版の得点と SPOT90 合計得点との相関係数 r を算出したところ、語彙 ($r =$ 最小 .80–最大 .94)・文法 (.84–.93) とともに高値を示した(図3)。母語別に推定を行なった結果、漢字圏受験者では語彙 (.56–.94)、文法 (.81–.93)、非漢字圏受験者では語彙 (.79–.99)、文法 (.85–.97)であった。このうち、 r が他よりも低く示された(.56)漢字圏受験者の語彙 AB 版では、受験者の正答率がプロット右上に集中する高止まりの傾向が観察された(図3)。当該版を割り当てられた漢字圏受験者の平均学習歴が他の群より長かったことから推察されるように(表1)、この標本の能力分布が他より比較的高かったために、緩やかな天井効果が生じた結果、相関係数がそれほど高く示されなかったと考えられる。この版を除けば、全体的に.80以上の高い相関が本テストと SPOT90 との間で得られており、両テストが非常に良く対応していることが示された。したがって、本テストが、日本語能力を十分に反映しており、一定以上の妥当性を有していることが示唆された。

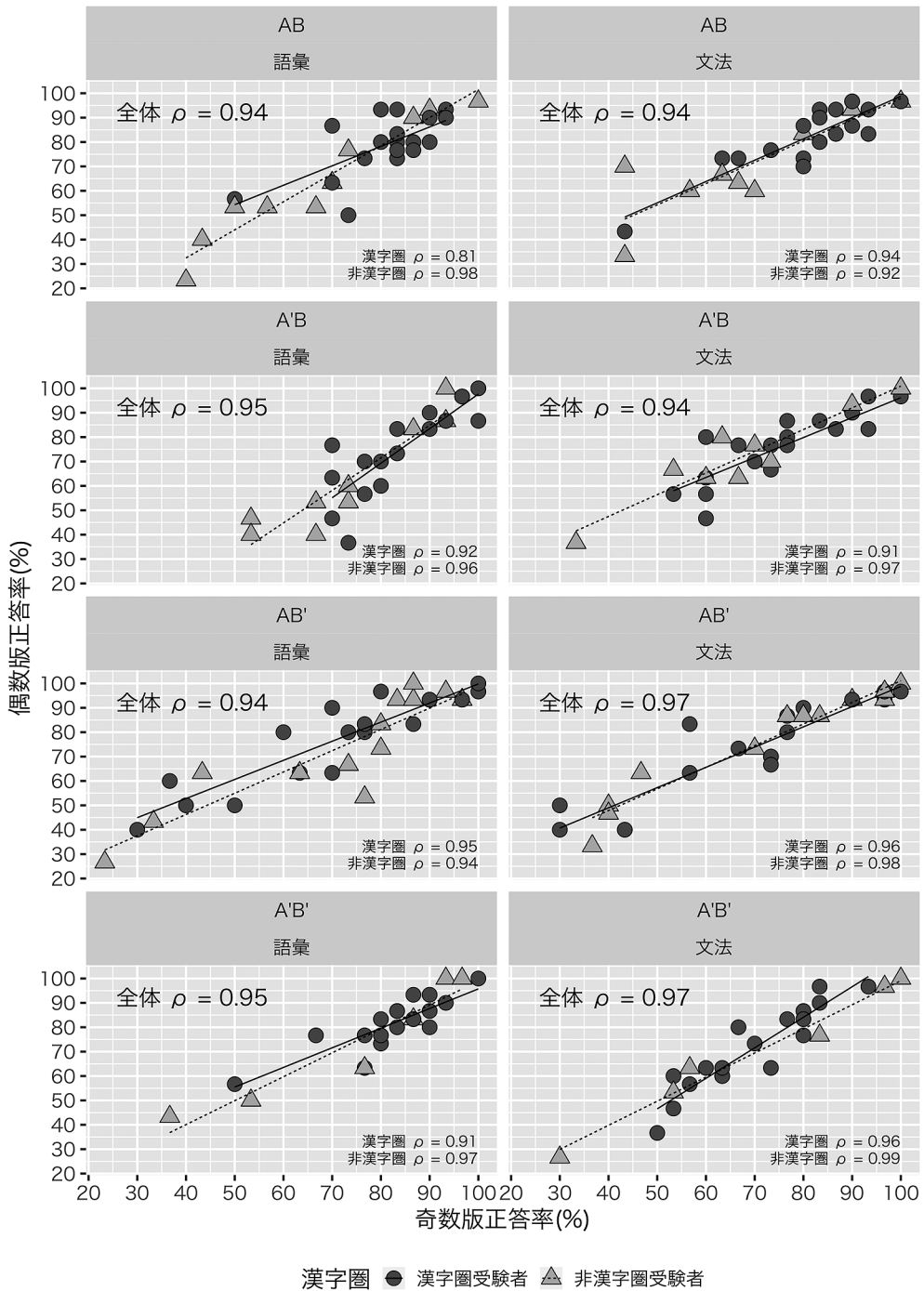


図2 折半法によるテスト信頼性の検証結果
(直線・点線は回帰直線)

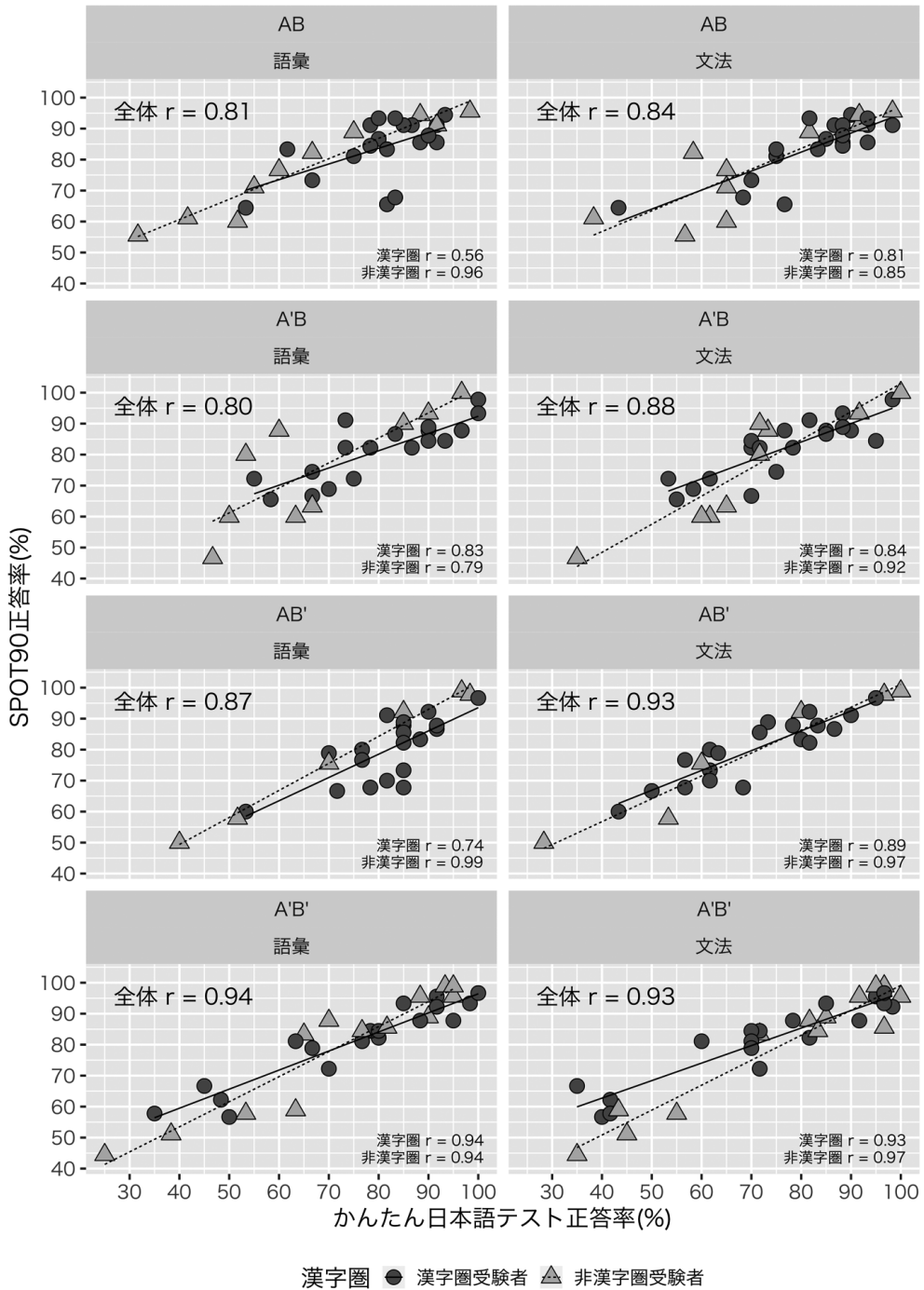


図3 SPOT90 との相関分析による基準関連妥当性の検証結果
(直線・点線は回帰直線)

次に、本テストの正答率と学習歴との関連を検証した（図4）。相関分析の結果、語彙では $r=.33$ （漢字圏 .39, 非漢字圏 .44）、文法では .39（漢字圏 .44, 非漢字圏 .38）と、弱い正の相関が示された。すなわち、本テスト得点は参加者の学習歴と緩やかな関連を示した。ただし、本調査で聴取した学習歴（期間）は実際の累積学習時間とは異なる可能性があり、さらに言えば、一般的な学習プロセスを考慮すると、成績と学習時間との間には、曲線的な関係を想定する方が妥当かもしれない。今後の調査では、同一受験者の学習経過に伴う本テストの時間的推移について、縦断的検討によって明らかにする必要がある。

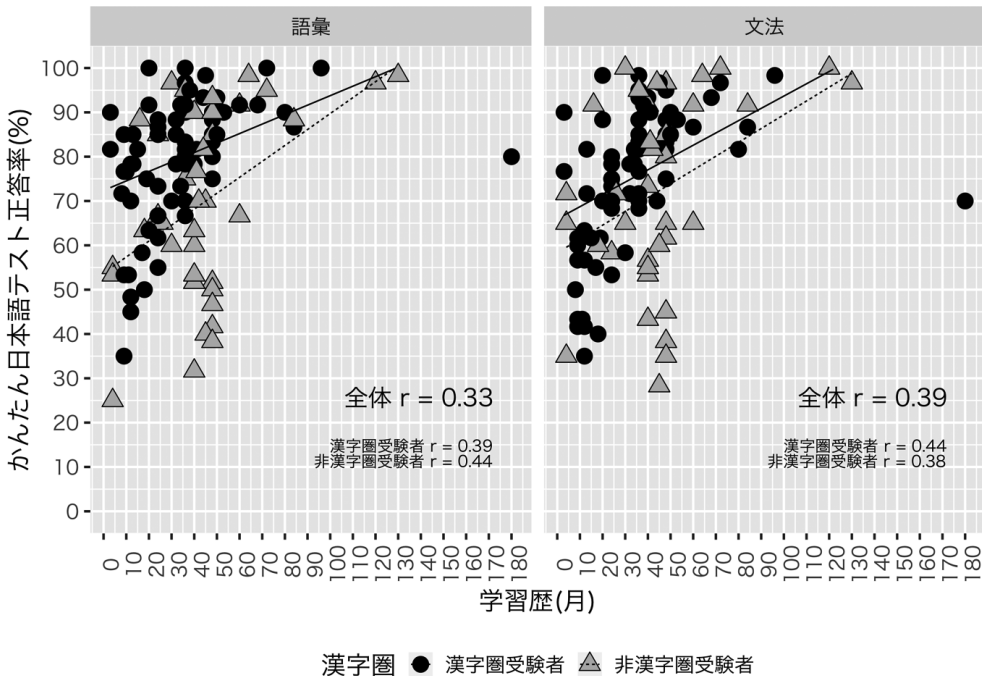


図4 学習歴グループごとの本テスト正答率
(直線・点線は回帰直線)

5.3 同定性検証のための4版の正答率の多重比較

最後に、4版の正答率が一定に保たれていることを検証するため、4版の平均正答率の多重比較を行なった。分析の結果、いずれの版の組み合わせにおいても統計学的有意差は認められなかった。したがって、各版の難易度は同程度であることが示唆された。

各版・各版の正答率の散布図を図5に示した。図5からも、各版の正答率には大きな違いは見られなかった。さらに、各版の正答率は約20%から100%の範囲で広く分布しており、天井効果もほとんど観察されなかった。この結果は、本テストが幅広い能力の測定に有用である可能性を示唆しており、今後の活用を期待させるものである。

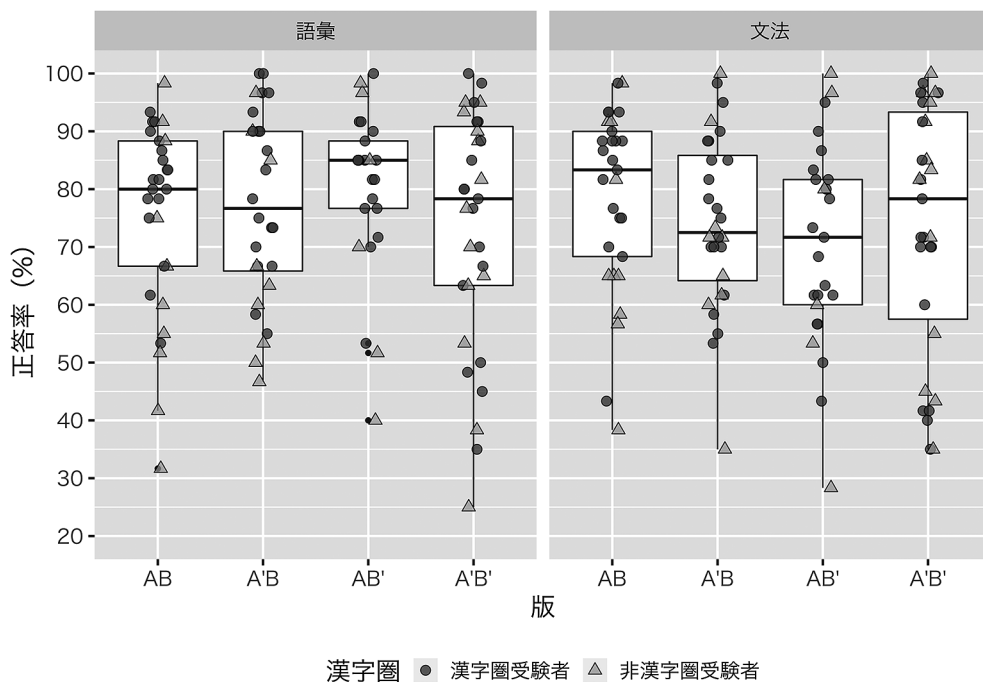


図5 各版の正答率の分布

6. おわりに

本研究では、研究者が自由に利用・拡張できる「かんたん日本語テスト」の新たな版を開発し、その妥当性と信頼性を検証した。検証結果はいずれも良好であり、さらに、作成した4版の難易度に関する同等性も示唆された。

研究の限界として、以下の3点が挙げられる。まず、本研究は一時点での横断研究であり、学習に伴うテスト得点の推移については今後の縦断的な調査で明らかにする必要がある。ただし、学習歴を用いた相関分析の結果から、本テストが学習期間と一定の関連を有することは既に示唆されている。次に、本テストの4版は互いに半数程度の項目を共有し合っているため、複数回受験した場合、正答が受験者の記憶に残る可能性も否定できない。再出題項目の正答率についても、今後の調査で検討する必要がある。最後に、low-stakes testの性質として、本テストを受験した参加者の動機づけは、公的なhigh-stakes testの受験時ほど高くなかった可能性がある。今後の調査では、参加者の動機づけも検討指標に組み入れる必要があるだろう。

本テストの特色は、簡便に実施可能なlow-stakes testであるだけでなく、研究目的に沿った拡張性を有する点である。本研究で用いたように、テスト項目の半数ずつを作題（改変）する手法によって、作題資源の限られた小規模の研究グループにおいても、容易に版を追加作成することが可能である。今後は4版を1年ごとに実施し、測定の実験性を縦断的に検討する予定である。

参考文献

- Choi, Inn-Chull, Kyoung Sung Kim and Jaeyool Boo (2003) Comparability of a paper-based language test and a computer-based language test. *Language Testing* 20: 295-320.
- Cronbach, Lee J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3): 297-334.
- Green, Anthony (2013) *Exploring language assessment and testing: Language in action*. London: Routledge Taylor & Francis Group.
- 市江愛・吉田暁・石黒圭 (2023) 「日本語教育研究のための『かんたん日本語テスト』の開発—テスト開発経緯と項目分析結果を中心に—」『国際学報』1: 19-27.
- 今井新悟・伊東祐郎・中村洋一・菊地賢一・赤木彌生・中園博美・本田明子・平村健勝 (2009) 「項目応答理論に基づくテストの得点—J-CAT の得点換算・解釈・利用法について—」『大学教育』6: 93-106.
- 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論集』20: 67-82.
- 小林典子・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.
- 国際交流基金・日本国際教育支援協会 (編著) (2002) 『日本語能力試験出題基準 改訂版』東京: 凡人社.
- 国際交流基金・日本国際教育支援協会 (編) (2012) 『日本語能力試験公式問題集 (N1 ~ N5)』東京: 凡人社.
- 国際交流基金・日本国際教育支援協会 (編) (2018) 『日本語能力試験公式問題集 第二集 (N1 ~ N5)』東京: 凡人社.
- 李在鎬・小林典子・今井新悟 (2015) 「テスト分析に基づく「SPOT」と「J-CAT」の比較」『第二言語としての日本語の習得研究』18: 53-69.
- 日本語記述文法研究会 (編) (2003-2010) 『現代日本語文法 全7巻』東京: くろしお出版.
- Ockey, Gary J. (2009) Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal* 93: 836-47.
- 岡田謙介 (2015) 「心理学と心理測定における信頼性について」『教育心理学年報』54: 71-83.
- 小塩真司 (2015) 「心理テストは信用できるのか」『心理学ワールド』68: 13-16.
- 島田めぐみ・野口裕之 (2021) 『統計で転ばぬ先の杖』東京: ひつじ書房.
- 高本真寛・服部環 (2015) 「国内の心理尺度作成論文における信頼性係数の利用動向」『心理学評論』58(2): 220-235.
- 豊田秀樹 (2002) 『項目反応理論 [入門編]』東京: 朝倉書店.
- 吉田暁・市江愛・石黒圭 (2022) 「日本語教育研究のための「かんたん日本語テスト」の開発—信頼性と妥当性の検証—」『2022 年度日本語教育学会秋季大会予稿集』246-249.

関連 Web サイト

- 国立国語研究所 「現代日本語書き言葉均衡コーパス」 <https://clrd.ninjal.ac.jp/bccwj/> (2023 年 3 月 1 日確認)
- 国立国語研究所 コーパス検索アプリケーション『中納言』 <https://chunagon.ninjal.ac.jp/> (2023 年 3 月 1 日確認)
- 筑波大学 『筑波日本語テスト集 | TTBJ』 <https://ttbj.cegloc.tsukuba.ac.jp/p2.html> (2023 年 3 月 1 日確認)

Developing a New Japanese Language Proficiency Test in a Simplified Method: Report on Preliminary Implementation of the “Simple Japanese Test”

YOSHIDA Akira^a ICHIE Ai^b

^aJosai University / Project Collaborator, NINJAL

^bTokyo Metropolitan University / Project Collaborator, NINJAL

Abstract

This study developed a new version of the “Simple Japanese Test,” a low-stakes test that can be freely used and extended by researchers. The study verified its performance through preliminary implementation. The test includes vocabulary and grammar sections, and has four versions covering four years of university study. Reliability and validity were assessed using the answers of 113 preliminary survey respondents, which indicated overall promising results. Furthermore, the results for Kanji-speaking and non-Kanji-speaking participants were favorable. Therefore, this test is adequate for Japanese language educational research and promising for future use.

Keywords: Japanese language proficiency tests, vocabulary tests, grammar tests, reliability, validity