J-CAT(Japanese computerized adaptive test)の得点と Can-do スコアの関連付け

今井　新悟
山口大学

要旨

本研究では受験者自身の can-do 記述の判定とテストの得点との関連付けを行い、テスト得点が意味する能力を具体的・明示的に示すための方法を示す。本研究の対象とするテストは J-CAT（Japanese computerized adaptive test）である。J-CAT は項目応答理論により項目が等化されて得点に不変性があるアダプティブテストである。J-CAT 受験者に対して can-do 記述文のアンケートを 7 件法で実施した。J-CAT 得点と can-do 記述文ごとのスコアの相関に基づき、関連づけを行う can-do 項目を抽出した。次に単回帰分析を行い，can-do スコアに相当する J-CAT 得点を求めて分割点とした。これにより、J-CAT 得点と can-do スコアの関連づけが行われ，受験者の能力を一般の利用者にも理解されやすい can-do 記述文を利用して具体的・明示的に示すことが可能となった。また、J-CAT の得点は受験者集団に依存しないため、受験者間比較および同受験者間での縦断的な比較も可能である。

【キーワード】　J-CAT、アダプティブテスト、日本語テスト、Can-do

1　J-CAT（Japanese-computerized adaptive test）の概要

はじめに、本研究の対象となった、J-CAT について見ておきたい。

Japanese Computerized Adaptive Test（略称 J-CAT）は WEB 上でいつでもどこからでもアクセスでき、日本語学習者の日本語能力がリアルタイムで測定できるテストである。文字語彙、文法、読解、聴解の 4 セクションからなり、受験時間は概ね 60 分から 90 分である。成績がテスト終了と同時に画面上に表示され、成績証が PDF 形式で自動で作成され、印刷やダウンロードが可能である。個人登録方式と団体受験方式が用意されている。前者では、WEB 上から必要な情報を入力して登録すれば、パスワードが発行されるので、そのパスワードでログインして、受験できる。後者では、試験実施者用にパスワードが発行され、そのパスワードで受験者は受験できる。前者は各自が自分の日本語能力の伸びなどを確認するのに適しており、後者では、大学等のプレースメントテストなどでの一斉利用が可能である。また、団体受験の場合には、受験者全員の成績が一覧表となって、試験実施者に送付される。以上のように時間と場所の制約を受けずに日本語の能力を測定できるテストになっている。

J-CAT はアダプティブテスト（適応型テスト）になっている。これは受験者の解答のでき、不出来によって出題される問題が変化するものである。受験者の能力に従い、難易度の異なる問題を出題することで、効率的に能力測定を行う。よって、従来の紙と鉛筆によるテストに比べて、短時間でより高い精度での能力測定が可能である。

アダプティブテストの仕組みは視力検査のメタファーによって理解されやすい。視力検査では、まず適当な大きさの環（ランドルト環）や文字を指し、それが見えたら、より小さい環や文字を指すし、一方見えなかったら、より大きい環や文字を指し示す。

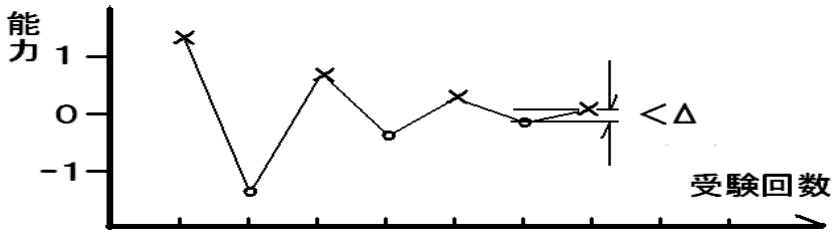

図１　能力推定のイメージ

このようにして、視力検査では、かろうじて見える大きさのランドルト環を探る。そして、そのランドルト環に付いた 1.0 や 1.2 が視力となる。ランドルト環はアダプティブテストの問題項目に相当し、かろうじて見えるというのは、50%の確率で正答できることに相当する。

ランドルト環の持つ数値は、テストの各項目に付けられた困難度パラメータに相当する。ただし、視力検査では、判別できる最小のランドルト環の数値が視力となるが、アダプティブテストでは、受験者の能力値を仮定し、そのとき予想される正誤の解答パターンが、実際の正誤のパターンに近づくように仮定の能力値を調整しながら、推定能力値を探る。テストの始まりでは、受験者の本来の能力がまったく見当が付かないことから、受験者の能力レベルと出題される問題の困難度のレベルが合っていないが、テストが進むにつれて、だんだん、能力の推定ができてきて、測定誤差が一定範囲内に収まるようになったときにテストは終了し、受験者の推定能力値が与えられる。

J-CAT は他にも従来の日本語テストでは実現できなかった以下のように機能を持っている。

・テストの標準化を図り、絶対評価をし、すべての能力を一つのスケールで測定する。
・日本語能力試験ではカバーしきれない範囲までカバーする。旧日本語能力試験で言えば１級以上、新日本語能力試験でいえばＮ１以上の能力の判定も行う。
・すべての問題を画像として画面表示することにより，オペレーションシステムの種類や言語に影響されず、ルビなどの特殊表記にも対応する。
・画面のコピー・ペースト機能を制御して問題の流出を防ぐ。
・音声のみならず、カラーのイラスト、写真、動画形式の問題もあり、真正性を高めている。

２　J-CAT の得点について

J-CATでは，項目応答理論を用いて，各問題（アイテム）の項目困難度（難易度）と項目識別力の値を算出している。困難度と識別力の値が付加された問題（アイテム）が，アイテムプールに収納されている。項目応答理論により算出された困難度と識別力は受験者集団やアイテムの難易度に左右されない不変的な指標となる。そしてそこから推定されてくる能力値もまた，受験者集団やアイテムの困難度に左右されない値である。

パラメータの尺度が等間隔になっていることは保証されることから，たとえ受験者母集団の能力の平均・分散が未知であっても，能力推定値そしてJ-CATの結果として示される得点の比較は常に可能である。古典的テスト理論においても点数は間隔尺度であるが，天井効果や床面効果を考慮すると，その尺度は能力の違いを適切に等間隔で反映していると

はいい難い。一方，得点の天井に近い者にはボーナス得点を加算し，得点の床面に近いものにはペナルティとして得点を差し引くような，別の言い方をすれば得点にウェートをかけるような操作をしているのが項目応答理論での能力推定値である。このように，異なる受験者集団間での得点の比較が可能となることから，「項目応答理論の推定値が受験者集団の影響を受けない」と言える。

能力値は，項目応答理論で算出され，原理的には0を中心とした、正負無限大までであることになるが、J-CATにおいては，概ね - 3.5から+3.5の範囲となることがこれまでのデータから分かっている。しかし，この値のままでは，一般の受験者およびテスト利用者はその意味を理解することはできないので,一般になじみのある100点満点の値に変換することが望ましいと考えた。項目応答理論で求められる能力値は0を原点とする比例尺度である。よって，それを一次変換しても間隔は変らないので，偏差値と同じようにして得点換算ができる。J-CATでは，以下の換算式を用いて得点を算出している。

得点=最終能力値×15＋50

これで，聴解，語彙，文法，読解の４セクションの得点が，概ね0点から100点の間に入る。そして総得点は0点から400点の間に入る。

この式内の１５の係数は任意のものであるが，それが小さすぎると換算される得点の幅が狭く，得点に差がつかず，出題されたアイテムに正答を繰り返しても高得点が取れないということになる。また，係数が大きすぎると，設定したい得点幅（ここでは，0点から100点）を超えてしまう例が多く出てくることになり，不都合が生じる。そのような不都合が最も少なくなるような係数を探るため，困難度と識別力の実データを用いてのシミュレーションを行い、その結果から導きだされたものである。（詳しくは今井他2009を参照）

3　先行研究

本研究のテーマの先行研究としては、Alderson (2005) 、三枝（2004）、島田他（2006）があるのでそれぞれについて以下で概観する。

Alderson (2005) では DIALANG の英語のテストスコアとスキル領域ごとに 18 の can-do 記述のスコアとの相関を調べている。なお、can-do のスコアは項目応答理論により産出している。その結果、各スキル領域の can-do スコアとそれに対応するスキルのテストスコアとの相関はそれぞれ、Reading 領域で 0.487、Writing 領域で 0.550、Listening 領域で 0.495 となっており、中程度の相関が認められる。なお、DIALANG でいう Writing テストとは、自由記述ではなく、穴埋め問題である。

三枝（2004）は日本語のテストを扱い、can-do 記述と日本語能力試験および大学で使用したプレースメントテストとの相関を調べている。can-do スコアは 1 から 7 までの 7 段階で回答してもらったものを素点として 1 点から 7 点として使っている。日本語能力試験の各セクションと can-do 記述スコアの各領域との相関は以下の通り（三枝 2004: 32）である。

表 1　日本語能力試験 1 級（2001 年）における相関

Can-do / JLPT	読む	書く	話す	聞く	総点
文字語彙	0.432	0.300	0.254	0.319	0.354

聴解	0.318	0.259	0.261	0.350	0.323
読解文法	0.347	0.279	0.222	0.302	0.312
総点	0.400	0.312	0.267	0.356	0.363

表 2　日本語能力試験 2 級（2001 年）における相関

Can-do / JLPT	読む	書く	話す	聞く	総点
文字語彙	0.382	0.093	-0.018	0.052	0.144
聴解	0.275	0.143	0.154	0.259	0.242
読解文法	0.387	0.134	0.044	0.074	0.183
総点	0.414	0.146	0.064	0.131	0.216

いずれの結果においても相関は低い。このことについて、三枝（2004:31）は「日本語能力試験が級別試験であり、必然的に回答者の日本語能力幅が狭く」なることに起因していると分析している。島田他（2006）では、これを「輪切り現象」と呼んで、同様の分析を行っている。

これに対して、ある大学のプレースメントテストのスコアと can-do スコアの相関は以下のように高かったことを三枝（2004:66）および島田他（2006:81）は報告している。

表 3　プレースメントテストにおける相関

Can-do / Placement	読む	書く	話す	聞く	計
聴解	0.758	0.669	0.623	0.725	0.729
語彙	0.711	0.645	0.648	0.729	0.717
文法	0.709	0.609	0.583	0.645	0.669
読み	0.755	0.672	0.588	0.667	0.706
漢字	0.823	0.775	0.705	0.759	0.805
総合	0.834	0.748	0.697	0.780	0.804

この結果について、三枝（2004）・島田他（2006）は日本語能力試験が「輪切り現象」を起こしていたのに対し、プレースメントテストでは、能力の幅が大きかったからであるとし、さらに、輪切り現象の影響を差し引けば、can-do スコアとテストスコアの相関は高く、can-do が「日本語能力を反映する尺度としての有効性が示された」と結論づけている。

確かに、can-do スコアとテストスコアの相関があることは間違いないだろう。しかし、それがどの程度の強さなのかは、次の理由から慎重にすべきである。

三枝（2004）・島田他（2006）が扱ったデータの受験者の分布は図 2 のようになっていて、正規分布からの逸脱が大きい。このような分布は、正規分布と比べて相関が高く出る。通常のテストスコア分布は正規分布となることが多いのに比べ、ここで扱われているデータ分布のようにそれが特殊である場合、その結果・解釈が歪んでしまう危険性がある。また、このデータは、床効果も相当程度予想される点にも注意が必要である。以上からプレ

ースメントテストでの強い相関とそこから導かれるcan-doスコアによる日本語能力の尺度としての有効性についての判断は慎重に行われるのが望ましい。

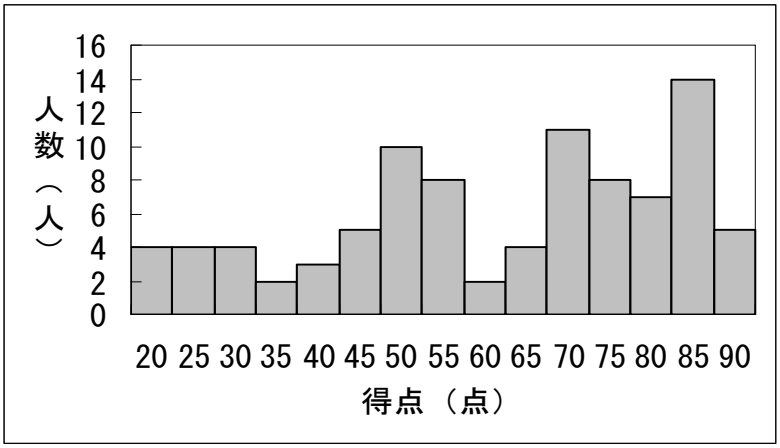

図2　あるプレースメントテスト（2003年）総合得点の分布　三枝（2004:65）から

4　方法

本研究の方法は以下の通りである。

①J-CAT受験者に対してcan-do記述文のアンケートを7件法で実施

②J-CAT得点とcan-doスコアの相関に基づき、can-do項目を抽出

③単回帰分析を行い，J-CAT得点とcan-doスコアの関連づけ

①のcan-do記述文のアンケート用紙は島田・谷部両氏のご好意により、日本語に加えて英語・中国語・ハングルの訳がそれぞれ併記された版をご提供いただいた。本研究に合わせて、一部項目を削除して使用し、「読む」に関する21項目、「書く」に関する17項目、「話す」に関する21項目、「聞く」に関する18項目について、WEB上でJ-CATを受験した者に対して、その直後にアンケート用紙で調査した。すべての項目について1から7の7段階で自己評価をしてもらった。対象者は3つの日本国内の大学でのプレースメントテストを団体受験した留学生（研究生を含む）である。

②では、まずJ-CATの得点とcan-doスコアの相関を産出した。後述するように、項目ごとに見ていくと相関の高くない項目が相当数あった。それらは本研究の目的である関連付けには適さないため、除外することとした。

③では、②ので残った項目について単回帰分析を行い、J-CATの得点からcan-doスコアの予測式を導いた。これによって、J-CATの得点とcan-doとを関連付けることを試みた。

5　結果と考察

Can-do記述アンケートで回答があった206人から無回答および重複回答が3項目以上ある者を除外し、2項目以下は欠損値として扱い、分析対象としたのは119人である。平均得点は232.7点、最低得点は63点、最高得点は366点であり、総合得点の分布は図3の通りであり、正規分布に近い。

紙幅の関係で、本稿では、「読む」に関する項目のみを考察対象とする。表4にcan-do記述文の7段階のスコアとJ-CATの4セクションの得点および総合得点との相関を示す。

「読解」以外のセクションと can-do の「読む」の項目との相関も参考のために示す。表中の網掛け部分は相関が 0.6 以上で中程度以上の相関が認められるものである。

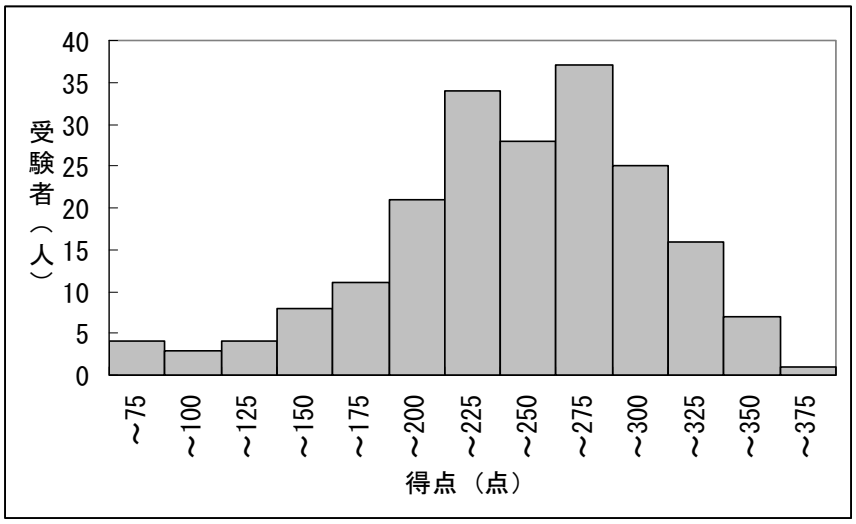

図 3　J-CAT 総合得点分布

表 4　相関と回帰式による得点

	Can-do 記述	聴解	語彙	文法	読解	総合	回帰
R1	新聞の社説を読んでわかりますか。	0.534	0.617	0.433	0.594	0.628	257
R2	学内の掲示板のお知らせ・ポスター等の印刷物を読んでわかりますか。	0.580	0.584	0.471	0.641	0.660	233
R3	学校の規則を読んでわかりますか。	0.548	0.547	0.440	0.597	0.619	232
R4	図書館の本棚にある本の背表紙を見て、必要な本を探すことができますか。	0.506	0.559	0.405	0.554	0.588	235
R5	小説を読んでわかりますか。	0.597	0.633	0.521	0.652	0.696	255
R6	駅や旅行会社においてあるちらしを読んでわかりますか。	0.557	0.539	0.429	0.587	0.613	235
R7	勉強に必要な本や論文を読んでわかりますか。	0.555	0.614	0.471	0.594	0.646	245
R8	電車やバスなどの車内の広告がわかりますか。	0.530	0.554	0.424	0.558	0.598	233
R9	病院で診察を受ける前の質問票を読んでわかりますか。	0.574	0.573	0.438	0.585	0.629	245
R10	掲示板や黒板などに手書きで書かれたものが読んでわかりますか。	0.525	0.549	0.405	0.567	0.593	238
R11	新聞の社会面（事件・事故などの記事）を読んでわかりますか。	0.509	0.597	0.460	0.583	0.621	246
R12	ガス・水道・電気の明細書をみて必要なことがわかりますか。	0.516	0.511	0.322	0.532	0.546	230
R13	パソコンや機械の使い方の説明書（マニュアル）がわかりますか。	0.504	0.542	0.361	0.521	0.558	244

R14	学校・区役所 （市役所）などからの通知（お知らせ）がわかりますか。	0.540	0.587	0.432	0.594	0.623	234
R15	就職情報（求人広告・アルバイト情報誌など）を読んでわかりますか。	0.589	0.635	0.492	0.624	0.677	235
R16	カタカナで書かれた国名、都市名が読めますか。	0.070	0.058	-0.019	0.028	0.040	*
R17	日本語で書かれた授業名がわかりますか。	0.554	0.587	0.456	0.581	0.630	223
R19	日本語のウェブページを見て、求めるページに到達できますか。	0.477	0.524	0.380	0.532	0.554	234
R20	銀行や郵便局で、窓口の標示を読んでわかりますか。	0.540	0.594	0.441	0.602	0.630	234
R21	スーパーの売り場標示を読んでわかりますか。	0.578	0.578	0.421	0.619	0.637	229
R22	大学キャンパスの案内板を読んでわかりますか。	0.538	0.579	0.406	0.577	0.608	224
平均		0.520	0.551	0.409	0.558	0.590	

＊R18 は内容が本調査にそぐわないため、調査項目から削除した。

総合得点と Can-do スコアとの相関は概ね中程度であり、「読解」得点と can-do スコアとの相関もある程度見られる。また、「語彙」得点と「読む」の can-do スコアにある程度の相関が認められるが、「読む」技能に語彙量が多分に関わるのは自然だと思われる。総合得点および「読解＝reading」のテストスコアと can-do スコアについて先行研究と比べると、三枝（2004）・島田他（2006）の日本語能力試験の場合よりははるかに相関が高く、Alderson (2005) の DIALANG の場合よりも若干高く、三枝（2004）・島田他（2006）のプレースメントテストの場合よりは低い。（プレースメントテストのデータでは相関が高く出やすいことについては前述した。）

R16 のカタカナの読みに関する項目は、相関が著しく低い。受験者の中にカタカナが読めるかどうかという低いレベルの者がいなかったためだと思われる。また、J-CAT に単純にカタカナの読みを問う問題がなかったことも影響を与えている。

Can-do スコアを説明変数、J-CAT 得点を目的変数として単回帰分析を行った。R1 を例にとると以下のような結果になった。なお、相関係数が 0.628 であったので、決定係数は 0.395 である。

J-CAT 得点＝23.53×can-do スコア+136.56　(p<0.001)

Can-do 記述文の内容が「できる」と「できない」のカットスコアを 5 ととる。つまり、5 以上を当該内容が「できる」とみなす。これに従い、回帰式に can-do スコアの 5 を代入した結果、つまり can-do スコアの 5 に相当すると予測される J-CAT の点数が表 4 の右端の欄「回帰」に示されている。なお、R16 は can-do スコアと J-CAT 得点の相関が低いので対象外とした。この結果から、回帰分析による、can-do スコアとテスト得点の関連づけの可能性を確認できた。これにより、より具体性を伴ったテスト得点の解釈が可能であることが示された。一方で、今回の結果からは、得点の差があまりないことが読み取れる。最も高

いR1の257点と最低点であるR17が223点であり、その差は34点しかない。これは、can-do記述文の内容に難易差がつきにくいものであったことが原因だろう。

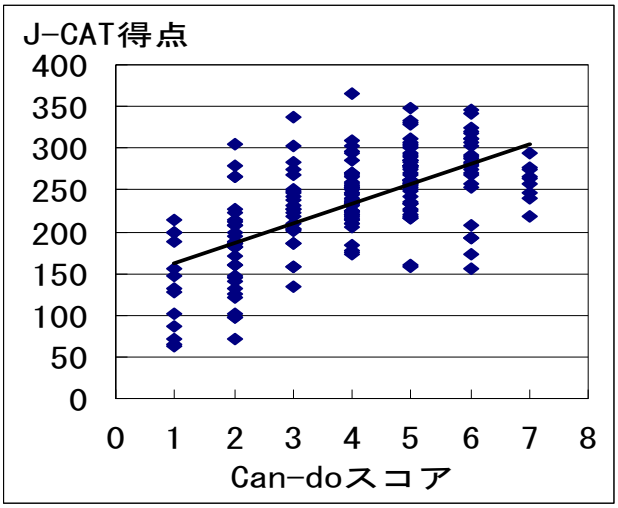

図4　Can-doスコア（R1）とJ-CAT得点の関係

特に、漢字が「分かる・分からない」がcan-do記述文の内容が「できる・できない」をほぼ決定づけていると思われる。例えば、R17の授業名でも、専門科目の漢字表記は大変難しく、R22の案内板でも固有名詞を伴うものは大変難しい。この漢字の影響でcan-doの難易度に傾斜がつきにくかったと言えよう。この点から、今後のcan-do調査では、can-do記述の内容の難易度が分散するような配慮が必要であることが明らかになった。

今後、今回得られた上記の知見を基に、国際交流基金で研究・開発が進められている日本語スタンダーズとcan-do記述の成果を参照しながら、不変性のあるJ-CATの得点との関連付けを試みる予定である。

謝辞

Can-do statements調査用紙をご提供いただいた島田めぐみ氏・谷部弘子氏に感謝いたします。J-CATの開発・研究メンバーである、伊東祐郎氏、中村洋一氏、菊地賢一氏、中園博美氏、本田明子氏、赤木彌生氏に感謝いたします。また、紙幅の関係でお名前・機関名を省略させていただきますが、プレテストおよびcan-do調査にご協力いただいた協力者および機関各位に感謝いたします。 なお、本研究は山口大学研究教育後援財団(2005年）、科学研究費補助金基盤（A）(18202012)（2006～2009年）の助成を受けています。

＜参考文献＞

Alderson J.C. (2005) *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment.* New York: Continuum.

三枝令子（2004）『日本語Can-do-statements尺度の開発　研究成果報告書』（科学研究費補助金基盤研究（B1）課題番号13480068）

島田めぐみ・三枝令子・野口裕之（2006）「日本語Can-do-statementsを利用した言語行動記述の試み―日本語能力試験受験者を対象として―」『世界の日本語教育』第16号, pp.75-88, 国際交流基金.