コンピュータ適応型語彙テストの開発と有用性の検証

――オープンソースプラットフォーム Concertoを利用して――

水本　篤

はじめに

コンピュータ適応型テスト (computerized adaptive testing: CAT) は項目応答理論を用いて，より少ない問題数でより高い精度の能力測定が可能であることが知られているが，研究者や実践者が自作の CAT を開発するには技術的なハードルが高い。そこで本稿では，比較的設置が簡単で，オープンソースプラットフォームである Concerto を使用し，自作のコンピュータ適応型語彙テストを開発した経緯と有用性の検証結果，そして今後の適用可能性などについて報告する。

1. 言語テスト理論とコンピュータ適応型テスト

1.1 古典的テスト理論と項目応答理論

言語テスト理論は，素点をもとにしている古典的テスト理論 (CTT: classical test theory) と，古典的テスト理論の限界を改善した項目応答理論 (IRT: item response theory) の 2 つに大きく分類することができる（竹内・水本，2012）。

表 1 は大友 (2009) による CTT と IRT の比較である。古典的テスト理論のテスト得点は，本来は順序尺度の情報しか持っていないため，足したり引いたりという計算は間隔尺度であるとみなして行っている。そのため，40 点と 50 点，80 点と 90 点は両方とも同じ「10 点差」という，直感的にも等間隔ではないであろうものを等間隔とみなすことになる。一方，項目応答理論では，素点を間隔尺度のロジット (logit) という間隔尺度に変換するため，純粋な間隔尺度上での点数の比較が可能になる。

また，受験者能力（得点）やテストの項目特性（難易度など），そして測定の精度も，古典的テスト理論の場合は受験者集団や使用するテストに依存するため，結果として得られた点数の評価については，能力の高い（もしくは低い）受験者集団が受験していたせいなのか，テスト項目が難しい（もしくは簡単な）ためその点数になったのか，テストを実施してみるまでわからない。一方，項目応答理論では，このような古典的テスト理論の問題点を克服できるため，現在では多くの大規模テストが項目応答理論をその開発・分析に用いている。

異なる受験者集団に異なるテストを実施して，そのテスト得点が比較可能となるようにするには，等化 (equating) とよばれる方法が用いられるが，古典的テスト理論ではそのような等化は非常に困難である。一方，項目応答理論では，等化が容易にできるため，いくつものテストフォームを用意し，テスト得点を等化することによって，どのテストフォームを受験しても，同一尺度上での比較が可能となる。

さらに，項目応答理論を用いると，テスト項目をたくさん集めたアイテム・バンク (item bank) を作り，その中から受験者個人の能力に応じた問題を出題し，能力を測定するコンピュータ適応型テスト (CAT) も実施可能になる。

表 1
古典的テスト理論と項目応答理論の比較（大友，2009 を一部改変）

領域	古典的テスト理論	項目応答理論
テスト得点	順序尺度	間隔尺度
得点	素点（偏差値など）	ロジット値
受験者能力（得点）	テストに依存する	テストに依存しない
項目特性	受験者集団に依存する	受験者集団に依存しない
測定の精度	受験者集団全体を対象	受験者個人ごとを対象
等化・CAT	極めて困難	容易に可能

1.2 コンピュータ適応型テスト (CAT)

コンピュータ適応型テスト（以下，CAT と省略）は，研究自体，40 年以上の歴史があり (Thomson & Weiss, 2011)，言語テストの分野では過去 30 年間行われてきている（小山，2010）。CAT は上記に述べたような項目応答理論の利点を十分活かした形式であるため，米国の GRE (Grade Record Exam)，GMAT (Graduate Management Admission Test)，SAT (Scholastic Aptitude Test) などの，テスト結果が受験者の今後に影響を与えるハイステイクス (high-stakes) な大規模テストでも利用されている。

また，CAT は，1998 年から 2006 年に実施された TOEFL (Test of English as a Foreign Language) の CBT (computer-based testing) でも，Listening と Structure セクションで利用されていた。CAT と CBT は似ているため混同しがちであるが，CAT が出題される問題が毎回違うのに対し，CBT は出題内容と順番が決まったコンピュータを使用したテストを指す。ちなみに，現行の *TOEFL iBT*(internet-based testing) は，項目応答理論を用いているものの，CAT ではなく CBT になっている。

日本でも，CASEC (Computerized Assessment System for English Communication) や，J-CAT (http://www.j-cat.org/)，そして大学入試で活用できる 4 技能英語検定である GTEC CBT (http://www.benesse-gtec.com/cbt/) などのテストでも CAT が活用されており，それが普及していることがわかる。

図 1 は CAT で問題がどのように出題されるかのイメージを示したものである。（出題ルールによるが）1 問目が出題されて正解の場合には，2 問目により難しい問題が出題される。この出題と能力推定を繰り返し，最終的に，設定された問題数が出題されるか，設定された標準誤差 (standard error: SE) よりも誤差値が小さくなった場合に出題がストップする。このCAT における出題方法は，視力検査でランドルト環のすきまが空いている部分が正しく答えられれば，次に提示されるものは小さくなり，間違えれば大きくなるという例えで考えればわかりやすい（中村，2002)。図 1 中の上下に伸びたエラーバーは標準誤差を示しており，問題が出題されるにつれて標準誤差が小さくなっていて，より正確な能力推定ができるということがわかる。

CAT では，このような出題が受験者個人の能力に応じてなされるため，決まった問題数が含まれる従来の紙版のテストと比べて，理論的にはテスト時間が短縮でき，より正確な能力推定ができることが知られている。また，CAT の大きな利点として，受験者個人を測定の対象としているため，受験

者集団の無作為抽出の必要がない。これは，研究や教育実践で測定を行う際には非常に望ましいことである。

CAT を開発する場合，(a) 使用する CAT プログラムの選択，(b) アイテム・バンクの構築と調整 (calibration)，(c) 出題ルールや推定方法，フィードバック方法のそれぞれのステップを決定し，その後，CAT を実際に実施するという流れになる。

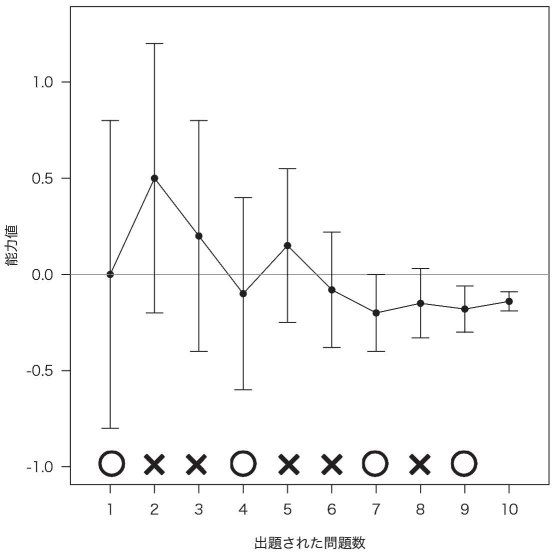

図 1　CAT の出題イメージ（エラーバーは標準誤差）

2. 使用する CAT の構築

2.1 使用する CAT プログラムの選択

これまでに述べてきたような CAT のテストとしての利点にもかかわらず，研究者や実践者が自作の CAT を開発したという事例は少ない。これは，開発には技術的なハードルが高く，テスト会社が中心となって CAT を開発するしか方法がなかったためであると考えられる。

また，CAT プログラムを選択する場合に，高価な商用ソフトを利用すると，金銭的な問題だけではなく，使用できる OS が限定されたり，プログラムの裏側でどのような処理がされているかが分からない設定になっていたりすることもある。これらは，研究者や実践者が個人レベルで CAT を構築しようと考えたときに大きな障害となる。

しかし，近年，研究者や実践者個人が CAT を構築できるような無償のオープンソースプログラムが開発されてきている。たとえば，e ラーニングプラットフォーム Moodle のプラグインとして UCAT Module for Moodle (https://github.com/VERSION2–Inc/moodle_ucat) のようなものも無償で公開されている (Kimura & Akiyama, 2009)。Moodle をすでに使用している場合には，このプラグインを利用すれば CAT の実装が可能である。

本研究では，Moodle を使用していなくても CAT を開発する場合を考えて，ケンブリッジ大学の The Psychometrics Centre が開発した，オープンソースプラットフォームである Concerto (http://www.psychometrics.cam.ac.uk/newconcerto) を使用する。Concerto は，フリーのデータ解析環境 R をサーバー上で動かし，HTML でテスト実施画面を作成できるようにしているため，R と HTML のコードが少し書ければ，自分で CAT を構築できるという便利なプラットフォームである。Concerto を利用するには，1 ヶ月 150 人までの受験が可能というフリーアカウントを使うか，Amazon Machine Images や自分のサーバーにインストールして使うことになる。本研究では CentOS 6 サーバーに Concerto をインストールして利用した。

2.2 アイテム・バンクの構築と調整

アイテム・バンクは水本 (2006) で使用された語彙テストを基にした項目を使用した。このテストは語彙リストの SVL12000(ALC, 2001) のうち，8,000 語までの各レベル 30 語（8×30）の 240 問で構成されている。問題は以下のような，日本語に対応する英単語を 4 択から選ぶというシンプルな形式である。

Q: 心の，精神の　(A) essential　(B) creative　(C) loose　(D) mental

水本 (2006) では，この 240 問の語彙テストを日本人大学生 716 名に実施した。今回の研究では，このテストの項目をアイテム・バンクとして使用するために，項目分析（古典的テスト理論とラッシュ・モデル）で 150 問を

選定したあと，2 パラメータ・ロジスティックモデル (IRT) で困難度と識別力を再度推定した。そのため，最終的には 150 問がアイテム・バンクとして使用された。本研究の分析にはすべて R Version 3.1.1 (R Core Team, 2014) を用いた。また，項目応答理論に基づく分析は，R の ltm パッケージ (Rizopoulos, 2006) を使用した。図 2 はこのアイテム・バンク 150 問の項目困難度を箱ひげ図で示したものであり，項目困難度 0 を中心としてこの付近に多くの項目が存在している，CAT で受験者の能力を弁別するには理想的な分布になっていることがわかる。

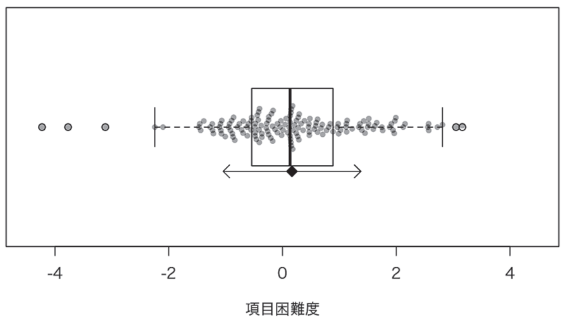

図 2　アイテム・バンク 150 問の項目困難度を箱ひげ図で示したもの
（矢印部分は平均と +/–1 標準偏差を示す）

2.3 出題ルール，推定，フィードバック方法の決定

CAT では，(a) 第 1 問目の選択方法，(b) 問題ごとの能力値推定方法，(c) テストの終了規則 (stopping rule)，(d) 最終能力値推定方法を決めておく必要がある。Concerto では，R コードが使用できるため，平均的な能力を 0 と想定し，–1 から 1 の値の中から乱数を 1 つ発生させることで，初期能力値を決定した。そして，その初期能力値に従って，第 1 問目が出題されるようにした（これにより，第 1 問目に出題される問題はある程度ランダム化できる）。問題ごとの能力値推定とテスト終了時の最終能力値推定は，R の catR パッケージ (Magis & Raîche, 2012) の thetaEST 関数でベイズ理論に基づく推定 (Bayesian modal estimation) を用いた。また，能力値の標準誤差の推定も catR パッケージの semTheta 関数を用いて行った。

catR パッケージではシミュレーションが可能であるため，テストの終了

規則も検討することができる。一般的に CAT が目標としている標準誤差 0.33（古典的テスト理論の信頼性係数で 0.90 相当）を使用した場合，平均的な能力の受験者が今回作成している CAT を回答すると，（モンテカルロ・シミュレーションにより）問題数は 30 問以下になることがわかった。よって，今回実施する CAT については，どの受験者にも一律に項目が 30 問出題された時点でテストが終了するように設定した。

図 3 は今回作成した CAT のテスト実施画面と，テスト終了時のフィードバック画面である。フィードバック画面で表示される語彙レベル，英検，TOEIC，TOEFL，IELTS などの推定得点・レベルは，水本 (2006) で作成された回帰式を元に，CAT の結果を換算することで表示させるようにした。

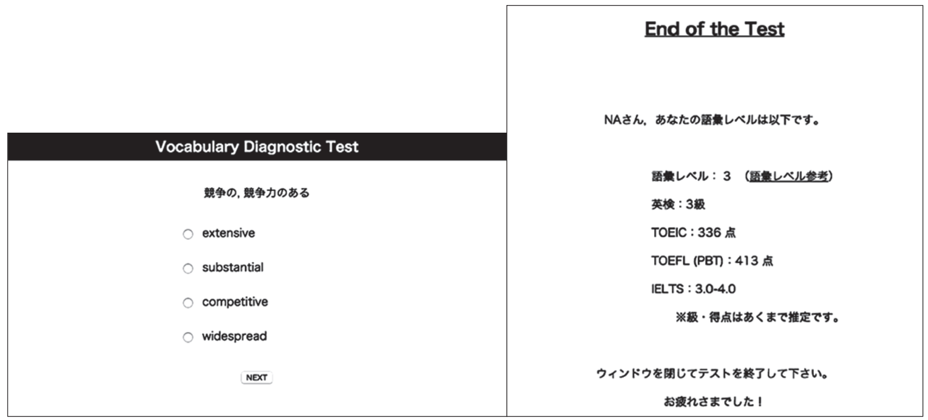

図 3　CAT のテスト実施画面（左）とフィードバック画面（右）

3. CAT の実施

3.1 参加者と手順

上記で説明した手順で作成された CAT を，関西の私立大学 1 年生 268 名に対して実施した。そして，1 ヶ月以内に同じ参加者に（150 項目のアイテム・バンクに含まれている）68 項目の紙版のテストを実施した（68 項目は難易度に偏りのないようにまんべんなく選択した）。2 つのテストを実施した理由は，CAT と紙版によって推定される能力値に違いがあるかどうかを

検討するためである。CAT と紙版テストは，受験するテスト形式の影響を相殺するために，半数の参加者が CAT，残り半数の参加者が紙版をはじめに受験し，2 回目の実施のときには形式を逆転させて実施した。テスト得点は，2 つのテスト形式を同じ受験者が回答したため共通受験者法を用いて等化された。

また，CAT 終了後に，フィードバック画面を見た参加者に「この語彙テストの結果は，あなたの現在の語彙力をどの程度正しく測定していると思いますか？」という質問を提示し，「まったく正しくない」から「とても正しい」の 6 件法で回答を求めた。

3.2 結果と考察

図 4 は CAT と紙版テストの点数の分布と相関を示したものである。相関係数 (*r*) は 0.92 (95% CI [0.90, 0.94]) で高い相関が得られた。能力推定値の平均は，CAT が 1.71 (*SD* = 1.13)，紙版が 1.72 (*SD* = 0.95) でほぼ同じ値となった。平均値差の効果量は，*d* [95% CI] = 0.01 [–0.03, 0.06] であったため，2 つのテスト形式から得られた能力推定値の平均にほとんど違いがないということがわかる。

また，2 つのテスト形式から得られた能力推定値の標準誤差の平均は，CAT が 0.28 (*SD* = 0.11)，紙版が 0.39 (*SD* = 0.11) で，効果量は，*d* [95% CI] = 0.98 [0.89, 1.07] であったため，標準誤差は CAT のほうが紙版テストよりも小さかった。これにより，CAT が紙版を測定の精度という観点から上回っていたことがわかった。

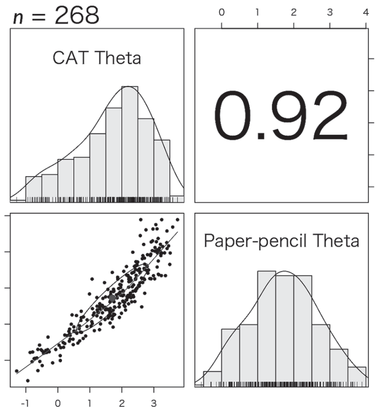

図 4
CAT と紙版テストの点数の分布と相関

CAT は受験者個人の能力に応じてランダムに 30 問出題されるのに対し，紙版テストは決まった 68 問が全員に出題されたことを考えると，問題数が半分以下の CAT でほぼ同じ能力値推定が行え，標準誤差も紙版より小さくなるという結果は，今回の研究で構築した CAT が紙版テストよりも測定の精度が高いということを示している。

図 5 は，CAT 終了後に表示されるフィードバック画面の結果に対して，参加者がどのように感じたかを調べたアンケートの結果である。能力推定平均値や標準誤差では，CAT が紙版テストと比べて優れているという結果が得られたが，フィードバックに対しての参加者の反応は「正しくない」と「正しい」の半々に分かれたことが図 5 からわかる。

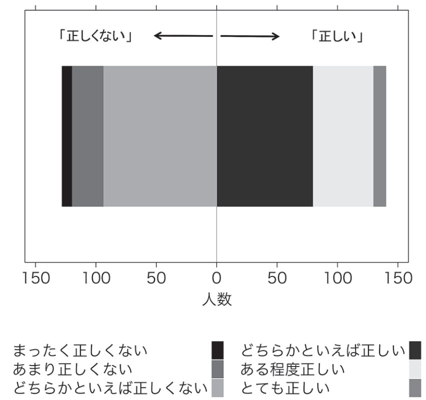

図 5　CAT の結果とフィードバックに対するアンケート回答

この結果の理由としては，フィードバック画面で表示される，語彙レベル，英検，TOEIC，TOEFL，IELTS などの推定得点・レベルが，受験者が自分で感じているレベル（特に総合的英語能力）とは違っていることが多かったことによるものと推察される。この結果から，今回開発した CAT のフィードバック方法は，改善の必要があるといえるだろう。

今後は，フィードバック方法の改善に加えて，アイテム・バンクに含まれる問題を増やしたり，問題形式の多様化を検討していく予定である。特に，今回 CAT で使用した語彙テストの項目は，部分的測定テスト (discrete-point test) 形式であり，リーディングなどのその他のスキルを含んだ統合的

測定テスト (integrative test) 形式で測定を行いたい場合には，アイテム・バンクもかなり大きなものにしなければ対応できない。実際，過去に実施されていた TOEFL CBT でも，CAT が使用されていたのは，Listening と Structure セクションだけで，現行の *TOEFL iBT* では CAT がまったく使われていないということは，統合的測定テストにおいて CAT を利用するということが，問題数や問題自体の複雑さなどの点で現実的・技術的に難しいということを示している。

CAT は項目応答理論を全面的に用い，測定の精度という観点からも優れているが，その測定対象が今回のように「言語知識」ではなく，「言語運用能力」だった場合には使えるのか，また，プレイスメント，診断，熟達度など，さまざまなテストの目的に応じて，CAT がより効果的に活用できるのはどういう場面なのかなどという点についても，今後検討を続けていきたい。

おわりに

本研究では，オープンソースプラットフォームである Concerto を使用して，自作の CAT 版語彙テストを開発し，その有用性について検討した。紙版テストと比較した結果，半分の問題数で同じ能力推定値を，より小さい標準誤差で得ることができることが明らかになり，開発した CAT の有用性が確認された。しかし，フィードバック方法については，改善の必要があることがわかった。（今回作成した CAT 版語彙テストは，次の URL からアクセス可能である。http://langtest.jp/#cat）

これまで一部のテスト会社の作成したテストを使用する以外では，高価な商用ソフトを使うことによってのみ構築が可能であった CAT を，（R や HTML のコードを書かなければならないというのは，まだ少し敷居が高いが）Concerto のようなオープンソースプラットフォームを使うことにより，研究者や実践者が自作できるようになったのは大きな進歩である。外国語教育研究分野での研究や実践において，これまでに多くのテストが実施されてきていることは想像に難くない。本研究で示したように，より精度の高い測定ができる CAT で，これまでに開発されたテストを再利用できる可能性があるのなら，それらのテストを無駄にせず共有し，さらなる精緻化の方法を共に積極的に模索していくべきであろう。

引用文献

ALC (2001). Standard Vocabulary List 12000 (SVL12000). Retrieved from http://www.alc.co.jp/vocgram/article/svl/

Kimura, T., & Akiyama, M. (2009), Reinforcing development of a Moodle-based English test with multiple choice maker, TDAP block, and CAT module. *Language Education & Technology*, *46*, 233–245. Retrieved from http://ci.nii.ac.jp/naid/110008441797

小山由紀江（2010）「テストの歴史的変遷とコンピュータ適応型テストの意義」『名古屋工業大学共通教育 New Directions』, 28, 13–26. Retrieved from http://repo.lib.nitech.ac.jp/bitstream/123456789/3639/1/ndnit2010_13.pdf

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1–31. doi:10.18637/jss.v048.i08

水本篤 (2006)「語彙サイズテストは何を測っているのか？―語彙サイズテストの開発における問題点」『統計数理研究所共同研究レポート 190：言語コーパス解析における共起語検出のための統計手法の比較研究』, 71–80. Retrieved from http://www.mizumot.com/files/VocSizeMeasure.pdf

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』東京：大修館書店 .

大友賢二 (2009)「項目応答理論――TOEFL・TOEIC 等の仕組み」『電子情報通信学会誌』, 92(12), 1008–1012. Retrieved from https://www.ieice.org/jpn/books/kaishikiji/2009/2009121.pdf

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.1) [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org/

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. doi: 10.18637/jss.v017.i05

竹内理・水本篤 (2012)（編著）『外国語教育研究ハンドブック――研究手法のより良い理解のために』東京：松柏社 .

Thompson, N., & Weiss, D. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*, 1–9. Retrieved from http://www.pareonline.net/pdf/v16n1.pdf