学内日本語学習者のレベル判定指標のための Computer Based Test 開発
－テスト運用までの過程と課題に関する実践報告－

寅丸真澄（早稲田大学）、岩下智彦（早稲田大学）、伊藤奈津美（早稲田大学）、
沖本与子（早稲田大学）、井下田貴子（早稲田大学）、三谷彩華（江戸川大学）

Development of a Computer Based Test for Japanese Language Learners for Institutional Use -A Practical Report on the Process and Problems of CBT Administration-

Masumi TORAMARU, Waseda University　Tomohiko IWASHITA, Waseda University
Natsumi ITO, Waseda University　Tomoko OKIMOTO, Waseda University
Takako IGETA, Waseda University　Ayaka MITANI, Edogawa University

要旨：本稿は、学内日本語学習者の自律的な学習環境の整備を目的とした Computer Based Test（CBT）開発とその運用過程、および課題に関する実践報告である。本報告は、知見の共有により、オンライン化が進む言語教育現場で期待されるカリキュラム対応型の CBT 開発に寄与することを目的とする。本 CBT 開発の課題は、次の 4 点であった。1）学内 LMS 外から受験可能な環境構築、2）総合日本語科目のレベル別学習項目との整合性の確保、3）対外的に説明可能なテストの質の担保、4）継続運用が可能な体制の構築である。また、今後検討すべき課題として、カリキュラム変更や学習者の日本語能力の変化への対応を挙げた。

キーワード：テスト開発、CBT（Computer Based Test）、学習環境、言語知識、目標規準準拠テスト

1. はじめに

早稲田大学日本語教育研究センター（Center for Japanese Language，以下「CJL」）では、日本語学習者（以下、「学習者」）の自律性の育成を目的に、自律的な学習環境整備の一環として、学習者の自己判断で日本語履修レベル、および履修科目を選択できるカリキュラムを採用している。そのため、学習者は自身の日本語能力と興味や関心、必要性等を踏まえ、週 650 コマに上る科目の中から科目を選択、履修することになる。選択に際しては、シラバスやテキストに加え、学習者の日本語レベルと履修科目のレベルの適切性を判断するための指標として、JCAT（Japanese Computerized Adaptive Test）（今井他 2010）が使用されてきた。学習者は教室外で JCAT を受験し、その得点と CJL 日本語レベルの対照表を参照することにより、自身の日本語能力を客観的に判断していたと言える。さらに、四技能をバランスよく学ぶ総合日本語科目（入門～6 レベル）では、CJL の学習項目に合わせたレベル別小テストを実施し、学習者の自律的なレベル・科目選択を支援していた。

しかし、2018 年度に、初級学習者に対するレベル判定の信頼性の問題や CJL 各レベルの学習項目との整合性等の観点から、JCAT とレベル別小テストを組み合わせるというシステムを見直すことになり、新たな Computer Based Test（CBT）の開発を決定した。本 CBT における課題は、1）学内 LMS 外から受験可能な環境構築、2）総合日本語科目のレベル別学習項目との整合性の確保、3）対外的に説明可能なテストの質の担保、4）継続運用が可能な体制の構築という 4 点であった。

近年、高等教育機関では、外国語としての日本語能力を推定するための独自のテスト開発が行われており、テスト開発の経緯や内容に関する検証が報告されている（藤田他 2017、小森他 2017)。本稿では、新たに開発する CBT を「CJL レベルチェックテスト」と称し、その開発と運用過程、課題について報告する。特に先行研究において言及が少なかった、カリキュラムに即した CBT 開発の知見を共有することにより、オンライン化が進む言語教育現場で期待されるカリキュラム対応型の CBT 開発に寄与したいと考える。

２．開発の方針

本プロジェクトでは、2018 年度秋学期～2020 年度秋学期までの 2 年半に、「漢字」、「文法・語彙」、「読解」、「聴解」の 4 種のテストを開発することとした。開発にあたっては、上述の 4 つの課題の解決と、2021 年度春学期からの運用の実現が最も留意された。以下、4 つの課題に対してどのように解決を目指したか順に述べる。

1）学内 LMS 外から受験可能な環境構築：CBT を前提に、テストの配信・開発支援等を行う民間企業、テストコンテンツの作成ソフト、学内 LMS の機能、学内のテストシステム等に関わる情報を広く収集し、検討した。その結果、学内 LMS から独立したシステムであることに加え、開発・運用・維持にかかるコスト、個人情報に対する安全性、継続的に使用可能である点を勘案し、学内で管理するテストシステム（EtestingSystem）を使用することとした。このシステムは、セクション別、各項目レベル別の得点表示が可能である。そのため、それまでの日本語学習歴を問わず、CJL における各レベルの学習内容がどの程度理解できているか、総合的にどの程度の日本語レベルであるかを示す受験結果を学習者に提供できるという点で、自律的なレベル・科目選択に適した仕様であった。

2）総合日本語科目のレベル別学習項目との整合性：テスト開発において、テストの種類、出題基準、基準設定の問題として検討した。言語テストの種類を区別する観点としては、カリキュラム等のある目標に対する受験者の達成度がわかる 「目標基準準拠テスト」（Criterion Reference Test；CRT）と、ある程度能力に幅のある集団内において、受験者が相対的にどの程度に位置しているかを測る「集団規準準拠テスト」（Norm Reference Test；NRT）という観点がある（ブラウン,J.D. 1999)。総合日本語科目の学習内容に準拠し、その理解度を診断するためのテスト開発を目的とした場合、CRT の開発が求められるといえる。しかし、本テストの受験者は、本学の在学生だけではなく、新入生も想定されており、そうした学習者の日本語レベルを判定するためには、NRT として

開発する必要があった。そのため、項目作成においては、原則として、総合日本語科目における学習項目を考慮した出題基準を用いて作成し、各レベルで必要と想定される項目レベルを設定した。また、これに加え、開発過程においては、NRT 開発の手続きに沿って、テストの試行調査、項目分析を行い、各項目が事前に想定したレベルの難易度と合致しているかを検証した上で、テスト項目のプールに加えることとした。また、基準設定においては、項目レベル別の正答率と各レベル別受験者の得点分布を考慮して得点区分を作成した。この 2 つの段階を経ることで、CRT の要素を含んだ NRT 開発を目指した。

3）対外的に説明可能なテストの質の担保：テストの質については、内容的妥当性、基準関連妥当性の観点、および古典的テスト理論、項目応答理論（Item Response Theory；IRT）における項目の難易度、識別力、および信頼性の指標に基づいて項目の質に関する検証を行った。内容的妥当性は、各レベルコーディネーターを含む常勤教員全員による検討を行い、基準関連妥当性については、「漢字」および「文法・語彙」において、それぞれ類似したテスト得点との相関係数を確認した。

4）継続運用が可能な体制の構築：本テストが項目固定型であることから、項目の繰り返し使用による項目露出が最大の課題であった。そこで、数回に分けた試行調査の後、全ての解答データに対して IRT を用いた等化分析を行った。その上で、事前に想定した項目レベルの観点に、等化後の共通尺度上での難易度の観点を加えた項目レベルを改めて付与し、項目プールを作成した。そこから各レベルの既定項目数を年度ごとにランダムに選択肢、出題する体制を整え、項目の繰り返し使用による信頼性の低下を最小化することを目指した。

上記（1）から（4）の方針に加え、各学期約 70 クラスを対象とした大規模な試行調査においては、学内での日時をずらした一斉受験のため、各教室で受験を案内するための指示や、不測の事態に対応するための準備も必要であり、受験にあたっての教師用、学習者用マニュアルの作成、ICT 環境の確認なども計画に含められた。

3．CJL レベルチェックテスト開発の概要

3.1　テストの種類と出題形式

CJL における日本語レベルは 0 から 8 レベルまであるが、開発するテストは「漢字」、「文法・語彙」、「聴解」、「読解」の 4 種で、受講希望者が科目登録をする際の指標として、「漢字」は漢字科目 1～5、「文法・語彙」は総合日本語科目 0～6 のレベルを判定するものである。一方、「聴解」と「読解」は、7 レベル以上かどうかを判定するためのものである。いずれも EtestingSystem の仕様とテスト受験時間の制約を考慮した多肢選択式テストとし、「わからない」を含む 5 肢選択式とした。なお、開発テストの試行調査は、原則として開講時にクラスオリエンテーションの一部として実施され、事後の分析にはテストシステム上で同意を得た学習者の解答のみを開発に使用した。「聴解」と「読解」については、授業外に任意の協力者を募り、その解答を開発に使用した。

3.2 各テストの概要と試行調査

以下、各テストの概要と開発にあたって行った試行調査について述べる。

「漢字」は、既に紙ベースのテストを当時の学内 LMS へ CBT 化する作業を終えていたため（岩下・沖本 2018）、漢字 CBT の質の検証も含めた計画を立てた。出題基準は、テキストの学習漢字語彙と語彙レベルに関する情報が付与された語彙表（日本語学習辞書支援グループ 2015）を利用し、漢字と語彙双方の難易度を考慮した出題基準を作成した。

「文法・語彙」は、全 164 項目を作成し、各項目に項目レベルを設定した。その後、全ての項目について内容的な妥当性の確認を行った上で、各学期 3 分の 1 の項目を入れ替える形で複数のテストセットを作成し、試行調査を行った。そして、試行調査ごとに項目分析を行い、項目の質及び学習者の能力に適したテストであるか適切性を検証した。

一方、「聴解」と「読解」は、CJL7 レベル以上を判定する目安の 1 つとして日本語能力試験（以下、「JLPT」）の N1 を想定した項目を作成し、中級以上の学習者を対象とした試行調査を行った。JLPT の出題基準は公開されていないため、公開されている語彙表（日本語学習辞書支援グループ 2015、松下 2011）の重複語彙をもとに、適切性を定性的に検討し、上位約 2 万語を抽出した独自の出題基準（全 22,002 語）を作成した。

3.3　開発スケジュール

テストの開発スケジュールは図 1 のとおりである。4 種のテスト開発を同時進行させるのは人員面で困難であったため、「漢字」、「文法・語彙」、「聴解」と「読解」の順で開発時期をずらして進めた。

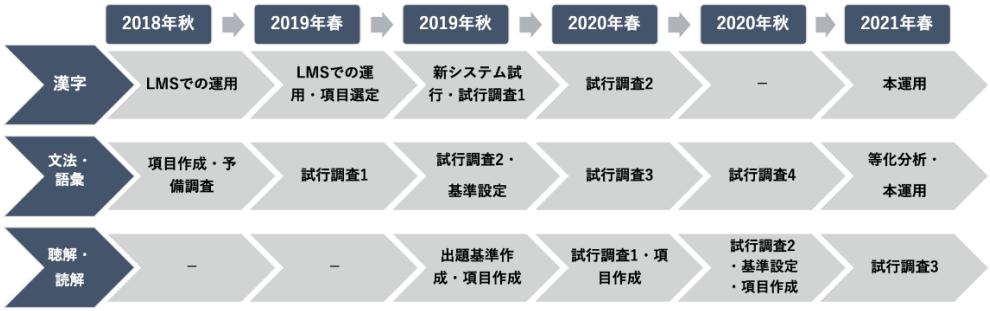

図 1　開発スケジュール

3.4　レベル判定の基準設定

基準設定についてはアンゴフ法等様々な方法を検討、試行したが、作成した得点区分と実際の履修者の得点分布を対照したところ、いずれも乖離があった。そこで、まず「文法・語彙」については、学習者の履修レベル別の得点分布、および到達度評価としての観点から項目レベル別に望ましい正答率を定め、これらを総合し、定性的に基準を作成し

た。その上で、実際の履修レベルを予測できるか回帰分析による検証を行い、基準設定の質を担保した。加えて、IRT による等化分析の中で、テスト特性関数における得点区分と受験者特性値の対応関係を確認した。受験者特性値から期待されるテストの素点を確認することで、本得点基準の妥当性を検証し、概ね問題がないことが明らかになった。

「聴解」と「読解」については、ブックマーク法を参考に、まず項目を古典的テスト理論に基づいた正答率順に並べたテスト冊子を用意し、テスト開発グループ内でレベル 7 以上に相当する学生がどの項目まで合格しているのが妥当か検討を行い、合否分割点となる得点基準を設定した。その合否分割点が適切であるかについては、実際の調査協力者の得点分布および JLPT の N1 保持者の得点分布を確認した。

3.5　テストの妥当性・信頼性の検証

いずれのテストも内容的妥当性、基準関連妥当性の観点から検討した。まず内容的妥当性については、いずれのテストにおいても当該科目の内容を熟知した常勤教員が複数名で内容を精査し、項目レベルと内容の適切性、正答、誤答の適切さなどを確認した。

基準関連妥当性については、延べ 1000 名以上の調査協力者のうち、構成概念が類似した習熟度テストの得点が得られた者の解答を対象に、その相関係数を確認した。「漢字」は本プロジェクトに先立つ CBT 化の際に、K-SPOT との相関係数に、強い相関（r=0.7～0.9、N＝16）があることを確認していた。「文法・語彙」は、JCAT との相関係数に、強い相関（r=0.7、N=25）があることを確認した。

信頼性については、α 係数を算出し、「文法・語彙」、「漢字」はいずれも問題のある値ではないことを確認した（α=0.9、「文法・語彙」: N＝326～608（全 4 回）、「漢字」: N＝244～262（全 2 回））。ただし、「聴解」と「読解」については、相対的にやや低い値であった（α=読解＝0.7、聴解 0.8、N＝25）。これは項目数の影響もあると考え、今後、項目数を増やす方針を固めている。加えて、「文法・語彙」は、IRT によるテスト情報量曲線を確認し、相対的に上級相当の能力推定における誤差が大きい点を確認した。

4. まとめ

本稿では、自律的な学習環境の整備の一環として、レベル・科目選択を支援するためのカリキュラムに即した CBT 開発に関して報告を行った。冒頭で挙げた 4 つの課題を解決し、カリキュラムに即し、かつ一定の質が担保された CBT を開発することができた。

ここまで開発および運用過程について述べたが、今後、検討すべき課題としては、カリキュラム変更や学習者の日本語能力の変化への対応が挙げられる。今回開発した CBT がカリキュラムに即しているがゆえに、カリキュラムの変更の影響を受けやすいことが懸念される。必要に応じて、今回開発した CBT がどの程度機能するか検証し、テスト項目の更新を行う必要がある。

最後に、実践知の共有を目的として、開発および運用場面で生じた事案とその対応につ

いて述べる。まず、テスト冊子の数と項目数についてである。当初「文法・語彙」は、受験者負担の軽減を目的に、初級用 60 問、中級以上用 90 問の 2 版が作成された、しかし、テスト冊子の管理および項目管理の煩雑さ、運用時の混乱等が課題となったため、2020 年度以降は 90 問 1 セットで運用することとなった。開発側の負担と受験者側の負担のバランスを考慮した開発計画の重要性を認識する必要があるといえる。

また、一定以上の規模で試行調査を行う場合、受験時に想定外のトラブルが発生する。学内 LAN、受験媒体を含めたテスト受験環境の確認、受験用教室の手配、受験者用の確認ガイド、担当教員用の受験方法の説明資料等に翻訳を付けて作成するなど、実施時の混乱がないように留意することは重要である。また、テストシステムの選別から分析におけるデータの管理、データ形式に関する共通認識を関係者で統一しておくことも継続的な開発・運用には必要である。

参考文献

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. 2010.『J-CAT Japanese Computerized Adaptive Test 日本語能力をコンピュータで測る』山口大学留学生センター.

岩下智彦・沖本与子. 2018「漢字習熟度に応じた強化が必要な要素の解明―漢字診断テストを用いて―」.早稲田大学日本語教育学会 2018 早稲田大学日本語教育研究センター30 周年特別企画.早稲田大学日本語教育学会.36-37. http://gsjal.jp/wnkg/dat/2018at/proceedings.pdf(2021 年 6 月 24 日最終閲覧)

加納千恵子. 2008.「レベル別漢字語彙処理能力テストの問題形式―WEB 漢字テストのマルチレベル化に向けて」『筑波大学留学センター日本語教育論集』第 23 号. 1-13.

小森和子・柳澤絵美・安高紀子. 2017.「日本語プレイスメント・テストの開発と問題項目の分析―国際日本学部の ET 日本語科目における試み―」『明治大学国際日本学研究』9. 31-61.

日本語学習辞書支援グループ. 2015.「日本語教育語彙表 Ver 1.0」https://jreadability.net/jev/(21 年 6 月 24 日最終閲覧)

藤田恵・平山紫帆・栗田奈美・金庭久美子・数野恵理. 2017.「Web による日本語プレイスメントテストの開発―外国人留学生の受け入れ拡大にむけて―」『立教大学ランゲージセンター紀要』37. 77-83.

ブラウン, J.D. 1999. 和田稔訳『言語テストの基礎知識』. 大修館書店.

松下達彦（2011）「日本語を読むための語彙データベース（VDRJ）Ver. 1. 1 （研究用)」http://tatsuma2010.web.fc2.com/(2021 年 6 月 24 日最終閲覧)