教育情報学研究　第1号（2003）
Educational Informatics Research, 2003, No. 1, 14-16

日本語基礎能力テストの特性（2）－項目反応理論に基づく測定性能の分析－

佐藤 洋之，[*] 伊藤 博美，[**] 倉元 直樹[***]

[*] 東北大学大学院教育情報学教育部
[**] 東北大学大学院文学研究科・秋田県本荘高校
[***] 東北大学アドミッションセンター・東北大学大学院教育情報学教育部

要旨：本研究の目的は、伊藤他（2003）で内容的妥当性を確認した日本語基礎能力テスト（平他、1998）の語彙理解力項目について、項目反応理論に基づき、その測定性能を分析することである。伊藤他（2003）の分析の結果、要注意の「レベルC」と判定された項目が若干存在したものの、項目作成から10年余りを経た現在でも、全体として十分機能することが示された。本研究では、「レベルC」の項目を除く444項目を分析の対象とする。

テスト情報曲線から、項目プールは全体としては能力尺度値θの尺度値で「84」に最大の識別性能を示すことが分かった。これは、小学校2年生～3年生のレベルである。しかしながら、元々作成された11の版では、それぞれ測定対象とする集団の平均付近に最大の測定性能を有することが確認された。したがって、本研究の結果からは、統計的観点においても尺度が十分機能することが確認された。

キーワード：項目反応理論、テスト情報量、項目プール、日本語語彙、ロジスティック・モデル

1. 目的

日本語基礎能力テスト（平・前川・小野・林部・内田、1998）は、項目プールが完成してから5年、最後の項目が作成されてから11年という歳月が経過しており、項目の実用性を内容的に再検討する必要性が迫られていた。語彙テスト部分については、伊藤・佐藤・倉元(2003)よって国語教育的見地からその点が再評価が行われ、ほとんどの項目は現在でも問題なく使用できることが確認された。

本研究ではその分析結果を元に、項目プール、および、冊子形式となっている11の版の測定論的性質について分析を行い、その測定性能を再吟味することを目的とする。

2. 項目反応理論の概要

最初に、本研究が分析対象とする日本語基礎能力テストの項目プールが依拠している項目反応理論の概要についてまとめる。

2.1. ロジスティックモデル

項目反応理論（*Item response theory*）は、あるテスト項目に対しての被験者の正答反応の確率をあらわすモデルが基盤になっている。そして、項目の性質はそれをあらわす「項目パラメタ」、被験者についてはその尺度（テスト）が測定することを目的とする潜在特性をあらわす「被験者パラメタ」で表現される。すなわち、ある特定の項目に対するモデル上の正答確率が被験者パラメタの単調増加関数として定義される。その関数で定義される曲線を項目特性曲線（*Item Characteristic Curve*）と呼ぶ。

次式は2パラメタロジスティックモデル(2PLモデル)の項目特性曲線の定義式である。項目パラメタが困難度をあらわすb、識別力をあらわすaの2つであらわされている。なお、本稿で取り扱う項目プールはこのモデルを用いて尺度化されている。

$$P(u_{ij}=1|\theta_i)=\frac{1}{1+\exp\left(-Da_j(\theta_i-b_j)\right)} \quad (1)$$

ここでθは被験者iの潜在能力をあらわすパラメタである。またDは1.7の定数であり、jは項目の番号をあらわす。また、uは正答のとき1，誤答のとき0となる変数である。aは項目の識別力と呼ばれる項目

パラメタであり、被験者の潜在能力を文字通り識別する指標となるパラメタである。また b は項目の困難度をあらわす項目パラメタであり、潜在能力値 θ と同尺度上にある。被験者の潜在能力値が当該項目の b の値と等しいとき、正答確率が 0.5 となる。

2.2. 潜在特性値 θ の推定

一般に特性値が θ_i である被験者 i が n 項目に対し反応パタン $\mathbf{u}_i$ を示す確率は

$$P(\mathbf{u}_i|\theta_i) = \prod_{j=1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))^{1-u_{ij}} \tag{2}$$

である。u_{ij} は前述の通り正答のとき 1、誤答のとき 0 をあらわす変数であり、$\mathbf{u}_i$ はそれらによる反応パタンをあらわす。ここで反応パタンの $\mathbf{u}_i$ は調査で得られる既知のデータであることから、(7) 式を特性値 θ の関数とみなしたとき、これはいわゆる尤度関数となる。

$$L(\mathbf{u}_i|\theta_i) = \prod_{j=1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))^{1-u_{ij}} \tag{3}$$

ここで用いる θ の推定方法は被験者の反応パタンについてこの尤度の値を最も大きくするような θ の値を求めるもので、いわゆる最尤推定法と呼ばれる。最尤推定法は解析的に求めることができないので、数値計算によって求めることになるのだが、まず尤度関数の対数を取り、これを θ に関して微分し、

$$\frac{\partial \log L}{\partial \theta} = 0 \tag{4}$$

とおいて解を求める。なお 2PL モデルでは、

$$\sum_{j=1}^{n} a_j P_j(\theta_i) = \sum_{j=1}^{n} a_j u_{ij} \tag{5}$$

となり、識別力パラメタ a_j を重みとした得点が θ_i の十分統計量となる。

2.3. 情報関数によるテストの評価

最尤推定法による推定値 θ の標本分布は、標本数(項目数)が大きいときに漸近的に正規分布に従う。その標本分布の分散をもって、推定値がどの程度真の θ に近いものであるか評価するために用いる。すなわち分散が小さければ実際に得られた推定値 $\hat{\theta}$ が真の θ に近いことが期待され、また分散が小さいほどテストのもたらす情報は価値があると言える。推定値 $\hat{\theta}$ の漸近的な標本分布は $N(\theta, 1/I(\theta))$ になる。その平均は真の θ、分散は情報関数と呼ばれる $I(\theta)$ の逆数である。情報関数は尤度関数 L を用いて

$$I(\theta) = E\left[\frac{\partial \log L^2}{\partial \theta}\right] \tag{6}$$

とあらわされる。2PL の場合、任意の 1 項目についてこの情報関数を求めると、

$$I(\theta_i) = D^2 a_j^2 P_j(\theta_i)(1 - P_j(\theta_i)) \tag{7}$$

となり、これは項目ごとの情報量をあらわす指標となる。また n 項目のテストでは

$$I(\theta) = D^2 \sum_{j=1}^{n} a_j^2 P_j(\theta_i)(1 - P_j(\theta_i)) \tag{8}$$

となる。

2PL モデルの場合、情報関数は $\theta = b$ のとき最大値を取り、対象となる被験者にとって正答確率が 0.5 である項目を多く含むテストを実施した場合に精度が最も高くなる。また (7) より情報関数の最大値の大小は識別力に関係することがわかる。すなわち P(θ) が等しければ、識別力が高い項目を集めたほうが精度の高い測定が可能となる。これらのことから、例えば、学齢などの情報により被験者集団の潜在特性値分布があらかじめ予想がつくような場合、情報関数のこれらの性質を利用し、目的に応じたテストを作成することも可能であることがわかる。

2.4. 項目プール

テストで測定したい内容領域について、項目パラメタが既知である項目を集めた項目プールを準備することで、その中から必要な項目を選んで目的に応じたテストを自由に構成でき、また情報関数を基にした測定の文脈や測定対象に適した項目から構成されるテストを容易に作成することができる。

困難度や識別力などの項目パラメタは項目反応理論に基づいて同一尺度上に等化される。このことで項目プールの中の異なる項目群から構成された複数のテストの測定結果について相互に直接比較可能となり、測定対象に応じて項目プールの中から適切な

項目を逐次的に選び出してテストを実施する適応型テストが実現可能になる。

2.5. 適応型テスト

最近のコンピュータの発達と日常化はテストの様々な側面に大きな影響を与えている。例えば、従来のテストは紙のテスト冊子にペンで答える形式（P&Pテスト）が主な形式であったが、PCをテスト実施メディアとして用いるCBT（*Computer Based Test*）が実用化され始めている。コンピュータを積極的に利用することで、マルチメディアを利用したテストの作成の「可能性」、テスト実施過程の制御についても被験者ごとに解答状況に応じたきめこまかい制御ができる「可能性」がでてくる。

このような新しい可能性の一つとして適応型テスト（CAT:*Computer Adaptive Test*）がある。これは、当該被験者を測定するのに最適な項目を項目プールの中から逐次的に選択して実施する手順を繰返すテスト方式であり、結果的に被験者がそれぞれ異なる項目からなるテストを実施することになる。実施した項目セットが異なるにもかかわらずテストによる測定結果が同一の測定尺度上の値で表示されるためには項目反応理論を用いることが不可欠な条件である。どのようにして次に実施するのに最適な項目を選択するかが問題になるが、適応型テストでは実施済みの項目を元に被験者の潜在特性値を推定し、それに見合った項目を提示するという手続きを逐次繰返し、推定結果が安定したところでテストは終了となる。適応型テストの利点には、1) 全ての被験者に対して高い精度の測定を実施できる、2) 精度を落すことなく被験者1人あたりに実施する項目数を減らすことができ、そのためテストの実施時間を短縮できる、3) 難しい項目が続いて被験者にフラストレーションや不安を起こさせたり、易しすぎる項目が続いて飽きさせたりしない、4) 被験者の都合に合わせて柔軟にテストを実施できる（オンデマンド）、などがあげられる。適応型テストでは各個人ごとに実施される項目が異なり、項目数も異なる。しかし、項目反応理論の特性尺度上の値で被験者個人の測定結果を表示することによって被験者ごとに実施した項目が異なっても測定結果を相互に比較可能な同一尺度上に表示できるし、テスト情報量を用いて個人ごとに測定精度を確認することができる。

3. 日本語基礎能力テスト

本研究の分析対象となる日本語基礎能力テストの項目プールは平他（1998）により作成されたものであり、本研究で分析の対象とする「語彙理解力」の尺度と「漢字読み取り能力」の尺度からなる。項目プールは項目反応理論に基づき尺度化された小・中学生用のテスト（小野・繁桝・林部・岡崎・市川、1889）と高校生用のテスト（平・前川・小野・林部・米山、1995）を等化し、共通の尺度上にあらわすことで広い能力範囲に適応可能なものとして機能するようになった。なお、小・中学生用項目は、高校生用も含めて尺度化された項目プールを持つ語彙・漢字の2領域以外に助詞、文型など、全部で8領域の下位テストを持つ。

日本語基礎能力テストの各項目は元々11の版に編成されている。先述した項目反応理論の利点を生かすためには版毎の利用にこだわる必要はないが、被験者能力が特性値θであらわされること、その推定にはコンピュータとそのためのプログラムが必要であることなど、専門知識を持たない一般のユーザーにはやや利用しにくい面がある。そこで、本研究では既にセットとなっている版毎の測定性能の特徴に関しても、分析を加えることとする。

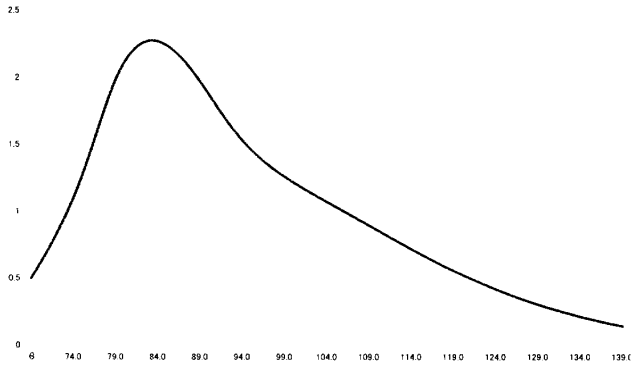
4. 項目プールの測定性能

平他（1998）によると、語彙、漢字の項目プールは標準化を目的とした線形変換が施されており、被験者の能力値がほぼ平均100、標準偏差15となっている。

4.1. 語彙

伊藤他（2003）により、平他（1998）の日本語基礎能力の語彙に関する項目プールについて国語学的見地

図1. 語彙項目プールのテスト情報量



からの再検討が行われた。その結果、もともと471項目あった項目プールのうち、除外される項目が1項目、「レベルC」の評価を受けた項目が26項目となった。そこで、本研究では、「レベルA」、および、「レベルB」の評価を受けた444項目のみを取り出し、そこから測定性能を論じることとする。

上記のような手続きで定義した444項目から成る項目プールについて情報関数を計算したところ、その結果は図1のようであった。

項目全体として$\theta = 84$付近にピークがあり、この項目プールを用いて潜在能力値を測定する場合、潜在能力値が84付近の被験者に対して最も精度が高い測定を行うことができることを意味する。

5. 版別分析結果

5.1. 小学校1年生版～中学生版

図2.1～2.7は各版ごとに調査対象として想定されている学年の被験者の$\hat{\theta}$の分布とそれぞれの版の情報関数を重ね合わせて図示したものである。なお、それぞれの被験者データは項目プール作成過程において収集されたものである。この図から、各版がそれぞれどのようなθの値の範囲において効果的な測定が可能であるかがわかる。

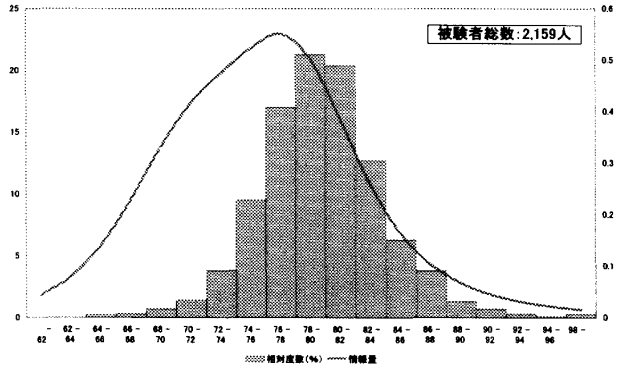


図2・1 小学校1年生版

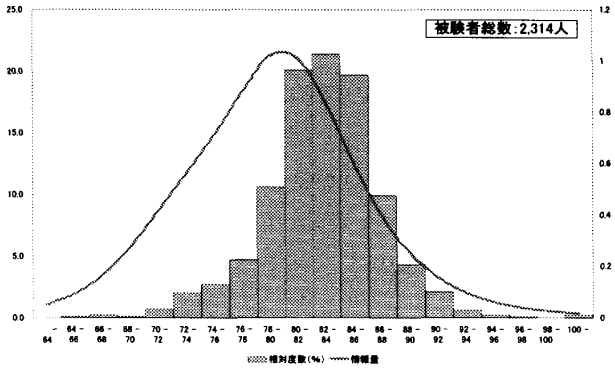


図2・2 小学校2年生版

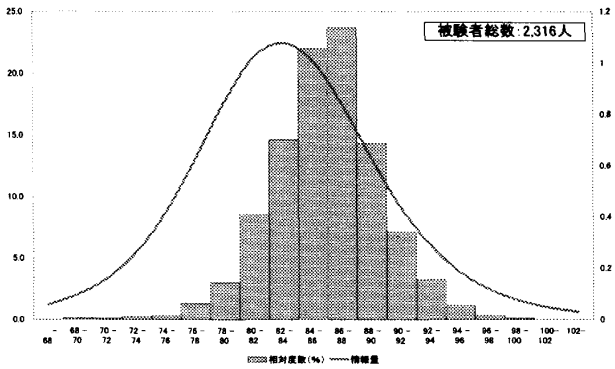


図2・3 小学校3年生版

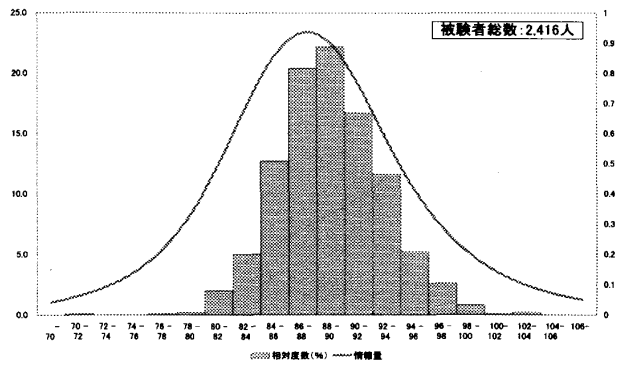


図2・4 小学校4年生版

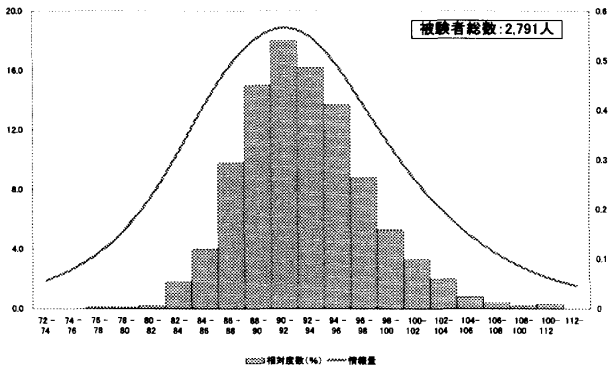


図2・5 小学校5年生版

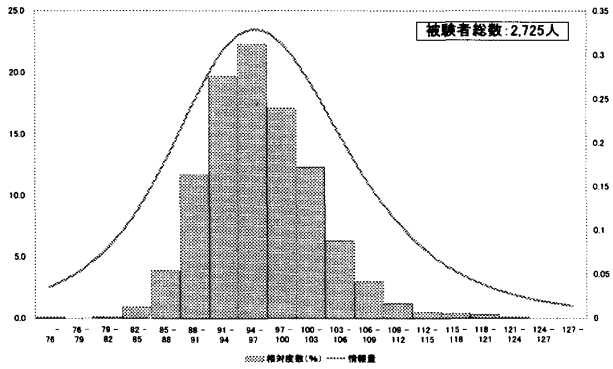


図2・6 小学校6年生版

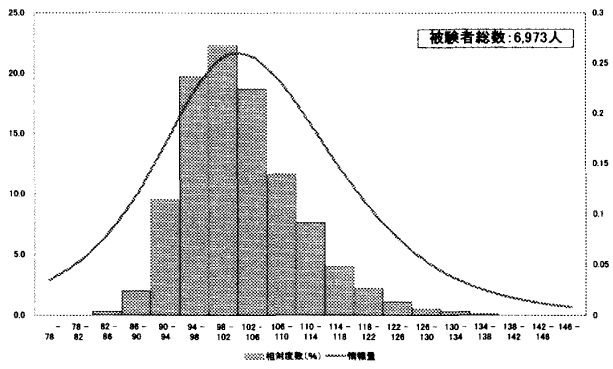


図2・7 中学生版

例えば、図2.7の小学校1年生版を見ると、当該する被験者の潜在特性値の平均は79.2、一方情報関数のピークは78付近にあり、ほぼ等しい値とみなすことができる。すなわち、実際の小学校1年生の語彙力を測定するテストとして十分機能するであろう。一方、情報関数の分布形を見ると、他の学年の小学生版および中学生版と比較するとなだらかな分布形となっており精度は多少劣るが、その分散は5.2となっていることから、測定用具として十分に機能を果たすテストであると言える。

他の小学生版および中学生版の情報関数を見ると、潜在能力の推定値の分布とほぼ重なることがわかる。また、分布形のピークを中心としたある範囲において情報量が集中して大きいことから、それぞれの学年相応の語彙理解能力を持つ被験者に対してはある程度の制度で機能すると考えられる。

5.2. 高校生91年度版～92年度版A・B・C

図3は前述の図2.1～2.7同様、高校生版として作成された語彙能力テストの被験者の潜在特性の推定値の分布および情報関数をあらわしている。高校生用の4つの版は全て高校生程度の語彙理解力の測定を目的として作成されたものであるので、各版の被験者を当該年齢の被験者集団として一つの分布にまとめてある。なお、情報関数については各版について分析を行い、その結果を載せている。

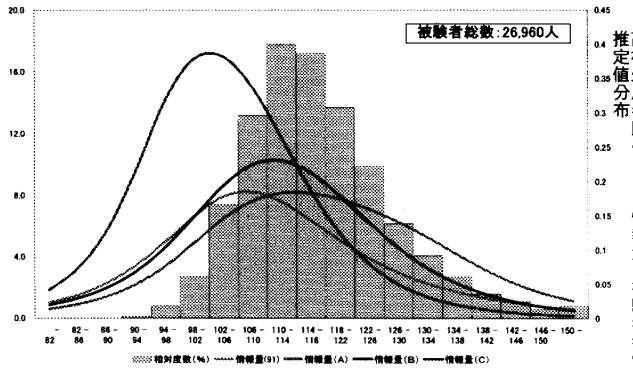


図3　高校生用各版のテスト情報量と被験者特性の推定値分布

それぞれの情報関数のピークは91年版が108、92年版Aが114、92年版Bが110、92年版Cが104となっており、92年度版Cは中学生版に近いところに分布のピークがある。一方で、92年度版Aはその分布のピークの山は低いが高い潜在特性値を持つ被験者を識別することに精度が高いテストである。高校生全体の潜在特性の推定値$\hat{\theta}$の分布と比較すると、この分布の平均は117.2である。したがって項目プール作成時点においては、高校生の日本語語彙能力を測定する用具として、91年度版および92年度版A,Bが適しており、92年度版Cは中学生レベルの語彙力を持つ被験者に適していたと思われる。

伊藤他(2003)によると、1) 92年度A版は主として進学校での使用を想定したものであり、使用面に関して通時的に安定し、かつ質の高い言語使用能力に関連した語彙項目を中心に構成され、2) B版は社会的生活での使用を想定した項目の比率が高く、3)C版は日常に必要と思われる基礎的な語彙の項目を多く含み、日常的な言語使用能力の測定という目的に合致している版であることが指摘されている。このことから、C版についても目的によっては高校生レベルの基礎的な語彙力を調査する上では有効な版であるといえる。

6. まとめと課題

以上、「日本語基礎能力テスト」について項目プールの情報関数に焦点を当てその測定精度について論じてきた。項目プール全体としての測定精度、各版ごとの当該被験者の潜在能力値とその測定精度は語彙テストとして実用可能なものであることがいえる。本研究では、伊藤他（2003）による項目ごとの評価によりレベルCと判定された項目を除き分析を行ってきたが、その特殊性を考慮した上で各版に取り入れることで、目的に沿ったテストを作成することが可能であろう。

今後の課題として、専門知識を持たないユーザが「日本語基礎能力テスト」を実際に利用しやすいような簡易得点化を行うことも必要となるであろう。また、CBTによる適応型テスト作成の試みなどの課題も残されている。

本研究で「日本語能力テスト」の語彙に関する項目の測定精度の比較として取り上げた被験者データは11年前のものであり、各版ごとに想定した当該年齢の被験者の潜在能力値が、現時点でそのまま適応できるものであるかどうか検討の余地がある。当該年齢の学生の語彙能力の経時的変化を捉えることも含め、今後、更なる研究が進むことが望まれる。

文献

伊藤博美・佐藤洋之・倉元直樹 2003. 日本語基礎能力の特性(1) －国語教育から見た項目内容の評価－

(印刷中)
小野　博・繁桝算男・林部秀雄・岡崎　勉・市川雅教・木下ひさし・牧野泰美 1889. 日本語力検査の開発、昭和61-63年度科学研究費報告書、試験研究(1)課題番号618103(未公刊).
平　直樹・前川眞一・小野　博・林部英雄・米山千佳子 1995 高校生程度の日本語能力テストの開発、－語彙理解テスト・漢字読み取りテストの尺度化－、教育心理学研究 43.6-73.
平　直樹・前川眞一・小野　博・林部英雄・内田照久 1998. 日本語基礎能力テストの項目プールの作成、大学入試センター研究紀要、No.28,1-12.

Properties of the Broad-Range Japanese Fundamental Language Skills Test II :Statistical Analysis of Vocabulary Subscale Based on Item Response Theory

Hiroyuki Satou, [*] **Hiromi Itou,** [**] **Naoki T. Kuramoto** [***]

[*] *Education Division, Tohoku University Graduate School of Educational Infomartics*
[**] *Tohoku University Graduate School ofLetters; Honjo Akita Prefectural High school*
[***] *Admission Research Center of Tohoku University;*
Education Division, Tohoku University Graduate School of Educational Infomartics

The purpose of this study is to reexamine the statistical properties of vocabulary items, which consists of the subscale for the Broad-Range Japanese Fundamental Language Skills Test (BRJFLST) developed by Taira et al. (1998). Ito, Sato and Kuramoto (2003) showed the content validity of the item pool, even after 10 years since item construction. However, the present analysis excluded Level C items, which had been evaluated to be utilized with care. Finally, 444 items were included in the present study.

According to the test information curve, the item pool as a whole showed the largest discriminating power around the theta scale point of 84, which corresponded to the level of second or third grade of elementary school pupils. Each of the original 11 testlets showed the peak of measurement efficacy at the average level of its target population. The present study showed that the scale was also effective from statistical viewpoints.

Key words: Item Response Theory, test information, item pool, Japanese vocabulary, logistic model