教育心理学研究, 2003, 51, 413−424

日本語語彙理解力テストの妥当性についての検討

―― ルールスペース法を用いた認知論的分析 ――

倉元直樹* スコット寿美** 笠居昌弘***

Tatsuokaのルールスペース法（RSM）は，テストから学力診断的な情報を取り出すことを目的に，項目反応理論に基づいて考案されたテスト分析法である。ルールスペースの2つの軸は，テストで測定される特性を表すθと反応の非典型性を表すζで示される。「項目に正答するには，必要なアトリビュートを全て備えていること」という測定パラダイムに基づき，テスト結果がルールスペース上の知識ステートとして説明される。アトリビュート，および，それらと項目との対応関係を示す照応行列が適切に作成されるならば，個々の解答パターンは確率に基づいて知識ステートに分類することが可能であり，テスト結果から診断的情報を抽出することができる。本研究では，日本語語彙理解力テスト（平他, 1998）のデータを分析し，被験者の多い知識ステートを樹状図で表すことによって，テストの妥当性について検討した。30項目に対する1,500名分のデータについて，2名の専門家の協力を得て作成したQ行列を用いて分類したところ，80.5%の分類成功率を得た。また，それを用いた樹状図による分析では，日本語語彙理解力テストの妥当性が検証された。

キーワード：ルールスペース法，アトリビュート，照応行列，知識ステート，日本語語彙理解力テスト

問　題

テストは，人や機関が何らかの評価や意志決定を行う場合にその判断に根拠を与えるための大切な道具として用いられている。その一方で，「テストは一体何を測っているのか」という測定の妥当性に関わる問いは，素朴な疑問でありながら，科学的・実証的基盤からこれに明瞭な回答を与えるのは意外に難しい。本研究は項目反応理論（以後，IRTと略記）に基づいて尺度化されたテストによる測定結果に対し，個人の認知的状況に対する診断情報を提供することを目的として開発されたルールスペース法（Rule Space Methodology, 以後, RSMと略記）（Tatsuoka, 1990；倉元・龍岡, 2001）[1]を用いて測定の妥当性について検討する[2]。

本研究で分析の対象とするのは，日本語語彙理解力テストである。1980年代後半から1990年代前半にかけて小・中学生用，高校生用の項目がそれぞれ作成され，項目反応理論の2パラメタモデルによって別々に尺度化された後，1つの項目プールとしてまとめられたものである（小野・繁桝・林部・岡崎・市川・木下・牧野, 1989；平・小野・林部, 1992；平・小野・前川・林部・米山, 1995；平・前川・小野・林部・内田, 1998）。現時点で470問から成る項目プールを持つ（伊藤・佐藤・倉元, 2003）。最新とは言えないが，適切な項目を選定してテストを構成すれば，被験者の日本語語彙理解力を評価することは可能と考えられる。しかしながら，ア・ポステリオリに測定の意義自体が検証されてきたとは言いがたい。すなわち，日本語語彙理解力の評価という測定意図は明白である

* 東北大学アドミッションセンター（旧姓：平）
〒980-8577 宮城県仙台市青葉区片平2-1-1 東北大学アドミッションセンター ntkuramt@mail.tains.tohoku.ac.jp
** マクロメディア（旧姓：齋藤） Macromedia Inc., 101 Redwood Shores Parkway, Redwood City, CA 94065, USA
hscott@macromedia.com
*** ノースイースタンイリノイ大学評価テスト部門
Northeastern Illinois University, 5500 North St. Louis Avenue, Chicago, IL 60625-4699, USA
M-Kasai1@neiu.edu

[1] 現在，まとまった形でRSMを紹介した文献は存在しない。日本語では倉元・龍岡（2001）がRSMに関する唯一の概説的な紹介論文である。本稿におけるRSMの用語の日本語訳は，原則として倉元・龍岡（2001）の付表に基づく。

[2] Messick (1992) は，「テスト遂行中に示される行動」の類推には「応答や行動の一貫性の証拠が必要であり，内容についての判断だけの問題ではない」としているが，具体的な方法論の考案は困難である。一方，Messick (1992) は，「専門家の判断が内容と形式の適切性を証明するための重要な材料であることは明らか」とも述べている。エキスパート判断を重視するRSMの方法論は，ミクロな認知論的観点に立った解答プロセスのモデルを含む。まさに，上記のMessickの観点に立脚したテストの妥当性検証法と位置づけられる。

が，実際にテストがそれを測っているのかどうかという点を実証する手立ては得られていなかった。なお，この点に関しては，本研究の素材である日本語語彙理解力テストのみならず，通常の評価ツールにとって不可避的に共通の問題と言える。

一方，RSMは，学習者の認知的過誤を診断してIRTを用いて尺度化されたテストの結果の裏にある認知的過程や知識を調べ，テスト結果から診断的情報を抽出することを目的に開発された測定論的方法である（Tatsuoka, 1995）。理論の概要は以下の通りである[3]。エキスパートの評定によって作成された「項目×アトリビュート（attributes）」の照応行列（Q行列：incidence matrix, Q matrix）（Tatsuoka, 1990）を元に，項目の正誤パターンから，当該のテストで測定される解答者の認知的状況「知識ステート（knowledge state）」を同定し，診断的な情報を統計的にフィードバックするテスト解析法である。

RSMにおける「アトリビュート」とは，テスト項目に正答するために必要となる認知的要素のリストである。すなわち，特定の項目が測定している内容要素を具体的に記述したものがアトリビュートであると考えてよい。さらに，個々の項目解答に対するアトリビュートの要不要のパターンを「1」，「0」の2値行列の形で表示したものが照応行列（以下，Q行列と記す）である。「被験者がある特定のテスト項目に正解するためには，そのテスト項目に含まれるアトリビュートを全て習得していなければならない」ということがRSMの前提である。例えば，TABLE 1は，3項目に対して2つのアトリビュートが設定されている場合である。当該項目に正答するために，1がそのアトリビュートが「必要」，0が「不要」であることを表す。このケースでは，アトリビュートa_1，a_2の両方を習得している状態（$ks_{1,1}$），a_1のみ習得（$ks_{1,0}$），a_2のみ習得（$ks_{0,1}$），a_1，a_2のいずれも習得していない状態（$ks_{0,0}$）の4種類の知識ステートがこれらの項目に関してあり得るパターンである。$ks_{1,1}$の状態にある被験者が解答を行った場合には3問全てが正解となるが，$ks_{1,0}$では項目1，$ks_{0,1}$では項目3のみ正解，$ks_{0,0}$ならばいずれも不正解，というのがモデル上の項目反応パターン（ideal response pattern）となる。理論的にはそれ以外のパターンは生起しないはずであるが，確率的にモデルからの逸脱（slip）が起こるため，実際には観測されることがある。

TABLE 1 Q行列（2×3）と知識ステート，正誤パターンの例

	アトリビュートa_1	アトリビュートa_2
項目1	1	0
項目2	1	1
項目3	0	1

知識ステート$ks_{1,1}$：「a_1習得，a_2習得」 → 全問正答
知識ステート$ks_{1,0}$：「a_1習得，a_2未習得」→項目1正答
知識ステート$ks_{0,1}$：「a_1未習得，a_2習得」→項目3正答
知識ステート$ks_{0,0}$：「a_1未習得，a_2未習得」→全問誤答

知識ステートが表示される空間が「ルールスペース（Rule Space）」である。ルールスペースはθとξ（および，$\xi_1, \xi_2, \cdots, \xi_l$）を次元として持つ。$\theta$はIRTにおける通常の被験者パラメタであるが，$\xi$は反応の非典型性を表す標準化された指標であり，以下の(1)式によって表される。

$$\xi = f(\boldsymbol{x}) / \sqrt{\mathrm{var} f(\boldsymbol{x})} \tag{1}$$

ただし，

$$f(\boldsymbol{x}) = -\sum_{j=1}^{n}(P_j(\theta) - T(\theta))x_j + \sum_{j=1}^{n} P_j(\theta)(P_j(\theta) - T(\theta)) \tag{2}$$

$$\mathrm{var} f(\boldsymbol{x}) = \sum_{j=1}^{n} P_j(\theta)(1 - P_j(\theta))(P_j(\theta) - T(\theta))^2 \tag{3}$$

$$T(\theta) = \frac{1}{n}\sum_{j=1}^{n} P_j(\theta) \tag{4}$$

である。ここで，nは項目数，x_jは項目jに対する反応の正誤（正答ならば$x_j=1$，誤答ならば$x_j=0$），$P(\theta)$はθの条件付正答確率を表す。(2)式から明らかなように，ξは被験者にとって相対的に難しい問題に正答するほど正に大きな値を取り，易しい問題に正答するほど負に大きな値を取る。また，局所独立の仮定の下ではθとξは無相関である（Tatsuoka, 1985）。

ルールスペース分析（Rule Space analysis）を行う場合，最初に吟味する必要があるのは，Q行列の適切性である。それは，理論的に存在し得る知識ステートに対して，実際にデータとして得られた被験者の項目反応パターンがどの程度分類可能であるか，という点から評価される。

Q行列が不十分な場合には，アトリビュートの取捨選択が行われることがある。個別のアトリビュートの評価は，それぞれの被験者に対して計算されるアトリビュート習得確率（attribute mastery probability）と合計得点（またはθ）との相関係数，あるいは，被験者が分類された知識ステートを特徴づけるアトリビュート習得パターン（attribute mastery pattern）と合計得点（または

[3] 詳細は，倉元・龍岡（2001）を参照のこと。

θ) との点双列相関係数の値が目安として用いられる。通常の項目分析と同様に，適切に機能しているアトリビュートであれば，比較的大きな正の値が得られるはずである。また，分類成功率を上げるために，交互作用アトリビュート (interaction attribute) を加えることも可能である。

満足できる結果が得られたならば，観測されたデータが多い知識ステートをθとξの次元で表現される空間にプロットすることが可能になる。それらを樹状図の形でつないでネットワーク化していくことにより，アトリビュート習得プロセスのモデルを示すことができる。また，それを被験者個人の成績情報と付き合わせることにより，習得されていないアトリビュート，さらに，その中で次に学習すべきアトリビュートが何であるかを診断することが可能になる。また，アトリビュート習得確率を用いて各項目の困難度（正答率）を予測することも可能である。

以上が，RSM を用いたテストの診断的利用法の概略である。現在まで，米国の ETS で作成，実施されている TOEFL (Kasai, 1997 ; Scott, 1998)，SAT Ⅰの数学 (Tatsuoka, 1995)など，普及した大規模テストに対して RSM を適用した研究例がある。

本研究の目的は，RSM の方法論をテストの妥当性検証に応用しようとするものである。テスト項目に正答するために必要とされるアトリビュートには，テストのスペックによって定まる本質的にテストの測定目的に沿った能力と同時に，テストワイズネスに類するような測定目的以外の要素も含まれている可能性がある。測定目的に合致したアトリビュートを習得することにより，成績が飛躍的に伸びる場合にはそのテストの妥当性は高いと言える。逆に，非本質的なアトリビュートの習得によって成績が大きく伸びるようなテストは妥当性に疑念が持たれても仕方がないであろう。

本研究では，以上のような観点から RSM を用いて日本語語彙理解力テストの妥当性について，検証を試みる。

方　　法

データ

本研究においてルールスペース分析の対象とするデータは，項目プールの中でも平他 (1992) で作成，平他 (1995)，および，平他 (1998) で尺度化された高校生用語彙理解力測定項目73項目から抽出した30項目である。高校生による18,293名分の正誤データのうち，1,500名分の反応を無作為抽出して用いた。なお，本稿で用いられている項目番号は平他 (1992) に基づくものである[4]。

項目形式は五肢択一式である。すなわち，幹に示された語の意味に最も近い意味を持つ正答を，５つの選択肢の中から１つ選ぶ方式となっている。

再尺度化

分析の準備として，抽出した1,500名分のデータを用いて検討の対象とする30項目を BILOG (Mislevy & Bock, 1989) を用いて２パラメタ・ロジスティックモデル（池田, 1977 ; 南風原, 1991）による尺度化を行った。その結果，平他 (1998) で得られた項目パラメタとほぼ同等とみなせる（パラメタ a の相関係数は r＝.82，b は r＝.98）推定値が得られた。

予備分析

解答過程の分析　３名の日本語母語話者に対して，本研究で用いる30項目への解答を求めた。その際，頭に思い浮かんだことを全て発話するように教示し，それをオーディオ・テープに記録した。その発話記録（思考発話プロトコル：think aloud protocol）を予備分析のアトリビュート作成の参考に用いた。なお，被験者のうち２名は30代の女性，１名は男子高校生であった。

アトリビュート，Q行列の作成　３名の共同研究者が予備分析用のアトリビュートを作成した。その後，３名がそれぞれ独立にアトリビュートの要不要を評定して，一致度をチェックした。

解答のルールスペースへの分類　解答の正誤パターンをルールスペース上の知識ステートに分類するために，RSM の分析プログラム Pmain(K. K. Tatsuoka, C. M. Tatsuoka, & Varadi, 1995)[5]を用いた。

本分析

アトリビュート，Q行列の改良　２名の専門家（高校国語教師）の協力により，アトリビュートとQ行列を改良した。反応のルールスペースへの分類までのプロセスは，予備分析と同様である。

アトリビュートの評価　アトリビュート習得確率，および，アトリビュート習得パターンと合計得点との関係により，Q行列の評価を行った。

---

[4] ソフトウェアの制約のため，項目，被験者ともに一部データの抽出が必要となった。なお，項目の統計的な性質の保持のため，分析の対象とした項目の具体的な内容は一部を除いて詳らかにしない。

[5] ルールスペース分析用のソフトウェアは，現在，一般向けにリリースされていない。関心のある読者は龍岡菊美教授（現コロンビア大学，KikumiT@exchange.tc.columbia.edu にご連絡いただきたい。

樹状図によるアトリビュート習得プロセスの検討 分類された被験者の多い知識ステートをルールスペース上の樹状図に表し，アトリビュートの習得，未習得の状況と成績との関係からテストの妥当性について検討を加えた。

結　果

予備分析

アトリビュート，Q行列の作成 作成されたアトリビュートの内容は，TABLE 2に示す通りである。主に項目の形式的な性質から「幹：漢字」，「幹：ことば(語)の意味」，「幹：品詞」，「正答選択肢」，「まよわし(誤答選択肢)」という5カテゴリーに属する26のアトリビュートを作成した。

TABLE 2 予備分析におけるアトリビュート一覧

〔幹：漢字〕*
1．**難しい漢字だけが使われている項目** → 本分析に使用
2．**難しい漢字と易しい漢字が使われている項目** → 本分析に使用
3．**易しい漢字のみが使われている項目** → 本分析に使用
4．**ひらがなだけの項目**（不採用） → 本分析に使用
5．漢字の意味の一義性（不採用）
〔幹：ことば（語）の意味〕
6．日常会話によく出てくる語
7．書きことばにしか使用されない語
8．漢語表現
9．和語表現
10．比喩表現
〔幹：品詞〕
11．動詞を含む
12．形容詞・形容動詞を含む
13．名詞を含む
14．簡単に動詞化，または，形容動詞化できる名詞を含む
〔正答選択肢〕
15．幹と同じ漢字（不採用）
16．**幹と同じ漢字があるが，読み方が異なる** → 本分析に使用
17．**幹の漢字の意味から連想が容易である** → 本分析に使用
18．**幹の漢字の意味から連想が困難である** → 本分析に使用
〔まよわし（誤答選択肢）〕
19．幹と同じ漢字
20．幹と同じ漢字があるが，読み方が異なる
21．幹と同じ漢字はないが，熟語全体の意味からの連想で行き着くもの
22．一部の漢字の連想で行き着くもの
23．同音異義語の意味になっているもの（不採用）
24．正答と非常に近い意味を持つもの（不採用）
25．Plausible Distractor（もっともらしいまよわし）の存在（理由を明記）（不採用）
26．文脈の重なり

＊易しい漢字とは，基本的に小学校段階で習得が必要とされるもの

個々の項目に対するアトリビュートの要不要に関して3名の判断が一致した場合には，評定結果がそのままQ行列の要素として採用されたが，3名の判断が分かれた場合には，原則的に2名が必要と評定したものを「1（必要）」，それ以外を「0（不要）」と評定した。

さらに，本研究の項目に対する必要度や評定の信頼性が著しく低かったアトリビュートを不適切と判断して不採用とし，20のアトリビュートを分析に用いることとした。採用されたアトリビュートにおける評定の一致率は，3名の評定者の評定結果を表す2値データを変数とする行列におけるアルファ信頼性係数を指標としてα＝.59～1.0であった。

予備分析結果 Pmainを用いた分析の結果，知識ステートへの分類に成功したのは1,500名中僅かに629名分の反応であり，分類成功率は42%と低かった。したがって，Q行列，または，アトリビュートそのものの内容を再検討する必要があることが分かった。

本分析

専門家によるアトリビュート，Q行列の改良 分類成功率を上げるためには，Q行列の改良，アトリビュート自体の改良，という2つのアプローチがあり得る。本研究においては，予備分析に関わった3名の共同研究者の中に国語学や高校生の日本語能力についての専門家がいなかったことが，分類成功率が低かった最大の原因と考え，2名の高校国語教師の協力によりアトリビュート自体を作成しなおすこととした。

2名の専門家との協議の結果，予備分析で作成されたアトリビュートについては，TABLE 2,3に示す通り，幹の語に用いられている漢字に関するもののうち4つ，正答選択肢に関するもののうち3つを残すこととし，それ以外は全て作成しなおすこととした。なお，予備分析から引き継いだアトリビュートについては概ね妥当であると考え，再評定は行わずに予備分析のQ行列の該当部分をそのまま用いることとした。

次に，幹の語に関連して設定していた2つのカテゴリーを廃し，代わりに「高校生にとっての幹の難しさ」を評定するカテゴリーを設けることとした。また，誤った推論の結果，誤った選択肢にたどり着くプロセスをモデル化し，各項目に対してそれを当てはめることを試みることとした。

さらに，アトリビュートの表現を「習得することによって正答に達する」形式に統一することとした。

高校生にとっての幹の難しさ 「高校生にとっての幹の難しさ」については，幹に掲げられた語の性質が

TABLE 3　本分析におけるアトリビュート一覧

〔幹の漢字〕
1．難しい漢字（中学校以上）だけが使われている項目に正しく答えられる
2．難しい漢字と易しい漢字が使われている項目に正しく答えられる
3．易しい漢字（小学校の教科書に掲載）だけが使われている項目に正しく答えられる
4．ひらがなだけの項目に正しく答えられる

〔正答選択肢〕
16．幹と同じ漢字があるが、読み方が異なる正答を正しく選べる
17．幹の漢字の意味から連想が容易な正答を正しく選べる
18．幹の漢字の意味から連想が困難な正答を正しく選べる

〔高校生にとっての幹の難しさ〕
A．高校生が日常ほとんど見聞きすることがない言葉を知っている
B_1．教科書、テレビ、雑誌等改まった場面では使用される硬いことばを知っている
B_2．テレビ、雑誌、マンガ等では使用される若者文化となった硬いことばを知っている
B_3．大人の世代が日常的に使う古いことばを知っている
C．高校生が日常使用することがあることばを知っている
D．高校生にとっては状況に馴染みがうすく、ニュアンスが分かりにくいことばを知っている
E．文語で使われることばを知っている（不採用）
F．口語で使われることばを知っている（不採用）

〔誤答への推論〕
L．正答選択肢と意味や状況が重なる誤答との区別がつく
M．漢字の意味、形によって類推される誤答との区別がつく
N．その語を含む使用頻度の多い熟語から類推される誤答との区別がつく
O．同じ、あるいは、類似した漢字を用いる熟語から類推される誤答との区別がつく
P．語から想起されるニュアンスを共有した誤答との区別がつく
Q．語尾を共有している誤答選択肢との区別がつく（不採用）
R．同じ音の語、同音異義語との区別がつく

〔交互作用〕*
1-B_1
3-18
4-C
17-P

＊2つのアトリビュート双方が「1」の場合のみ「1」、それ以外は「0」

高校生にとってどのようなものか，2名の専門家が1項目ずつディスカッションによる検討を行って判断した。その結果 TABLE 3 に示す8つのアトリビュートが必要であるという結論に達した。さらに，母集団での正答率を参考として，個々の項目に関する評定を行った。

誤答への推論　「誤答への推論」については，幹と選択肢の双方に用いられている語を比較検討し，被験者がどのような推論の結果として誤答を選択する可能性があるか，検討を行った。その結果，TABLE 3 に示した7つの推論プロセスを考えるべきであるという結論に達した。さらに，個々の項目について推論プロセスをモデル化し，それを図示して検討を行った。2つの項目に関する検討例を FIGURE 1 に示す。

項目「60．無常」は，⑤が正答である。しかしながら，正答との「L．意味や状況の重なり」から②を選択する可能性があると考えた。また，「P．幹の語から想起するニュアンスを共有」している点でも②を選択する可能性があるとした。さらに，「無情」という「R．同音異義語」の存在から①，④を，「非常」という「O．幹と類似した漢字を用いる熟語の類推」から③を選択する可能性があると考えたものである。

項目「62．息災」は③が正答である。しかしながら，「息」という「M．幹と同じ漢字からの類推」で④，⑤を，「無病息災」という「N．幹の語を含む使用頻度の多い熟語から類推」で①，②，⑤を，「災害，災難」という「O．幹と同じ漢字の熟語」により①，②，⑤を選択する可能性があると考えたものである。

「誤答への推論」に関しては，1つの選択肢に対する

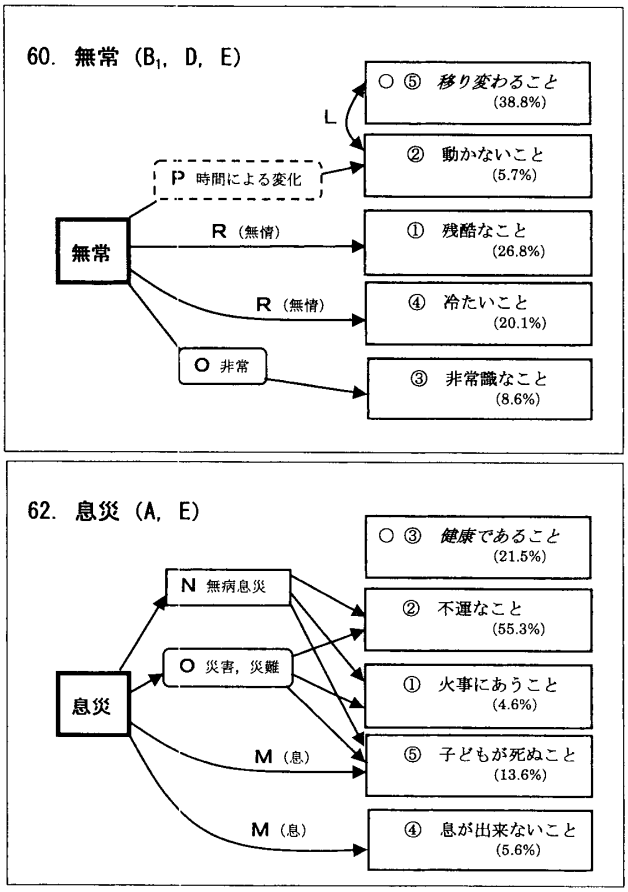

FIGURE 1「誤答への推論」に関するモデル例

複数のプロセス，1つの推論プロセスによる複数の誤答選択肢への到達の重複をそれぞれ許した上で，1つの推論プロセスによって選択される選択肢の合計選択率が20%以上のものを必要なアトリビュートであると認定した。すなわち，「60. 無常」に正答するために習得が必要なアトリビュートは，エキスパート判断によって選ばれた「B_1，D，E」に加えて「R（選択率合計46.9%）」であり，「62. 息災」に正答するためのアトリビュートは「A，E」に加えて「N(73.5%)，O(73.5%)」となる。

アトリビュートの取捨選択　再び，Pmainを用いて被験者の反応パターンを知識ステートに分類した。分類成功率の向上を期してアトリビュートの取捨選択を行い，数回の試行錯誤を繰り返した。その結果，3つのアトリビュートが除外され，4つの交互作用項が加えられた。最終的に，TABLE 4のQ行列による分析が最適と判断された。

Q行列の評価　TABLE 4のQ行列を用いた分析の結果，1,500名中1,208名の反応の分類に成功した。分類成功率は80.5%であった。ルールスペース分析の先行研究がない日本語語彙理解力という測定分野の特性，本研究の目的が被験者個人の診断ではなくてテストの妥当性の検討である，という2点を鑑みると，十分満足できる結果と言える。

合計得点のアトリビュート習得確率への回帰による説明率は$R^2=.93$，アトリビュート習得パターンへの回帰による説明率は$R^2=.89$であった(TABLE 5参照)。したがって，本分析に用いたアトリビュートとQ行列はこの点からも十分に機能していると考えられる。また，個々のアトリビュートと合計得点との相関関係も全て

TABLE 4　本分析のQ行列

	幹の漢字				正答選択肢			高校生にとっての幹の難しさ						誤答への推論						交互作用			
項目	1	2	3	4	16	17	18	A	B_1	B_2	B_3	C	D	L	M	N	O	P	R	1-B1	3-18	4-C	17-P
3	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	1	0	0	1
14	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0
15	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
16	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
23	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
34	1	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0
36	0	0	1	0	0	0	1	0	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0
37	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0
38	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
39	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
40	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
41	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
42	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
43	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0
54	0	1	0	0	1	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1
55	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
58	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0
59	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
60	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0
62	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
63	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0
64	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0
65	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1
66	0	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
69	1	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
71	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
72	0	1	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0
73	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0

TABLE 5 アトリビュート習得確率，アトリビュート習得パターンと総得点

	AMPr 平均	総得点と AMPr の相関	総得点と AMPt の点双列相関
1. 難しい漢字（漢字）	0.93	0.39	0.33
2. 難易漢字（漢字）	0.87	0.46	0.40
3. 易しい漢字（漢字）	0.97	0.26	0.22
4. ひらがな（漢字）	0.94	0.41	0.36
16. 幹と同じ漢字（正答）	0.52	0.46	0.37
17. 連想容易（正答）	0.95	0.34	0.25
18. 連想困難（正答）	0.99	0.25	0.19
A. 日常見聞せず（語の難度）	0.67	0.25	0.19
B_1. 硬いことば（語の難度）	0.91	0.42	0.38
B_2. 若者文化ことば（語の難度）	0.87	0.49	0.43
B_3. 古いことば（語の難度）	0.83	0.46	0.37
C. 日常使用する（語の難度）	0.99	0.22	0.20
D. 馴染みがうすい（語の難度）	0.71	0.49	0.44
L. 状況の重なり（誤答）	0.86	0.39	0.30
M. 漢字による類推（誤答）	0.57	0.36	0.22
N. その語を含む熟語（誤答）	0.36	0.44	0.32
O. 類似漢字の熟語（誤答）	0.78	0.43	0.38
P. ニュアンスの共有（誤答）	0.94	0.30	0.26
R. 同じ音の語（誤答）	0.52	0.44	0.38
1-B_1.（交互作用）	0.74	0.48	0.43
3-18.（交互作用）	0.90	0.31	0.29
4-C.（交互作用）	0.89	0.47	0.44
17-P.（交互作用）	0.77	0.35	0.33
		説明率 $R^2=.93$	説明率 $R^2=.89$

AMPr：アトリビュート習得確率，AMPt：アトリビュート習得パターン

正の値であり，全てのアトリビュートが機能していたと考えられる。

また，9つのアトリビュートの習得確率の平均値が90%を超えており，全体的には習得状況は高かったと言える。しかしながら，「N. その語を含む使用頻度の高い熟語から類推される誤答との区別がつく」が習得確率36%と低かったのを始め，「16. 幹と同じ漢字があるが，読み方が異なる正答を正しく選べる」，「R. 同じ音の語，同音異義語との区別がつく」，「M. 漢字の意味，形によって類推される誤答との区別がつく」という3つのアトリビュートの平均習得確率は50%台と低かった。

A～Cの5つのアトリビュートについては，Aが最も難しく，Cが最も易しく，B_1～B_3がその中間の難易度であることを想定して設定されたものである。アトリビュート習得確率の平均値は，その想定に沿った結果となっている。

アトリビュート習得確率による正答率推定 RSM分析のモデル上，当該項目に必要とされるアトリビュート習得確率平均値の積が正答率となる。

データの母集団と本研究で用いられたサンプルの正答率の差は最大3.5%，相関係数は $r=.998$ であった。そこで，母集団正答率と推定正答率を比較したところ，最大で50.0%の違いが見られた。実際の正答率と推定正答率との相関係数は，$r=.77$ であった。全体としては，実際より低く推定されている項目が多い（FIGURE 2参照）。実際の平均正答率が60.7%であるのに対し，推定正答率の平均値は49.6%であった。30項目中19項目は12%以内の誤差であり，比較的精度よく推定されている。一方，残りの11項目に19.5%以上の大きな誤差が見られた（TABLE 6参照）。したがって，アトリビュートの内容自体は概ね妥当であるが，誤差の大きな項目に関する課題は残った。

樹状図によるアトリビュート習得プロセスの検討 アトリビュート習得プロセスを分かりやすく図示するためには，主要な知識ステートを同定する必要がある。本研究では，10名以上の被験者が分類される知識ステートを主要な知識ステートとする。

Pmain の設定では，それぞれの被験者の反応に対して最大4つの知識ステートを分類先の候補とすることができる。分類に成功した1,208名分のデータのうち，4つの知識ステートが候補となっていたのは911名（75.4%），3つが63名（5.2%），2つが95名（7.9%），1つが139名（11.5%），通算で1,773パターンの知識ステートが現れた。第1候補のみを考慮した場合，主要な17知識ステートに分類される被験者は304名（25.2%）に止まった。また，1人以上の被験者が分類された知識ステートの種類も668パターンと多かった。

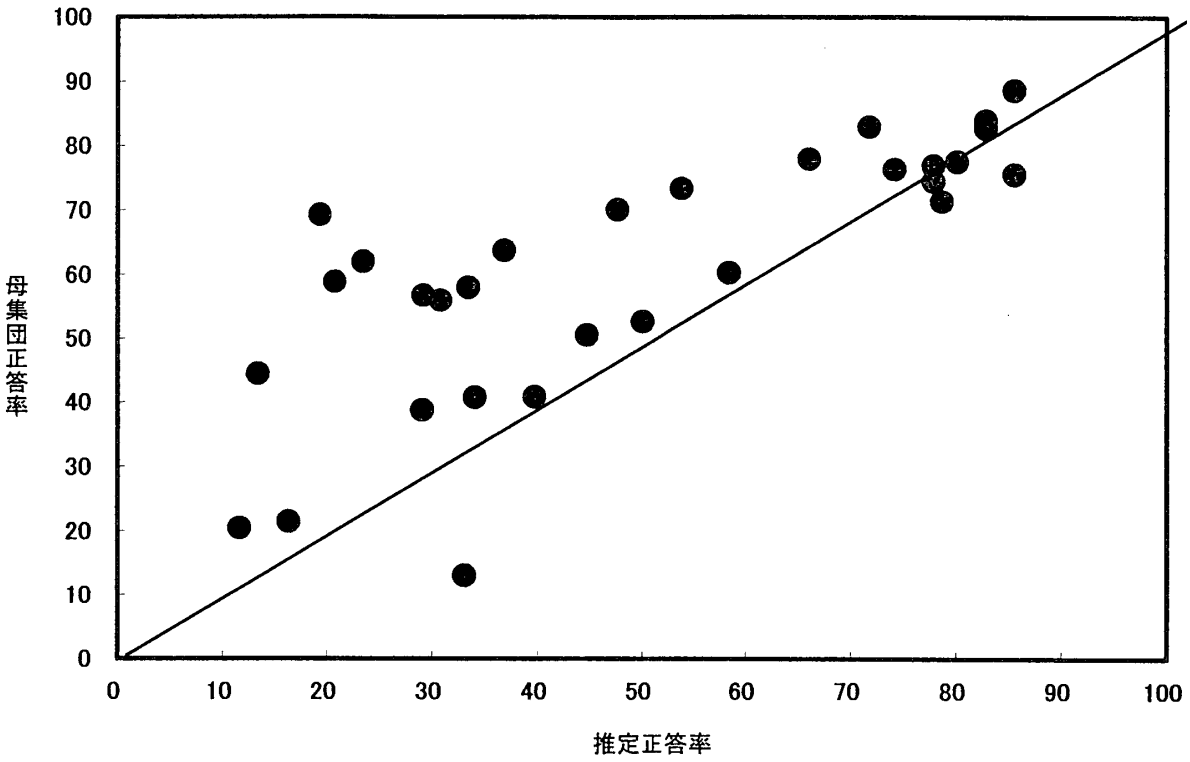

FIGURE 2 アトリビュート習得確率による推定正答率と実際の母集団正答率

そこで，分類確率を加味しながら，最大4つの候補から出現頻度の高い知識ステートに優先的に分類する基準を設けて分類先の知識ステートを決定することとした。その結果，1人以上の被験者が分類された知識ステートの数は504と当初の約3/4に減少した。さらに，410名(33.9%)が主要な知識ステートに分類され，改善が見られた。なお，主要な知識ステートの数は17と第1候補のみを考慮したケースと同じであったが，2つの知識ステートに入れ替わりが見られた。主要なアトリビュートのモデル上の項目反応パターンにおける θ, ξ はTABLE 7に示す通りである。なお，「完全習得」の被験者(11名，そのうち5名が満点)のモデル上の項目反応パターンは全問正解となるため，θ, ξ を定めることはできない。そこで，便宜上，$\theta=4.000$, $\xi=0.000$ と置いて図示することとした。

主要なアトリビュートを用いて習得プロセスを示す樹状図を作成した。縦軸に θ，横軸に ξ を取って図示したのがFIGURE 3である。知識ステートを示す記号番号は，未習得のアトリビュートを表す。また，図中の矢印は，逐次アトリビュートを習得していく場合の習得プロセスを示す。完全習得の状況に至る最後のステップをボールドで示している。

FIGURE 3から，アトリビュート「A」，「D」を習得した場合に特性値 θ が飛躍的に向上することが見て取れる。「A」も「D」も「高校生にとっての幹の難しさ」を表すアトリビュートであり，まさしくテストが測定しようとする概念そのものである。一方，「M」，「N」，「O」，「R」といったアトリビュートは「誤答への推論」である。これらも，ある種の能力と言えないことはないが，測定目的からすると本質的なものではない。「16」は正答選択肢の漢字の読みに関するアトリビュートである。語彙能力とは間接的には関係があるが，直接的なものではない。

以上の結果から，これらの30項目から成るテストは，全体として十分に妥当性のある測定が可能なものと考えられる。

考　　察

ルールスペース法(RSM)は，第一義的には受験者の属性に関する診断情報をテスト結果から抽出する目的で作られた方法論である。それを前提とした上で，本研究においては，テストが何を測っているか，すなわち，テストの妥当性に実証的に答える方法としてRSMを用いることが可能であることを示した。実際に利用されているテストにおいても，アトリビュートの発想を問題作成の過程に取り入れることによって，

TABLE 6　アトリビュート習得確率による正答率予測と実際の正答率

項目番号	推定正答率	母集団正答率	サンプル正答率	p-s差*	p-e差**
3	13.3%	44.5%	43.5%	1.0	31.2
14	77.9%	74.5%	73.0%	1.5	-3.4
15	78.6%	71.4%	67.9%	3.5	-7.2
16	85.5%	75.6%	73.9%	1.7	-9.9
18	74.2%	76.4%	77.6%	-1.2	2.2
22	82.8%	84.0%	82.3%	1.7	1.2
23	66.0%	78.0%	79.5%	-1.5	12.0
34	20.8%	58.9%	57.3%	1.6	38.1
36	23.5%	62.0%	60.8%	1.2	38.5
37	50.1%	52.7%	50.7%	2.0	2.6
38	30.8%	56.0%	53.4%	2.6	25.2
39	58.4%	60.3%	59.1%	1.2	1.9
40	53.9%	73.4%	73.1%	0.3	19.5
41	82.8%	82.8%	81.1%	1.7	0.0
42	33.0%	13.1%	12.8%	0.3	-19.9
43	29.2%	56.8%	55.3%	1.5	27.6
54	11.5%	20.5%	20.8%	-0.3	9.0
55	85.6%	88.7%	87.1%	1.6	3.1
58	33.4%	58.0%	55.1%	2.9	24.6
59	47.7%	70.1%	67.4%	2.7	22.4
60	29.0%	38.8%	38.1%	0.7	9.8
62	16.3%	21.5%	21.0%	0.5	5.2
63	77.9%	77.0%	77.7%	-0.7	-0.9
64	39.7%	40.9%	38.1%	2.8	1.2
65	36.8%	63.8%	62.4%	1.4	27.0
66	19.3%	69.3%	67.9%	1.4	50.0
69	44.8%	50.6%	50.7%	-0.1	5.8
71	80.1%	77.6%	75.9%	1.7	-2.5
72	34.0%	40.8%	39.5%	1.3	6.8
73	71.7%	83.0%	80.2%	2.8	11.3

* 母集団の正答率－サンプルの正答率

** 母集団の正答率－アトリビュート習得確率による推定正答率

TABLE 7　主要17の知識ステートの ξ, θ

コード	ξ	θ	未習得アトリビュート	分類人数
1	0.000	4.000	なし（完全習得）	11名
2	1.487	2.498	R	36名
3	-0.293	1.088	A	38名
4	-0.590	0.858	A,R	25名
6	0.856	0.832	D	10名
8	0.203	2.062	M	55名
21	-0.507	0.345	A,M,R	20名
27	1.011	2.267	16	15名
33	-0.886	0.037	16,A,O	16名
2274	-0.011	2.960	N	20名
2275	0.757	1.678	N,R	20名
2277	-0.897	0.678	A,N,R	24名
2279	-0.596	0.447	D,N	19名
2281	-0.410	1.396	M,N	47名
2283	-2.058	-0.014	D,M,N	33名
2304	-0.766	0.422	16,A,N,R	11名
2339	-0.671	0.652	16,M,N	10名

測定目的に沿ったテスト開発が可能になると思われる。

しかしながら，実用的な側面ではRSMにも不利な点がある。例えば，将来的にはともかく，現時点ではPmain等のルールスペース分析用のソフトウェアが一般のユーザーにリリースされていないという問題もある。また，ルールスペース分析の対象となるテストが項目反応理論によって尺度化されている必要がある，という点も一種の制約である。本研究で分析対象とした日本語語彙理解力テストのように1問1問が独立した形式になっているテストならば，項目反応理論に合致しやすい。しかし，我が国においては，教育場面等，様々な分野で用いられるテストの多くはより構造化された形式のものである。より広い領域において，目的志向性が高くアカウンタビリティの高いテストを作成し，テスト結果を有効に活用するためには，理論的な洗練度は低くとも単純な方法論が適しているかもしれない。例えば，オーストラリア・クィーンズランド州で実用化されているCCEsを測定するQCSテスト（山村, 1996）の考え方は，RSMのアトリビュートと項目の関係と極めて類似している。

いずれにせよ，測定の意味をより明確にするためのテスト開発，評価の方法論の研究は，今後，ますます社会的重要度を増すであろう。本研究がその一端を担うことができれば，幸いである。

引用文献

南風原朝和　1991　項目反応理論の概要　芝　祐順（編）　項目反応理論　基礎と応用　東京大学出版会，Pp.9-31.

池田　央　1977　テスト・スコアの理論　印東太郎（編）　心理測定・学習理論　森北出版，Pp.1-92.

伊藤博美・佐藤洋之・倉元直樹　2003　日本語基礎能力テストの特性(1)—国語教育から見た語彙理解項目の内容評価—，教育情報学研究，**1**，15-24.
(Ito, H., Sato, H., & Kuramoto, N. T. 2003 Properties of the Broad-Range Japanese Fundamental Language Skills Test I : Item Contents Review of Vocabulary Subscale, from the Viewpoints of First Language Education, *Educational Informatics Research*, **1**, 15-24.)

Kasai, M. 1997 *Application of the Rule Space*

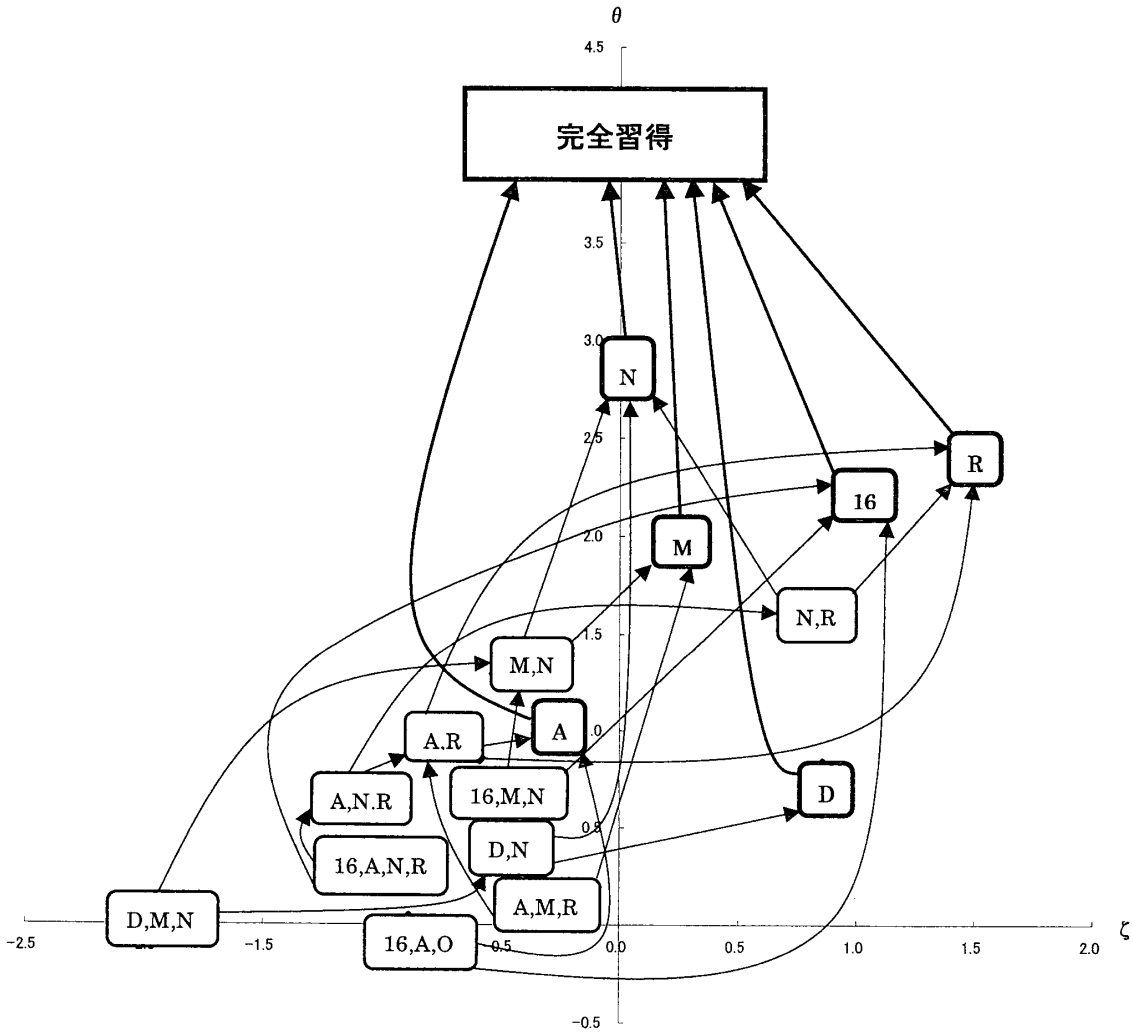

FIGURE 3 学習過程の樹状図

Model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL). Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, IL.

倉元直樹・龍岡菊美 2001 ルールスペース法—テストの診断的利用に対する測定論的アプローチ—平成11-12年度日本学術振興会科学研究費補助金（奨励研究(A)），研究課題番号 11710087，大学入試データによる学力の認知構造分析—総合試験設計のための基礎研究—，研究成果報告書
(Kuramoto, N. T., & Tatsuoka, K. K. 2001 Rule Space Methodology : A Psychometric Approach to Diagnostic Testing. *Research Report, Grant-in-Aid for Encouragement of Young Scientists (A), Japan Society for the Promotion of Science*, 11710087, 1999-2000.)

Messick, S. 池田 央（訳） 1992 妥当性 R. L. リン（編） 教育測定学原著第3版（上） C.S.L.学習評価研究所 Pp.19-145. (Messick, S. 1989 Validity. In R. L. Linn [Ed.], *Educational Measurement* [3rd *ed.*], The American Council on Education/Macmillan.)

Mislevy, R. J., & Bock, R. D. 1989 *PC-BILOG3: Item analysis and test scoring with binary logistic models*. Mooresville : Scientific Software.

Scott, H. S. 1998 *Cognitive diagnostic perspectives of a second language reading test*, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, IL.

小野 博・繁桝算男・林部英雄・岡崎 勉・市川雅教・木下ひさし・牧野泰美 1989 日本語力検査の開発，昭和61-63年度文部省科学研究費補助金試験研究(1)，課題番号6181003，研究成果報告書

平 直樹・小野 博・林部英雄 1992 高校生用日本

語語彙理解力テストの開発― (1)試作問題の精選，大学入試センター研究紀要，**21**，107-135.
(Taira, N., Ono, H., & Hayashibe, H. 1992 Development of a Japanese vocabulary comprehension test for senior high schoolers : (1) Selection of trial items. *Research Bulletin, The National Center for University Entrance Examinations*, **21**, 107-135.)

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，**28**，1-12.
(Taira, N., Mayekawa, S., Ono, H., Hayashibe, H., & Uchida, T. 1998 Development of the Item Pool for the Broad-Range Japanese Fundamental Language Skills Test. *Research Bulletin, The National Center for University Entrance Examinations*, **28**, 1-12.)

平 直樹・小野 博・前川眞一・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**，68-73. (Taira, N., Ono, H., Mayekawa, S., Hayashibe, H., & Yoneyama, C. 1995 Development of Japanese language tests for senior high schoolers' level : Calibration of vocabulary comprehension test and Chinese character reading test. *The Japanese Journal of Educational Psychology*, **43**, 68-73.)

Tatsuoka, K. K. 1985 A Probabilistic Model for Diagnosing Misconceptions in the Pattern Classification Approach, *Journal of Educational Statistics*, **12**, 55-73.

Tatsuoka, K. K. 1990 Toward an Integration of Item-Response Theory and Cognitive Error Diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Hillsdale, NJ : Erlbaum, Pp.453-488.

Tatsuoka, K. K. 1995 Architecture of Knowledge Structures and Cognitive Diagnosis : A Statistical Pattern Recognition and Classification Approach, In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum, Pp.327-359.

Tatsuoka, K. K., Tatsuoka, C. M., & Varadi, F. 1995 *Pmain (Tshell version)* [Computer software], Trenton, NJ: Tannar software.

山村 滋 1996 オーストラリア・クイーンズランド州における大学入学者選抜制度―中等学校側の評価資料の利用システムに焦点を当てて― 大学入試センター研究紀要，**25**，Pp.41-58. (Yamamura, S. 1996 University admissions system in Queensland, Australia, *Research Bulletin, The National Center for University Entrance Examinations*, **25**, 41-58.)

謝　辞

本研究の遂行において，現コロンビア大学教授の龍岡菊美先生には多大なご支援を賜りました。また，麗澤高等学校教諭の高草真知子先生，北海道立教育研究研究所研究員（当時）の多田久実子先生には国語教育のエキスパートとしてQ行列の改良にご協力いただきました。心から感謝申し上げます。

(2003.1.16 受稿, 9.29 受理)

Validity of a Japanese Vocabulary Test : Cognitive Analysis with Rule Space Methodology

Naoki T. Kuramoto (Admission Research Center of Tohoku University), Hisami S. Scott (Macromedia Inc.) and Masahiro Kasai (Assessment & Testing, Northeastern Illinois University) *Japanese Journal of Educational Psychology, 2003, 51, 413−424*

Rule Space Methodology (RSM) has been widely used as a promising method for providing diagnostic information in various testing areas. The purpose of the present study was to utilize RSM to verify the validity of the Japanese vocabulary test developed by Taira, Ono, and Hayashibe (1992). Rule Space analysis was applied to the responses of 1,500 high school student examinees ; a satisfactory classification rate of over 80% was obtained, using an incidence matrix that included 23 attributes. The 17 major knowledge states identified were used to draw a network of knowledge states in a 2-dimensional cognitive space, with axes representing overall ability level and the unusualness of the response pattern. The results indicate that students who master the important attributes among those that are being measured would receive a considerable increase in their score, whereas those who master attributes regarded as test-wiseness would get almost no increase at all. The results of this study suggest that Rule Space Methodology could be used for verifying the validity of a scale, as well as for diagnostic purposes.

Key Words : Rule Space Methodology, incidence matrix, knowledge state, Japanese vocabulary test, high school students