入学者選抜研究機構報告書７

独立行政法人大学入試センター
入学者選抜研究機構入試評価部門報告書

大学入試の標準化、多様化、および精密化

平成 24 年（2012 年）10月

独立行政法人大学入試センター入学者選抜研究機構

独立行政法人大学入試センター
入学者選抜研究機構入試評価部門報告書

大学入試の標準化、多様化、および精密化

目次

図目次

表目次

はじめに

大学入試研究機構には 3 つの部門があり，それぞれに 2 つのプロジェクトを持つ．よって，合計 6 つのプロジェクトがある．本報告書は，それらのうちの 1 つである「大学入試の標準化，多様化，および，精密化」プロジェクトの 2 つ目の活動報告である．

目次をご覧になればわかるように，本報告書は多様な論文の集成である．各論文が対象としている領域は，入試の問題項目の分析，入試の科目選択，調査書の利用，障害者支援などであり，また方法論的にも，概説的な紹介論文から，統計学的に先端的な新しい試みにまで至っている．しかし，中心的な方向性は，入試を客観的に分析しようという姿勢である．この姿勢は，入試情報の数量化，統計学的分析，意思決定論的分析などに通じている．

入試に関する議論は相変わらず数多くみられる．その中でも目立つのは，受験生の観点から入試で不合格とされ将来の可能性を摘まれることへの抗議や，入試の不公平さを批判する論議である．これらの議論は訴える力が強く，心して改善策を模索すべきである．しかし，同時に，すべての人を合格させるという方針は非現実的であり，また日本における公平性に関する議論は，すべての人が同じ条件で競争し同じ基準で評価されるという公平性に近視眼的に偏り過ぎているように思う．入試に関して，大局的に合理的なあるべき姿を論じるには，数量化や統計的な考え方が重要である．

本プロジェクトの 2 つ目の活動報告は，多様である面白みはあるものの，入試に対する具体的な提言に結びつけるには,いまだに統計学的な立論や抽象的な議論に止まっている．本プロジェクトに残された期間内に，具体的な提言や役に立つ技術の提供を完成させ，実践的な意味で役に立つプロジェクトとしたい．

大学入試センター入学者選抜研究機構
「大学入試の標準化，多様化，および精密化」プロジェクト
繁桝 算男，山形 伸二

研究組織

研究代表者

繁桝 算男　　大学入試センター入学者選抜研究機構 客員教授

研究分担者

山形 伸二　　大学入試センター入学者選抜研究機構 特任助教

研究協力者

宮埜 寿夫　　大学入試センター研究開発部 教授
大津 起夫　　大学入試センター研究開発部 教授
林 篤裕　　九州大学高等教育開発推進センター 教授
大森 拓哉　　多摩大学経営情報学部 教授
倉元 直樹　　東北大学高等教育開発推進センター高等教育開発部入試開発室 准教授
星野 崇弘　　名古屋大学大学院経済学研究科 准教授
岡田 謙介　　専修大学人間科学部 准教授
森 一将　　東京大学教養学部附属教養教育高度化機構 特任講師
橋本 貴充　　大学入試センター研究開発部 助教
立脇 洋介　　大学入試センター入学者選抜研究機構 特任助教

第 1 章　入試に役立つテスト理論

繁桝算男

本稿は，平成 24 年度大学入学者選抜研究連絡協議会における研究セミナーの内容を再構成したものである．確かに大学入試の中心となっているテスト得点は入学者選抜の役に立つ情報を含んでいるが，重要なのはその情報をいかに適切に引き出すかである．本稿はこの方法について，歴史的観点，およびテスト一般が備えるべきスタンダードの観点から考察を行う．

1.1 テストを考察する歴史的視点

テストが持つ意味は，それが用いられる歴史的文化的状況によって大きな違いがある．日本の大学入試に使われるテストもまた，現在の日本の状況を踏まえて検討すべき問題である．しかし，日本の現状を客観的に見るためには，各国のテストの歴史を理解することが有効である．

欧米で書かれたテキストでは，テストの歴史的記述はギリシア・ローマの時代から始まっている．しかし，常識的な意味でのテストに関する歴史は，中国から始めるのが正しい．国家あげての大規模なテストシステムは，中国の隋の時代，文帝によって始められた．科挙は，現代的にいえば公務員試験であり，この試験に通るのが出世のパスポートであった．テスト理論ではこのような種類のテストはハイステークスなテストと呼ばれる．

科挙は次のようにおこなわれた．清の時代の科挙にはいくつかの試験があるが，その最初の試験の貢擧は次のように行われた．箇条書きで概要を示す(宮崎，1984, 1987).

・期日：3 年毎，8 月 9 日～12 日，および，15 日と固定されていた．
・場所：各省の首府
・試験官：北京から派遣される．
・試験場：貢院 (一人ずつ個室で受験)
・日程
  ・8 月 8 日　挙士(受験生)入場 (書物，文字の書かれた紙の持ち込みは一切禁止)
  ・9 日～10 日　四書題３問，詩題１問　制限時間　翌日 10 日の夕刻まで
  ・11 日～12 日　五経５問，
  ・15 日　策題 (古今の政治を論じる)
・採点：まず筆跡から不公平が生じないように，数千人の写字生が書き写す．同考官がスクリーニングをする．スクリーニングによって薦とされたもののみ，正・副考官が採点
・合格者　各省ごとに約一万人に至る受験生の中から，40 人から 90 人

科挙の試験は，隋から清にかけて長い歴史を持つが，総じて暗記能力を問う試験(記誦の学)である．たとえば，経書の 3 文字を伏せ，当てさせる穴埋め問題が典型的である．受験者は，試験範囲の帖経をすべて覚えるとすれば，57 万字を暗記しなければならない．人生の浮沈がかかっている，ハイステークスなテストであるということと，暗記中心であることが，カンニングへの誘因となった．テスト実施者の立場からいえば，カンニング防止のために多大な努力をしなければならなかったことを意味する．科挙の長い歴史の中では，このような傾向に対する改革の試みもあった．書道で有名な王安石は，暗記よりも大義を問うべきだとし，また，優れた官吏であるためには実践的な技量が必要であると考えて，官吏任用試験として法律の知識を問う詮試を課した．さらに，例外措置ではあるが，新科明法科を試験科目とし法律の解釈や判決能力を試した．しかし，このような改革の努力は早い段階で伝統を重んじる勢力につぶされた．

このような歴史を持つ科挙をテスト理論の観点によって評価してみる．科挙の目的は，天子を助ける有意な人材を集めることと，すべての人が政治に関わる機会を持つ(と思わせる)ことであったといってよい．また，この目的はある程度達成されたといってもよいであろう．この科挙の長所と短所は次のように考える．

・長所：万人に開かれている．公明正大である(科挙に関するエピソードの数々を読んだ後では，意外な感もするが，長い歴史の大半において不正は厳しく取り締まられていたようである)．実力主義の気風を生む(科挙において非常に良い順位を得たもの，たとえば，トップスリー，状元，桿眼，探花とよんだらしいが，彼らには栄達の道が開けていた)．文官が軍を制する道を開いた．
・短所：総体的に記憶中心であり，潜在的能力，自然科学や実証の方法論，および，法律や経済などの実践的スキルへの配慮なし．学校の教育を重視しない．

西洋におけるテストの歴史についても概括してみよう．ホーガン(2005)がコンパクトにテストの歴史を 5 つの時期に分けてまとめているので，ここではそれをさらに簡潔にまとめる．

・テスト以前(1840 年まで)：人物評価をするならば，口述試験を重要視するという伝統があり，また，紙が高価なこともあって，いわゆる，試筆型のテスト(paper and pencil test)はなかなか普及しなかった．
・準備期(1840 年～1880 年)：精神障害に対する関心が高まり，障害を識別する必要が生じた．また，実験心理学が勃興し，心理事象を正確に測定するという考え方が生じた．これがテストの誕生につながる動きである．また，筆記試験は多くの例はないが，たとえば，　イエズス会の論述試験(16 世紀)やケンブリッジ大学の筆記試験(18 世紀)などにおいて，実現していたようである．

・黎明期(1880 年～1915 年) : テストらしいテストが出現する時期である．たとえば，1890 年に J.M.キャッテルの知能テスト，1905 年にビネの知能テスト出版された．また，マサチューセッツ州教育委員会が，1845 年に高校卒業試験に筆記試験を導入したが，これがアメリカにおける論述試験の始まりとされているようである．

・発展期(1915 年～1965 年) : 本格的なテストが多数出版され，厳密な標準化，信頼性・妥当性の検討が綿密になされた．
・反省と拡大(1965 年～) : 出版されるテストの数はさらに多くなり，その意味ではますます発展しているといえる．しかし，一方では，テストの使用について反省すべき点も各方面から指摘されている．たとえば，人種や民族の差別とかかわる問題や，司法や行政においてテストが果たす役割について，多くの論議があった．また，特筆すべき現代的な傾向は，何と言ってもコンピュータの普及との関連である．コンピュータの利用は，テスト項目の作成や実施，テスト得点の信頼性や妥当性の検討などすべての側面に大きく影響している．

先に挙げたマサチューセッツ州教育委員会が，なぜ口述試験に加えて論述試験を導入したかという理由を挙げている．それは，口述試験が，公平ではない，試験官の恣意的判断が入る，時間がかかりすぎる，規準がはっきりしないなどという理由である．これらの口述試験の欠点を是正し，客観性を求めた結果は，多肢選択式を典型とする客観式の試験の普及であった．客観テストが導入されたこの時期には，客観性が高い，幅広く知識を問うことができる，多様な統計的分析を行うことができる等の利点が強調された．

現在，大学入試センター試験に対する批判として，このような客観テストが「問と答えが近すぎる」，「論理的思考力を測定していない」，「長時間かけて問題を熟考し解決しようとする態度を抑制する」などの理由を挙げられていることは興味深い．繰り返しになるが，試験に完璧を期待することはできない．せいぜいのところ，目的に適うように最善を尽くすのみである．なお，客観的なテストの設計や実施について，池田(1992)が参考になる．

1.2 テストスタンダードからみた大学入試

本節では，テスト理論の枠組みによって，入試を評価することを考える．より具体的には，テストの規範を示すテストスタンダードの観点から，入試を考察する．テストスタンダードはテストの先進国では当たり前に存在するものであったが，日本では，2007 年にやっと日本テスト学会の編集によって誕生した(日本テスト学会編，金子書房，2007).

まず，テストとは何かを明らかにしなければならない.

(テストスタンダード)テストとは，「能力，学力，性格，行動などの個人や集団の特性を測定するための用具であり，実施方法，採点手続き，結果の利用法などが明確に定められて

いるものである」

入試が客観性をもとめ，説明責任(アカウンタビリティ)があることを考えれば，入試選抜資料の主要部分は，上記のような一般的な意味においてではるが，テストであるべきだと考える．

テストスタンダードに従い，以下，1．信頼性，2，妥当性，3．標準化，4．等化，5．公平性の 5 つの観点から，入試システムを評価する．

1.2.1 信頼性

(テストスタンダード)テスト開発者は，構成された尺度得点がどの程度安定しているかについて，統計指標を算出して検討し(この過程を「信頼性の確認」という)，その結果を開示すべきである．なお，テストが複数の下位テストから構成される場合は，それぞれの下位テストごとに検討し，その結果を開示すべきである．

信頼性の統計的定義は，テスト得点の分散のうち，真の得点の分散の占める割合である．ただし，この定義に含まれる「真の得点」はかなり問題の多い用語であり，測定対象の真の値を意味せず，統計的には，単なる平均(理論上の分布による平均なので，期待値というのが正確であるが)である．そもそも，信頼性という用語は誤解を生じやすい．信頼性が高いことが，このテストを信頼し，使用してよいということであることには直結しない．
信頼性の値を推定する一般的な方法としては，(１)同じ真の得点を持つテストを二つ作って，その間の相関係数を計算する方法，(2)下位にある尺度(あるいは項目そのもの)の内的整合性から評価する方法の二つがある．これらの方法は，成書(たとえば，リン，1992)に詳しいし，統計学的には分散成分の分割に当たるので補足で説明する．

1.2.2 妥当性

(テストスタンダード)テスト開発者は，構成された尺度が測定内容として定義された特性をどの程度適切に測定しているかどうかを多面的に検討し(この過程を「妥当性の確認」という)，その結果を開示すべきである．

信頼性と比べ，妥当性は文字通りの意味を示す．すなわち，妥当性は，ある特定の理論に従うときテスト得点の解釈が妥当であるとか，特定の目的のためにテスト得点を使用することが妥当であるということを意味する．解釈が妥当であるとは，ある理論を正しいとした場合に，その理論上のある特定の概念(構成概念と呼ばれる)を測定しているとみなしてよいということである．(測定には，その測定を成り立たせる前提条件を公理とし，その公

理から導かれる必然として定義される場合(たとえば，長さや質量)や，そのような公理系を背景としなくても，ただ定義することによって測定とする場合(たとえば温度)などもある．このような場合には妥当性の問題は生じない．物理学では測定とは何かという議論が生じないのに対し，構成概念を測定しようとする心理学で測定とは何かについて議論が絶えないのはそれが理由である．) テストの使用が妥当であるということは，その使用が何らかの便益をもたらすことであり，その便益はテストを使用するために要するコストとくらべて意味があるほどに大きくなければならない．

1.2.2.1 妥当性検証の方法

妥当性を評価する方法は一筋縄ではいかない．いろいろな方法を用いて妥当性を検証する努力を重ねるしかないと考える．まず，テストの項目や尺度の内容を考察してテストが妥当であることを評価する場合と，実証的データを用いて妥当性を評価する場合に分けられる．前者を表面的妥当性，後者を経験的妥当性という．表面的妥当性は，言葉の語感はよくないが，実際には必要とされることも多い．社会に受け入れられるには，そのテストが妥当であるということを，テストの使用者や受検者に納得してもらうことが望ましいからである．これに対して，経験的妥当性とは，empirical validity の訳語である．empirical，すなわち，理論ではなく，実験や調査に基づくという意味である．実験や調査などのシステマティックなデータを妥当性検証のために使う方法はいくつかある．妥当性とは何かを考えるための知識として，それぞれの方法を簡単に説明する．

(1)内容的妥当性：測定対象が明確に定義され，かつ，カテゴリー化されており，カテゴリー化された各分野とテスト項目との間の関連もはっきりしている場合に有効である．この場合に問うべきは，収集されたテスト項目群が，測定目標の諸分野をよく代表しているかどうかであり，この代表性は，測定目標の各分野における出題数の頻度表を検討すればよい．

(2)基準連関妥当性(併存的妥当性と予測的妥当性)

測定したい属性が実際に測定されるならばそれを基準とし，その基準とテスト得点との相関の強さが妥当性の証拠となる．基準がすでに測定されているならば併存的妥当性，将来において測定されるならば予測的妥当性というが，この区別は本質的ではない．問題は，測定したい属性の測度が簡単には手に入らないことである．簡単に測定できるならば，そもそもテストなど必要ないともいえる．実際には，「テストよりも費用が掛かる」，「測定対象の一部からしか得られない」，「完全には測定目標の属性を代表していない」ような基準の測定値しか得られない．したがって，この不完全な基準との相関係数をテストの妥当性とすることには危険を伴う．この危険性を認識し，基準とテストの関連について，的確に統計モデルとして表現し，真の関連にもっとも近い相関関係を推定する必要がある．

(3)構成概念的妥当性

テストの測定目標は構成概念であることが多い．根拠とする理論は観察可能な予測をす

ることができる．その予測が実証的にテストによって確かめられているならば，それは，テストが妥当性を持つことを意味する．

1.2.2.2 妥当性検証の現代的傾向

以上のように妥当性を伝統的な分類に従って解説した．概括するならば，理論的にそうなるはずのことをテスト得点が実現するならば，そのテストが妥当性を持つということである．その意味で，妥当性検証は構成概念妥当性の検証であると書かれることもある．しかし，この場合に注意しなければならないのは，理論から導かれる予測が正しいとは限らないということである．テスト得点を用いて，その予測が実現できないことを確かめたとした場合，それは理論自体に問題がある可能性もある．方法論的に言えば，理論の検証とテスト得点の妥当性検証は，同じことの表と裏の関係にあるといえよう．

理論的予測が実現しているかどうかを確認する方法も，統計学や認知科学の手法の発展に伴い，現代的な変化がみられる．テスト得点と基準との関連が理論に適っているかどうかを検証するためには，因子分析や構造方程式モデルが適している．多属性多方法の関連構造は，同じものを測定しているならば相関が高く，違うものを測定しているならば相関は比較的に低いという予測に基づく．一般的にもっと複雑な形で，テスト得点と基準との関連を予測するならば，その関連の構造をパス図であらわし，データとの関連を問うことができる．これが多変量解析的に妥当性を問う一般的な形である．

現在，脳の働きの可視化をはじめとして，心の内的プロセスを反映する生理的測度が発展しており，テスト得点と内的プロセスとの関連付けがより容易になっている．ある種のテストの妥当性がこのような測度との関連で論じることができる．

ここまで議論してきた妥当性は，テスト得点の解釈にかかわる妥当性である．すなわち，テスト得点を得て，その意味を解釈し，心理学の理論の発展に寄与しようとする目的や，受験者をテストし，そのテスト得点を解釈することによって，その受験者について，より良い知識を得ようとする場合を想定している．

一方，より直接的に，テストを使用することがどの程度妥当であったかを問うこともできる．テストの使用の妥当性を結果的妥当性(outcome validity)という．テストの使用が妥当であったかどうかは，当面する意思決定に役に立ったかどうかを問うことと同義である．この結果的妥当性については，1.2.4 項で論じる．

1.2.2.3 信頼性と妥当性の関係

信頼性と妥当性の関係を説明するために次のベン図(図 1.1)が役に立つ．

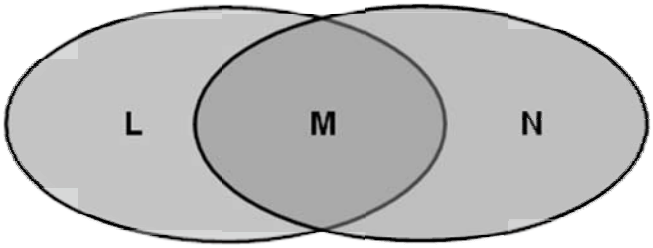

図 1.1 信頼性と妥当性の関係

ただしこの図において，ある構成概念(e.g. 学力，知能)を測ろうとするとき，三つの領域 L,M,N は，それぞれ，テストによって測定できない構成概念の部分(L)，テストによって測定される構成概念の部分(M)，テスト得点に含まれる構成概念と関係ない部分(N)である．

信頼性とは，テスト得点の値があまり変動しないという意味で信頼できることをいうのみであり，それは必要なことではあるが，そのテストを使ってよいということには直接つながらない．一方，そのテスト得点の解釈が妥当であり，かつ，そのテストの使用が目的に適っているならば，そのテストを使う理由として十分である．

たとえば，知能という構成概念を測るために，頭の周り(頭囲)を測ったとする．この頭囲は，だれが測っても，あるいは，今日測っても明日測ってもあまり値は変わらないし，測定者の違いにも時間的な違いにもかかわらず安定している．しかし，知能という概念を用いるどのような理論に従った解釈も頭囲に対しては妥当ではない．頭囲は，知能という構成概念の測定において，信頼性はあるが妥当性はないと言える．知能テスト得点は，すくなくとも頭囲よりは妥当性が高い．実際，知能が高ければ，学業成績が優秀になるという理論的予測はデータによって検証されている．ただし，個々人のテスト得点を，個々人の知能ととらえることができるほど信頼性は高いか，妥当性は高いかと問うならば，それは議論の余地がありそうである．

1.2.3 標準化・等化・公平性

標準化については，テストスタンダードでは次のように言及されている．

(テストスダンダード)汎用されるテストは，規準とする集団を明確に定め，その集団における相対的位置づけによって尺度化することが望ましい(この手続を標準化という)．標準化においては，用いた標本と標準化手続について記録し，開示する．また，標準化の結果は定期的にその有効性を確認し，改訂の必要性の有無を検討する．

テスト得点に意味を与える有力な方法が，基準とする集団における相対的な位置づけである．一般的な世論ではこのような相対的な位置づけは評判が悪い．偏差値という便利な統計指標の評判が悪いのは，それが競争を意味するからである．しかし，測定する範囲が

広く，かつ，複数の作題者が作るスト得点に意味を与える方法としてはこの相対的位置づけが適していることが多い．標準化による得点は基準とする母集団と標本(データ)の取り方が一定していれば，テスト得点のそれぞれの意味も安定している．

実際の入試で用いられるテストは，受験生の母集団が毎年変遷する可能性があり，また，教科科目の選択の仕方も一定とは言い難い．このような場合に，各年度のテスト得点や異なる教科科目の得点を比較可能にする努力は等化と呼ばれる．

(テストスタンダード)同一の特性を測定する2種以上のテストの採点結果を比較する場合や，実施時期が異なるテストの採点結果を比較する場合には，それらのテストが相互に比較できる尺度得点に変換されている必要がある(この手続きを等化という)．

等化は同じ学力を持つものが同じ得点になるように(あるいは換算表が用意できるように)することである．この問題を統計学的に解決しようとするならば，学力は潜在変数であり，潜在変数を含む階層型の統計複雑モデルを利用するのが自然である．このことに関してはまた別の機会に論ずることにする

最後に，入試において特に重要である公平性については，テストスタンダードは次のように述べている．

(テストスタンダード)受検者は，テストのすべての過程において年齢，性，国籍，障害の有無などによって差別されてはならない．また，質問項目の表現は，特定集団の成員に不快感を生じさせないように配慮されなければならない．

受検者が所属する，いくつかの下位集団に対する差別を念頭に置いた表現である．この点では，日本よりも深刻な多くの問題を抱えてきたアメリカの『教育・心理テストのスタンダード』(Standards for educational and psychological testing, AERA,APA,NCME, 1999)では，公平性(Fairness) について具体的に次のように述べている．

- 内容の適切さや，反応の内部構造等に関して，下位集団間の違いを示す確かな研究が報告された場合は，テスト得点の解釈と利用に際して，十分考慮されるべきである．
- 開発者は，人種，民族，性などの特定の集団に対して攻撃的とみなされるような言語，記号，語句，内容を識別して除去する努力をするべきである．
- 性，民族，年齢，言語，障害などの集団ごとの得点比較が公共に報告される際に，比較に意味がないという確かな研究結果がある場合は，そのことが警告されるべきである．

このような意味での公平性は必須の要件である．また，障害の有無による差別をなくすということは，テストスタンダードのように目標として掲げることはやさしいが，実際に

どのような処置をすれば公平といえるかという問題はチャレンジングな課題である．

1.2.4 意思決定理論と妥当性

意思決定理論はテストの結果的妥当性を明確に定義するために必要な枠組みである．いうまでもなく，入試は合格か不合格かを決める意思決定である．単純化して言えば，この意思決定の目的は，大学側として合格に値する学生を合格させ，大学が提供する教育リソースを活用できないであろう学生を不合格とすることである．この事態を図式化すると表1.1 のようになる．

表 1.1 意思決定としての大学入試

代替案	望ましい学生	望ましくない学生
合格(a1)	結果 1	結果 2
不合格 (a2)	結果 3	結果 4

この図式の状況において，大学はどのような決定をすべきか．意思決定論の答えは，大学側から評価して，予想される便益を最大化するということである．そのためには，次の二つが必要である．(1) 期待する受験生とはどのような受験生かを明らかにし，それぞれの受験生の望ましさの程度を数値化できるような入試資料を整えること，および(2) 入試資料によって，意思決定の結果のうち二つの許容できる結果(結果 1 と結果 4)の確率を最大化すること(あるいは，二つの“不都合な”結果 2 と結果 3 の確率を最小化すること) である．

当然のことと思われるかもしれないが，この単純な答えの中には解決しなければならない重要な課題が含まれている．たとえば，大学がそれぞれの受検者について，大学で教育を受け，成功するかどうかを大学入試という不完全な情報から評価しなければならない
このことを意思決定理論の用語に置き換える．便益は，その望ましさの観点から評価するとき効用(utility)とよばれ，また，その受検者が合格するかわからない不確定性の評価は主観確率によってなされる．意思決定論では，伝統的に，効用ではなく，その裏返しの損失で評価することも多かった．たとえば，図 1.1 の場合においては，結果 1 と結果 4 を損失 0 とし，それとの比較において，結果 2 と結果 3 の不都合な程度を損失とし，予想される損失，すなわち，期待損失を最小化する．損失は効用を適切に変換して得られるので，本質的には，期待効用の最大化は，期待損失最小化と同じ結果をもたらす．

ある特定の大学の入試選抜システムの望ましさを評価するためには，受験生の母集団を特定しなければならない．最適化される選抜システムは，この母集団全体に対する期待効用となる．この議論は数式を用いないとかなり煩雑な説明になるので，一般的な枠組みによる説明は補足に譲る．

1.3 テスト理論の専門家の仕事

以上の理論的枠組みを踏まえ，テスト理論の専門家ができることをいくつか例示する．

1.3.1 入試の試験項目の選択と選抜資料の組み合わせの最適化

各受験生の望ましさ(効用)を評価するためには，どのような選抜資料を採択するべきかという問題は，テストを実施することによって予想される便益(EVSI という指標を用いるが，EVSI については補足や繁桝(1985)を参照してほしい)を最大化する問題である．

テストを構成する項目のいずれを選択するべきかという問題も同じである．可能なテスト項目の組み合わせにおいて，EVSI を最大化する項目のセットが最適である．テスト実施のコストがゼロならば，可能なテスト項目はすべて採択するのがベストであるが，テスト実施にはコストがかかる．このコスト自体も効用評価に含めることも理論上は可能であるが, コストを考慮して, 項目の数を設定したうえでEVSIを最適化するほうが現実的である．項目の数はある程度大きい数なので，この場合，最適化する関数が明確に定義されていても，いわゆる「組合せ爆発」が起こり，計算の負担は大きすぎる．実際の項目選択のアルゴリズムは，回帰分析における変数選択と同様の工夫が必要である．

項目の特性をこの EVSI の最適化のアルゴリズムを適用可能にするには，数理モデルが必要である．各項目への反応をモデル化したものが項目反応モデルである．反応を正答か誤答かに分けた場合のモデルには，ラッシュモデル，2 母数や 3 母数をもつロジスティックモデルがある．反応が 3 つ以上の選択肢への応答であるとき，そのモデルは多重ロジスティックモデルである．

1.3.2 選抜方式の最適化

選抜資料に基づき，合格者を決める選抜方式にはいくつか異なる可能性がある．もっとも一般的なのは総合点方式である．個々の入試選抜資料を得点化し，それぞれにウェイトをかけて合計する．その合計点が高いものから合格とさせる方式である．また，個性が強くても得意な分野を生かしたいとか，潜在的な適性を重視したいという理由で，選抜資料のうち最も良い成績を重視したい場合には，最高の得点を選抜の資料とすることもある．これは最大点方式である．たとえば，国語，数学，英語の 3 教科得点のうち，最も高い得点を用いて選抜の資料とする場合である．この発想とは別に，いくつかの選抜資料すべてが必須である場合には，その最低の得点が問題となる．この場合は，たとえば最低点が最も大きい受験生から合格とすればよい．また，段階的に選抜することもできる．いわゆる足切りでまず予備的にスクリーニングをし，そのあとより精密に時間をかけて受験者を精査するという方式である． いずれにしろ，いろいろな方式が考えられるが，本質的には，それぞれの方式のもたらす望ましさ(損失)とコストを考慮して最適化することで選択することになる．

1.3.3 選抜システムの評価

合格者は正しく選抜されたかという事後的な評価は，合格者が入学後にどのような成果を上げたかを追跡することで可能になる．どのようなデータをとるかが重要であり，可能性のある指標は，大学の成績(GPA)，満足度，動機づけの強さ，などであろう．選抜資料はこれらの指標をよく予測しているであろうか．この問題は，予測的妥当性の問題である．先に述べたように，妥当性の検証の方式としては，比較的単純であるが，気を付けなければならないのは，検証するためのデータが本来の母集団である受検者全体から得られず，選抜された合格者のみから得られることである．この事実を適切に調整しないと，間違った結論に達することになる．計量心理学は，この調整の方法を提供することができる．

補論

A1.1 テストによる測定の困難さ

テストの歴史を紐解いてまでここで言いたかったことは，テスト得点とテストがそもそも測定しようとした性質・特性・属性との関係をモデル化することの難しさである．たとえば，科挙の試験は，天下のために役に立つ能力を測ろうとしている．科挙の試験の得点をx，その能力をθとする (観測される変数をローマ字で，観測されず推測するしかない変数をギリシア文字で示す伝統がある)．ここで，これを統計的分析に使えるモデルにするためには，xとθの関係を明示し，かつ，未知の部分にはある分布を特定しなければならない．たとえば，

$$x = \theta + \gamma \tag{1.1}$$

という加法的なモデルを考える．この単純なモデルですら，多くの問題をはらむ．このモデルは，(1) が本来の測定目標を含んでいるということ，(2) 測定対象のθと残差γが加法的に観測得点と結びついていることという二つの仮定を含んでおり，通常はこの仮定は満たされない．現実的なモデルとしては，このことを，同じ条件における測定の繰り返しの期待値とそれ以外に分けるモデルと比較してみる．そのモデル は，

$$x = \tau + \varepsilon \tag{1.2}$$

となるが，この場合は，残差の部分が正規分布であり，一つ一つの測定値が独立に分布するという仮定を無理なく設定することができる．xの分散のうち，τの分散が占める割合が信頼性といい，xと望む値θとの相関を妥当性というが，信頼性の分析のほうが統計的モデルに則って推論することが容易である．

以上のことを考慮すると，考察の対象となる統計モデルは次のようになる．

$$x = \tau + \varepsilon = \theta' + \eta + \varepsilon \tag{1.3}$$

すなわち，このモデルは，テスト得点xは，安定し構造化されている部分(系統的成分)と，無作為に変動する部分(誤差成分，残差成分)に分けられること，またさらに，安定した部分

は，望んでいる情報θと関連するθ'とそれ以外のηに分けられる．

科挙の例でいえば，科挙で選抜したいのは，国家有為の人材であり，測定したいのはそのための能力θである．このθが，テスト得点xの中にどの程度含まれているかが問題なのである．

A1.2 一般化可能性理論

信頼性は，本文にあるように，テスト得点の分散を意味のある成分に分解することによって，推定できる．たとえば，被験者iが，あるテスト実施条件jにおいて得られたテスト得点について，同じ状況が繰り返されることを想定すると，k回目の繰り返しにおいて，

$$x_{ijk} = \tau_i + \beta_j + \gamma_k + \varepsilon_{ijk} \tag{1.4}$$

というようにモデル化できる(ちなみに, このモデルは, 分散分析の変動モデルと呼ばれる．被検者や採点者が固定されておらず，ある母集団からのサンプルとみなされるからである．なお，この場合簡単のために交互作用は無視できるものとしている)．通常の場合，テスト得点の分散xのうち，τの分散のみが意味のあるものであり，その分散が占める割合が信頼性の指標となる．

また，採点者が少数者に特定されている場合には，母数βを推定し，各採点者の特徴を知ることもできる．このように分散分析として，テスト得点の分散を分割し，それぞれの成分の分散を推定し, 一般化の可能性のヒントとする方法を一般化可能性(generalizability)の理論という．

A1.3 妥当性検証の手段

妥当性とは，テスト得点xがどの程度θを反映しているかを推定する問題である．この問題に直接に統計的問題として答えることはできない．測定しようとしているθを含む統計モデルが構成できないからである．信頼性の推定と比較して，妥当性の評価が難しいのはこれが理由である．本文で説明したのは，入試の文脈で重要な妥当性は，テスト得点を構成概念の指標として解釈できるかどうかではなく，入試によっていかにアドミッションポリシーで示す目的を実現できるかどうかである．そして，この場合の妥当性の指標は，テスト得点を用いることによって得られる便益の予想値(すなわち，期待効用)とテストなしで毛手知する場合の便益の予想値(すなわち期待効用)の差であると考える．テストによって，どれだけ便益が大きくなるかの測度である. これは, サンプル情報の期待価値, (Expected Value for Sample Information, EVSI)であるという主張であった. このEVSIを明示的に示すために，想定される変数やパラメータの数をまず整理する．

妥当性の基準となる変数をyとし，一般的に扱うためにその数をqとする．大学入試において規準となるのは，たとえば，GPA や出席率である．また，妥当性の基準になる変数が観測されず，潜在している場合がある．たとえば，学力や動機づけの強さなどである．これをξで表し，問題とする構成概念の数をrとする(科挙の例では求める潜在変数をθとした．

基準を予測するための選抜資料をxで示し，その数をpとする．すなわち，

$y = (y_1, y_2, \ldots, y_q)^t$，基準　　例: GPA，出席率
$\xi = (\xi_1, \xi_2, \ldots, \xi_r)^t$，潜在変数　　例: 学力，適性，動機づけ
$x = (x_1, x_2, \ldots, x_p)^t$，選抜資料　　例: センター試験，個別試験，調査書

また，入試システムにおける代替案は，次の二つである．

a_1: 代替案 1 (合格という決定)
a_2: 代替案 2 (不合格という決定)

これらの決定をそれぞれの受験生iに適用した場合の効用を次に示す．

$u(a_1, y) = u_1(y)$: 基準yを持つ受験者を合格とした場合の効用
$u(a_2, y) = u_2(y)$: 基準yを持つ受験者を不合格とした場合の効用

この効用を潜在変数ξによって表現する場合，あるいは，直接に選抜資料xによって示す場合には，yのかわりに，ξやxを代置すればよい．

受験者iの効用を，規準yや潜在変数ξから評価しようとする場合，受験の時には未知である．未知ではあるが，推測することはできる．それが予測分布である．受験者iの効用は予測分布による期待効用である．すなわち，

$$\bar{u}_1(i|x_i) = \int u_1(y_i) p(y_i|x_i) dy_i \tag{1.5}$$

$$\bar{u}_1(i|x_i) = \int u_1(y_i) p(y_i|x_i) dy_i \tag{1.6}$$

となる．

n人の受験者に対して，受験者iが合格としたとき，$k(i) = 1$，不合格としたとき，$k(i) = 2$とする．同様に，選抜資料xを用いて，受験者iが合格としたとき，$k^*(i) = 1$，不合格としたとき，$k^*(i) = 2$とする．このとき，受験者集団全体に対する選抜資料の価値は，

$$\begin{aligned} I &= \sum_{i=1}^{n} \int u_{k^*(i)}(y_i) p(y_i|x_i) dy_i - \sum_{i=1}^{n} \int u_{k(i)}(y_i) p(y_i) dy_i \\ &= \sum_{i=1}^{n} \bar{u}_{k^*(i)}(y_i|x_i) - \sum_{i=1}^{n} \bar{u}_{k(i)}(y_i) \end{aligned} \tag{1.7}$$

によって表される．選抜資料xが利用できない場合のyの分布は知識がない場合の事前分布

と同じような問題をはらむ．この応用場面では，資料がない状態でどのような決定をしても効用は同じであることを仮定し，無知の状態を仮定する問題を回避する．すなわち，選抜資料の価値は，式(1.7)の第 1 項によって比較される．

ここで，予測分布の導出についてコメントしておく．大学入試において得られるデータには，際立った特徴がある．すなわち，入試選抜資料と基準の両者のデータのサンプリングが変則的であり，基準のデータは合格者のみに対して得られる．このデータに対して，この変則的サンプリングに影響されにくい回帰係数パラメータと条件付尤度を用い，ベイズ的な推論方法によって元の母集団における予測分布を導出することができる(岡田・繁桝, 2010).

ここまで，潜在変数ξによる効用評価について説明していない．原理的には，潜在変数(未知パラメータといってもよいが，被検者の数とともに増える特異なパラメータである)の事後分布を求めることによって, EVSI を評価することができる．しかし, 繰り返しになるが, ξをカギとする統計モデルは容易には組み立てられない．例えば，やる気や応用力というような，抽象的な言葉で受験生の望ましさを測り，また，選抜資料のそれぞれの価値を評価しようとする場合の難しさがここにある．すなわち，各受験生の望ましさを学力や動機づけの強さなどの構成概念ξで合格とした場合，および，不合格とした場合の評価の結果をそれぞれ$u_1(\xi_i)$, $u_2(\xi_i)$とすると，選抜資料の情報価値は,

$$I = \sum_{i=1}^{n} \int u_{k^*(i)}(\xi_i) p(\xi_i | x_i) d\xi_i - \sum_{i=1}^{n} \int u_{k(i)}(\xi_i) p(\xi_i) d\xi_i$$

$$= \sum_{i=1}^{n} \bar{u}_{k^*(i)}(\xi_i | x_i) - \sum_{i=1}^{n} \bar{u}_{k(i)}(\xi_i) \tag{1.8}$$

となる．上式において,

$$p(\xi_i | x_i) = \frac{p(x_i | \xi_i) p(\xi_i)}{p(x_i)} \tag{1.9}$$

であるが，構成概念を所与とした場合のデータ発生モデル$p(\xi_i | x_i)$の設定が難しいことが問題である．たとえば，ここに因子分析モデルや構造方程式モデルを仮定することになるが，それが正しいかどうかは慎重に見極める必要がある．

引用文献

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

ホーガン, T.P. (2005). 繁桝算男・椎名久美子・石垣琢磨(訳). 心理テスト 培風館.

池田央 (1992). テストの科学－試験に関わる全ての人に　日本文化科学社.
リン, R.L. (編), 池田央・藤田恵爾・柳井晴夫・繁桝算男(監訳).　教育測定学　みくに出版.
宮崎市定 (1984).　科挙　中央公論新社.
宮崎市定 (1987).　科挙史　平凡社.
日本テスト学会 (編) (2007). テスト・スタンダード 日本のテストの将来に向けて
　金子書房.
岡田謙介・繁桝算男 (2010).　小標本における選抜効果を補正する相関係数の推定について
　日本テスト学会誌, **6**, 63-74.
繁桝算男 (1985).　ベイズ統計入門　東京大学出版会.

第 2 章　センター試験における科目選択と受験者属性

宮埜寿夫

2.1 はじめに

大学入試センター試験は，高校段階における基礎的学習の達成度を測定を目的とする統一試験であり，多くの大学において入学者選抜の資料として利用されている．ここでは，入試のスタンダードを検討する際に必要とされる情報のひとつとして，センター試験において測定することが求められている学力の範囲について，高校ごとにまとめられたセンター試験の成績データに基づいて議論する．また，そのようなまとめられたデータの分析法についても議論する．

2.2 センター試験による学力測定

高校での履修の多様化にともない高校生の学習量には大きな違いが生じ，そのために学力差は拡大したと言われる．一方，高校進学率はセンター試験開始当初の約 94%から現在約 98%に達しており，大学進学率は約 36%から約 55%になっている．したがって，センター試験の測定すべき学力の範囲は，高校での履修の多様化による学力差の拡大もあるが，基本的には大学進学率の増大により，センター試験の開始された 1990 年頃よりも拡大したと言えるであろう．

センター試験から見た学力の範囲を調べるために，英語(筆記)の成績をある県について分析した．英語(筆記)を分析対象とした理由は，センター試験ではこの科目がほとんど全ての受験生により選択される科目であること，他の科目の学力との関係が深いことによる．この県には約 50 校の高校があり，英語の受験者総数は約 5,000 名であるが，高校によって受験者数は大きく異なっている．図 2.1 は，受験者数の比較的多い高校について現役受験者数と受験率の関係をプロットしたものである．図にプロットされている高校は 36 校であるが，36 校の受験者数は全体の受験者数の約 98%を占めている．残りの 15 校程度からの受験者数は，各校数名である．受験率 80%以上の学校数は 20 校程度あり，この県の現役受験率は全国平均の約 40%より高い．

センター試験が測定すべき学力の範囲は，学力差の拡大の真偽はともかく，大学進学率の増大により拡大したと言えるであろう．とくに，センター試験開始当初よりも低い学力層の方向に測定すべき範囲が拡大している可能性がある．図 2.2 は，選択された 36 校について現役受験者の英語の平均得点と標準偏差をプロットしたものである．平均得点が大きくなるにしたがって，標準偏差は小さくなる弱い傾向が見られる(相関係数-0.15)．この弱い相関は，学力が平均以下の高校間で標準偏差のちらばりが大きいことによる．また，平均得点が 60 点程度の高校があり，英語の試験はこれらの高校の生徒に対する学力測定が困難

であることを示している．なお，受験者数と平均得点には強い相関(0.87)がある(図 2.3).

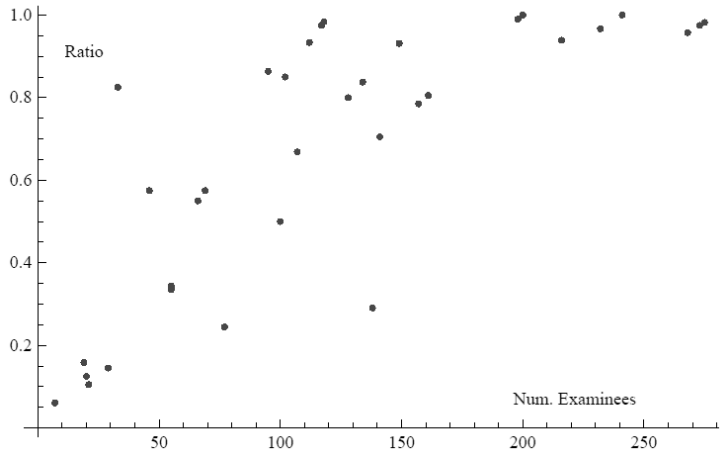

図 2.1 高校ごとの受験者数と受験率

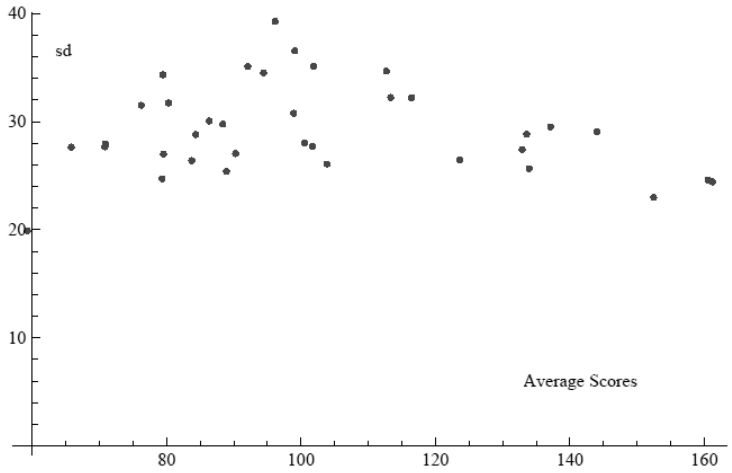

図 2.2 高校ごとの平均得点と標準偏差

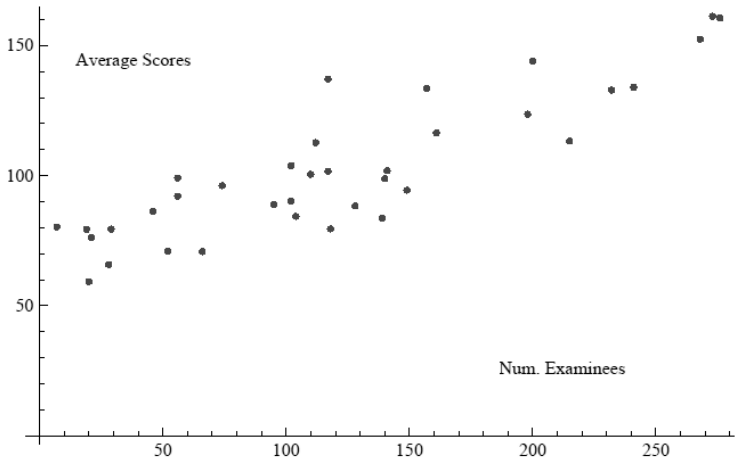

図 2.3 高校ごとの受験者数と平均得点

一般に，センター試験は，学力の低い層に対する学力測定が十分にできていない可能性は否定できないであろう．一方，学力の高い層に対する測定は，数学 IA を除いてうまく行われているように思われる．

大学への入学志願者数は，少子化を反映して毎年減少している．一方，入学者数も減少傾向にあるものの，志願者数と入学者数との差は毎年小さくなっており，大学を選びさえしなければ志願者の誰もが入学できる状態，大学全入時代に入りつつある．しかし，大学全入時代という言葉は，大学を選ばなければという前提の下で意味があり，センター試験の受験者に 15 万人程度の浪人生が含まれていることを考えると，この前提は成り立っているようには思われない．図 2.4 に，平均得点とセンター試験受験生に占める浪人生との関係を示す．この図より明らかなように，平均得点の高い高校ほど浪人生の割合の多くなっている(相関係 0.74)．このことは，当然のことではあるが，現実には大学は選ばれており，大学全入という言葉は学力の低い層にのみ意味があることを示している．

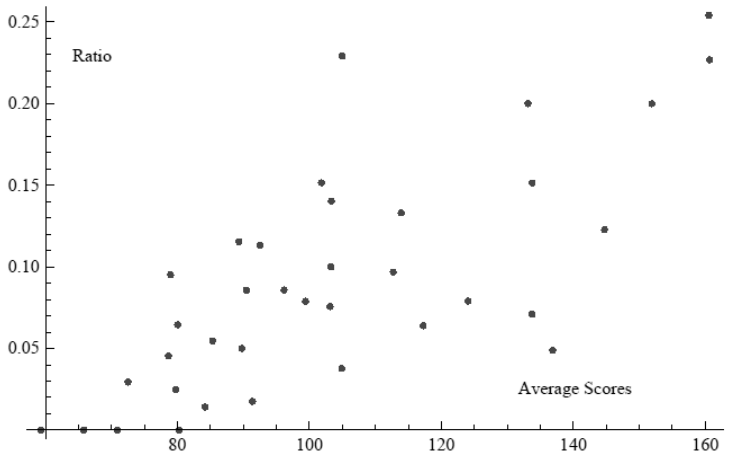

図 2.4 高校ごとの平均得点と浪人生の割合

2.3 センター試験における科目選択

センター試験は，入学者選抜の重要な資料として利用されているため，高校段階での教育に大きな影響を与えている．とくに，地理歴史，公民および理科は，高校間で平均的な学力差を反映した科目選択がなされていると思われる．図 2.5，図 2.6 は，高校ごとに地理歴史・公民および理科科目の受験者数を高校×科目の分割表にまとめたデータを対応分析により解析した結果であり，図 2.5 は科目の尺度値，図 2.6 は高校の尺度値を表している．理科科目では地学，公民科目では倫理の選択に高校による違いのあることが分かる．すなわち，この県の主要な進学校である 3 高校では，これらの科目が相対的に多く選択されている．

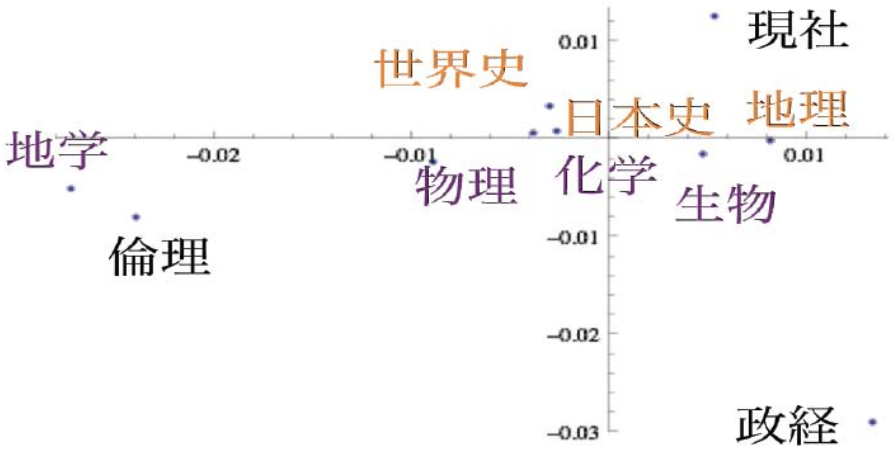

図 2.5 科目の布置 (対応分析)

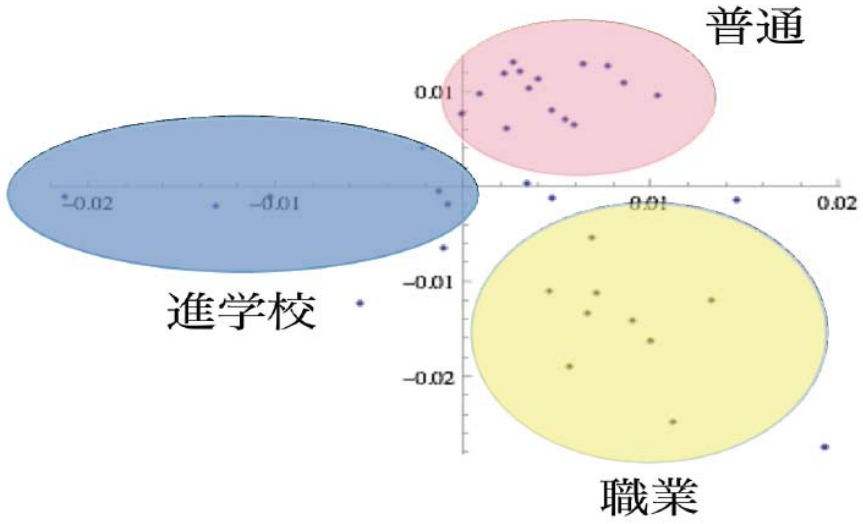

図 2.6 高校の布置 (対応分析)

2.4 高校単位でまとめられた試験成績の分析

高校と科目選択の特徴を把握するには，上に示したように対応分析などのカテゴリカルデータを対象とする分析法が有効と思われる．一方，選択された科目の成績を高校ごとにまとめたデータは，受験者数，平均および標準偏差であり，このようなデータを分析する方法はシンボリックデータ解析の枠組みで研究されているものの，まだ確立されていないと思われる. ここでは, 科目の平均得点および標準偏差が与えられた場合のデータ縮約法(主成分分析法)について検討する.

いま，科目j, $j = 1, 2, ..., p$に対する高校i, $i = 1, 2, ..., n$の受験者数，平均得点および標準偏差をそれぞれn_{ij}, $\bar{x}_{ij}, s_{ij}$とする．このようなデータに対しては，平均得点だけを取り出して主成分分析を適用することも可能であるが，標準偏差などで表されている平均得点の信

頼性を無視していることから，主成分得点の信頼性，すなわち変動の大きさが評価できないことになる．

一般に，平均得点，標準偏差および受験者数からなる多変量データを解析する方法は確立されていないが，各科目の受験者数がm_i =min(n_{ij})であること，高校内での科目間の成績の独立性を仮定すると，各高校の主成分得点は図 2.7 のように求められる．図 2.7 における楕円は，高校ごとに求められた等確率楕円を表しており，横軸が第 1 主成分，縦軸が第 2 主成分である．右方向にある高校ほど総合得点の高い高校であり，高校による学力差をよく反映している．

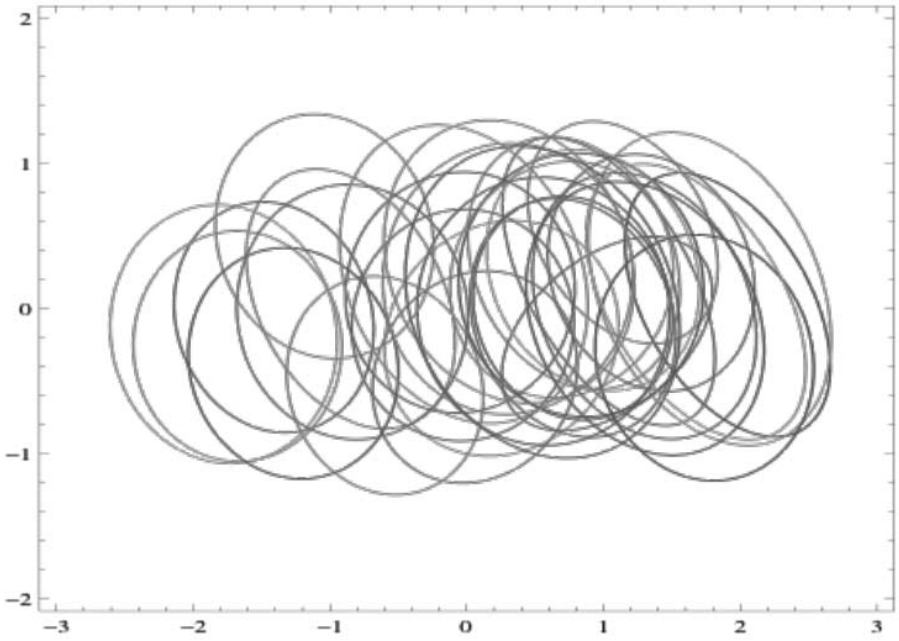

図 2.7 高校の布置(主成分分析)

2.5 おわりに

ここでは，入試のスタンダードを研究する上で必要とされる情報として，センター試験の測定範囲，選抜資料としての有効性などについて検討した．とくに，重要な受験者属性として，高校を取り上げ，その観点からの検討を試みた．属性として高校を取り上げた理由は，高校間の学力差は歴然としてあること，入試の適切さは高校単位で考慮する必要があることなどによる．また，高校単位等，属性単位での多変量データを縮約するためのひとつの方法を示した．なお，本稿は昨年秋の応用統計シンポジウムでの発表原稿に加筆修正したものである．

第 3 章　可変順序重み付マトリックス表示による科目選択データの分析

大津起夫

3.1 調査データの分析における視覚化の効用

社会科学をはじめ行動科学や疫学などにおいて得られる調査データの分析のためにさまざまな方法が考案されてきた．現在における学術的な分析方法の主流は，確率モデルにもとづく分析方法であるが，必ずしも確率モデルには基づかない直観的な分析を行うための方法も，実用的に広く用いられてきた．

現在用いられているデータ分析手法のうち，このような種類の代表的なものは多次元尺度構成法(multi-dimensional scaling,MDS)であろう．計量心理学において発展してきた一次元の尺度構成法は，確率モデルを前提として大小判断のデータから尺度構成を行うという側面が強いが(Togerson,1958)，多次元尺度構成は，アイデアを直観的な指標に置き換えて，分析の意図する構造を求めることを，指標の最適化によって実現するものとみなせる．このような傾向は，特に R. Shepard (1962a, 1962b)や J. B. Kruskal (1964a, 1964b)らによって開発された非計量的な手法においてより強まったとみなせる．

よりのちには，J. Ramsay (1977)のように MDS の枠組みに確率モデルにもとづく誤差分布を導入し，より確率モデルとしての特徴づけを明らかにしようとする研究もあったが，MDS の利用においては，確率モデルとしての特徴は多くの場合は強くは意識されておらず，計算によって得られた結果の視覚的表現が利用されていることが多い．このような手法については，数理統計学の研究者からは，懐疑的な評価が与えられたこともあった[1]．しかしながら，現在では MDS に類似の幾何学的表現は，特にインターネット上にあらわれるサイトや文書間の類似性の分析法として，なかば手法が再発明される形でしばしば利用されている．

記述的な多変量統計的分析法の，もう一つの代表的な手法は対応分析法(correspondence analysis)であろう(Benzecri, J. P., 1992)．対応分析法の計算手続きは離散データのための線形の次元縮約であり，第 2 次大戦後から 1950 年代頃にかけて Benzecri のグループも含め，社会科学におけるデータ分析に携わっていた複数の研究者によって独立に提案された(Guttman in Stouffer et al. 1950; Hayashi,1952)．日本においては，統計数理研究所の林知己夫らの研究グループによる「国民性調査」やその後の海外調査の分析に利用され，「パターン分析の数量化」または「数量化第 III 類」の名で知られている．また，教育社会学の領域ではブルデューの著書(1979)における調査データの分析例が著名なものである(邦訳中では

[1] 坂元編(1976)中の座談会における竹内啓による指摘など．

「照応関係の分析」と記されている).

スタウファーらの著書(Stouffer et al., 1949)の中では，個別の兵士の特質を示した小さなブロック(ドミノ)をマトリックス状に配置し，これらを行(兵士)と列(特性を示す変数)について並べ替えることにより，全体を直観的に把握しやすいように再構成する試みが紹介されている．これはスケーログラム分析として知られる試みである．ガットマンによる尺度構成法の研究は，これらの兵士の帰還順の評価方法の改善に動機づけられている．

類似のアイデアは，フランスの研究者によっても提案された．ベンゼクリ自身の研究ではないが，フランスの高等教育機関における地図とグラフ表示法の研究者であったベルタンによる著書(1967)は，データ視覚化の領域における極めて独創的な著述である．複雑な構造をデータから読み取るためのさまざまな工夫がこの著書において呈示されているが，提案されている手法のうち最も印象的なものが可変順序マトリックスによるデータ分析法である．これは次のような手順を用いてデータを表示する．まず，図の表示を縦横に格子状に区分し，それぞれのセルが対応するマトリックスの要素の数値を表示するように大きさなどが変化する矩形マーク等を表示する．次いで，スケーログラムアナリシスと同様に行および列を並べ替えて全体の構造が見いだせるようにする．ベルタンは，行がサンプル(個体)を表し列が変数を表す場合と，マトリックスが何かの量(生産高，人口等)の分割表である場合の両者について，それぞれ方法を示している．ベルタンの著書の中で頻繁に紹介されている分析法は，次のようなものである．まず，最初に列の表示順を決定し，次いで各行に対応するデータを一枚のカードに表示する．この際，各セルの値の大きさと表示すべきグラフの種類に応じて適切な数値の表示方法を選択する．その後，各行のカードの表示順を，全体の構造がなるべく明確になるように並べ替え，図の全体を作成する．

このような，対応分析を生み出す母体であったマトリックスの入れ替え表示と，対応分析の計算結果を組み合わせることにより，実用上有効な分析を構成することができる．対応分析の結果の解釈は，通常主成分分析や因子分析とのアナロジーで，因子負荷量に相当する解のベクトルの座標を用いて行われることが多い．しかし，以降で分析例を示すように，上記の手法の組み合わせは有益である．

次節では，対応分析法の計算方法について紹介する．第 3 節では，ベルタンの可変順序マトリックスを用いたセンター試験の科目選択データの分析例を示す．第 4 節では，R 言語を用いた可変順序マトリックス描画用の関数について紹介する．

3.2 対応分析(correspondence analysis)

対応分析法(林知己夫の用法では「パターン分類の数量化」や多重対応分析法(multiple correspondence analysis, 国内では「林の数量化 3 類」と呼ばれることが多い)は，変数間の関係を，(1) カテゴリーへのスコア付置，(2) スコアから得られる相関係数の分析，の 2 つの方法によって明らかにしようとするものである．

対応分析が 2 重分割表を対象とするのに対し，多重対応分析は主として各変数が名義尺

度である離散多変量データを対象とするが，これらは計算手続きとしては同一のものである．

2 重分割表のi行j列の数値をn_{ij}とあらわすことにする．また，行数をIとし，列数をJとする．また，対象(被験者)の総数を

$$N = \sum_i \sum_j n_{ij}$$

とする．

次に

$$P_{ij} = n_{ij}/N, i = 1, \dots, I; j = 1, \dots, J$$

とおくと，これらP_{ij}はそれぞれのセルへの相対頻度を表す．また，周辺相対頻度を

$$P_{i.} = \sum_{j=1}^{J} P_{ij} \ , P_{.j} = \sum_{i=1}^{I} P_{ij}$$

と表す[2]．

対応分析の目的は行と列とにそれぞれスコアを与えることにより，表の構造を直観的に理解できるようにすることである．ここで，行スコアをx_i，i = 1,..., Iとし，列スコアをy_j, j = 1,..., Jとすると, i行j列のセルには座標(x_i, y_j) が与えられる. この 2 次元上の位置にP_{ij}の確率が与えられているとすると，これによって定まる分布の平均は次のように表される．

$$\mu_x = \sum_i P_{i.} x_i \ , \mu_y = \sum_j P_{.j} y_j \tag{3.1}$$

以下では，これらの平均μ_xとμ_yとが，ともにゼロに制約されているとする．

この仮定のもとで，それぞれの分散は

$$\sigma_x^2 = \sum_i P_{i.} x_i^2 \ , \sigma_y^2 = \sum_j P_{.j} y_j^2 \tag{3.2}$$

となる．以下ではさらに$\sigma_x^2 = \sigma_y^2 = 1$の制約も，スコアが満たしているとする．

このときxとyとの相関は

$$r = \sum_i \sum_j P_{ij} x_i y_j \tag{3.3}$$

となる．この値は，行と列との関係を表す一つの指標であり，相関rが大きければ，行と列との間に強い関係があるとみなせる．

ここでは，$\{x_i\}$と$\{y_j\}$とに制約を加えてはいるが，一般的にこれらの制約を満たすスコア

[2] これらの値が，標本から求められたもののではなく，多項分布の母数であると考えても，以下と同様に対応分析の手続きを考えることができる．

は一通りではない．これらの中で相関を最大にするものを求めることができるなら，そのスコアがある意味で，データに内在する構造をよく表現しているとみなせるだろう．

そこで，rを制約のもとで最大にするスコアを求めるための計算方法を知る必要がある．主成分分析と同様に特異値分解を利用して，最適なスコアを求めることができる．

ここで，周辺相対頻度を要素とする対角行列(対角線上の要素のみが非ゼロの行列)を次のように定める．

$$\boldsymbol{F} = \mathrm{diag}(P_{1.}, \ldots, P_{I.}), \boldsymbol{G} = \mathrm{diag}\left(P_{.1}, \ldots, P_{.J}\right)$$

また，これらの要素の平方根を要素とする対角行列を，それぞれ

$$\boldsymbol{F}^{1/2} = \mathrm{diag}\left(\sqrt{P_{1.}}, \ldots, \sqrt{P_{I.}}\right), \boldsymbol{G}^{1/2} = \mathrm{diag}\left(\sqrt{P_{.1}}, \ldots, \sqrt{P_{.J}}\right)$$

とする．さらに各々の逆行列(対角行列なので，実際には対角要素を逆数としたもの)を$\boldsymbol{F}^{-1/2}$, および$\boldsymbol{G}^{-1/2}$とする．

相関(3.3)を制約のもとで最大化するスコアは，主成分分析(PCA) におけるのと同様に，次の特異値分解を利用して得ることができる．

$$\boldsymbol{F}^{-1/2}\boldsymbol{P}\boldsymbol{G}^{-1/2} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T} \tag{3.4}$$

ここで$\boldsymbol{D}$の対角要素はすべて非負であり，降順にならんでいるものとする．ここで，変則的だが$\boldsymbol{U}$と$\boldsymbol{V}$の列ベクトルを，それぞれ添え字ゼロから始まるものとし，$\boldsymbol{u}_k, \boldsymbol{v}_k$ ，(k = 0, 1, ..., K)とする．ここで，K = min(I − 1, J − 1) である．また，対角行列$\boldsymbol{D}$の要素についても添え字がゼロから始まるものとし，d_0, d_1, ..., d_kとする．

ここで，

$$\boldsymbol{F}^{-1/2}\boldsymbol{u}_k = \boldsymbol{x}_k, \boldsymbol{G}^{-1/2}\boldsymbol{v}_k = \boldsymbol{y}_k$$

とおくと，$\boldsymbol{x}_1$と$\boldsymbol{y}_1$とが(3.3)を最大にし，制約を満たすスコアになり，$d_1 = r_{max}$となる．また，$\boldsymbol{x}_0$と$\boldsymbol{y}_0$とは，要素が全て 1 のベクトルとなり，d_0 = 1 となる．添え字ゼロに対応するスコアは，重みつき平均がゼロの制約を満たさないため，対応分析の解とはならない．

特異値分解の(3.4)式の左から$\boldsymbol{F}^{1/2}$を乗じ，また右から$\boldsymbol{G}^{1/2}$を乗ずると

$$\boldsymbol{P} = \boldsymbol{F}^{1/2}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}\boldsymbol{G}^{1/2} = \boldsymbol{F}\boldsymbol{F}^{-1/2}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}\boldsymbol{G}^{-1/2}\boldsymbol{G} = \boldsymbol{F}\left(\sum_{k=0}^{K} d_k \boldsymbol{x}_k \boldsymbol{y}_k^{T}\right)\boldsymbol{G}$$

となる．

対応分析の解を考えることは，上の式でk = 0 とk = 1 の部分によって 2 重分割表$\boldsymbol{P}$の重みつきの近似を行っているともみなせる．また，対応分析の利用においては 1 次元より多くの解，つまりk = 2, 3, ...に対応する$\boldsymbol{x}_k$および$\boldsymbol{y}_k$を解釈の対象とすることがある．

これらのk = 2,...に対応する解は，第 1 次元ほど多くの部分を説明しないが，データのより詳細な部分についての情報を与える場合がある(そうでない場合もある)．$\boldsymbol{U}$が直交行列であることから，

$$\boldsymbol{x}_k^T \boldsymbol{F} \boldsymbol{x}_l = 0, (k \neq l)$$

であり，$\boldsymbol{V}$が直交行列であることから，

$$\boldsymbol{y}_k^T \boldsymbol{G} \boldsymbol{y}_l = 0, (k \neq l)$$

が成立する．

スコア$\boldsymbol{x}_2$と$\boldsymbol{y}_2$とは，平均と分散の制約に加え，上の直交条件を満たすもののうち，(3.3)を最大にするものである．またk =3 以降の場合も，同様の性質が成り立つ．

対応分析の第 1 次元のスコア($\boldsymbol{x}_1$と$\boldsymbol{y}_1$) は，周辺度数が極端に偏っていない限り，分割表の持つ構造の主要な部分を表現していると解釈できるが，第 2 次元目以後の解の解釈はかなり難しいことが多い．これは，対応分析の特徴から，データに内在する構造というより，むしろ，手法の特徴に基づく成分が得られることがあるためである(大津，2003).

3.3 センター試験科目選択データの分析

ここでは，センター試験における科目選択データを，対応分析とベルタンの可変順序マトリックス表示を用いて分析した例を示す. 1990 年(平成 2 年)と 2009 年(平成 21 年)に実施されたセンター試験における社会科目(地理・歴史・公民)および理科の科目選択の状況を表示する．

図 3.1 は，1990 年に実施されたセンター試験(第 1 回)における，社会(1 コマ)と，理科 3 科目の選択者数を可変順序マトリックスによって表示したものである．表示の対象としたのは，センター試験のいずれかの科目を受験した者であり，すべての科目を欠席したものは除いてある．各行の幅(表側の幅)は，社会の科目(欠席も一つ区分として扱う)の周辺度数に比例する大きさにとってある．一方，図の列は理科の選択科目を示している．理科は 3 つの時間にわたって実施されたため，最大 3 科目の受験が可能である．周辺度数が 5000 人以下の区分は表示から除いてあるため，2 科目以上の選択パターンは「物理，化学」および「化学，生物」の組み合わせのみが表示されている．1990 年に一科目以上センター試験を受験した者は約 40 万 8 千人いたが，ここでの表示対象となった者はこれらのうち約 39 万 7 千人である．各セルの横幅は該当する区分における理科科目の選択率を示しているので，各セルの面積は該当する 2 重分割表の該当人数に比例する．また，各列中央近くに表示されている破線は，社会科目の選択と理科科目の選択とが統計的に独立であったときに期待される値を示しており，これを超える部分の色を変えて表示してある．また，行と列の表示順は対応分析によって得られた第 1 次元の解の大小順に基づいている．

社会科科目の選択と理科の科目選択については，これらの 2 つの領域での関係が生じる必然性は特にはないように思われる[3]．しかしながら，図 3.1 に見られるように，実際にはこれらの間にかなり明瞭な関係が存在している．

[3] ただし 2009 年においては理系志願の受験生では，理科 2 科目以上と社会 1 科目を受験し，一方，文系志願の受験生では，理科 1 科目と社会 2 科目を受験することが多いと推測される．

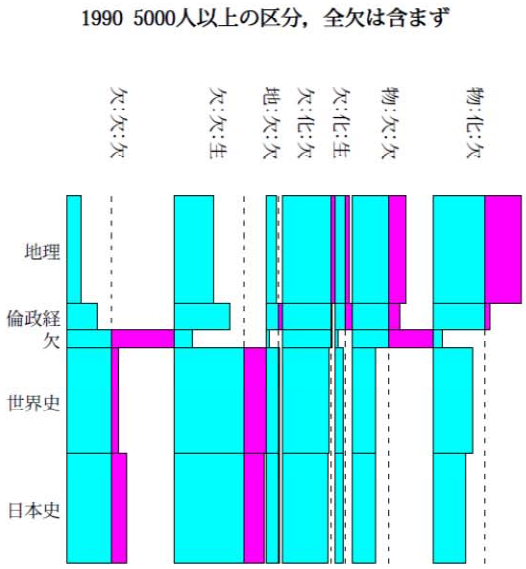

図 3.1　1990(平成 2)年センター試験
社会と理科の科目選択状況

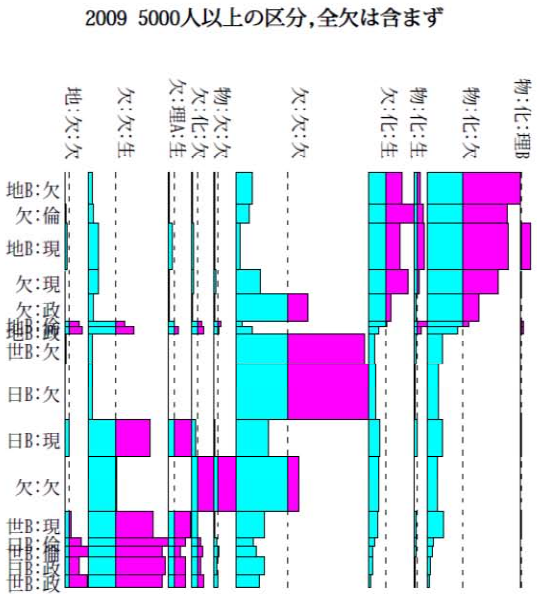

図 3.2　2009(平成 21)年センター試験
社会と理科の科目選択状況

図 3.2 は，2009(平成 21)年のセンター試験について図 3.1 と同様の表示を行ったものである．1990 年には 1 科目のみであった社会科目が，「地理・歴史」と「公民」の 2 コマ実施されるようになったため，行のパターン数が増加している．また，理科についても「理科総合Ａ」および「理科総合Ｂ」が追加されているために，科目の選択のパターンは増加した．ここでも図 3.1 と同様に，周辺度数が 5000 人に満たないものは，表示から除いた．2009 年には，1 科目以上を受験した者は，50 万 8 千人弱であるが，ここで表示対象となった受験者数はこのうち約 47 万 2 千人である．

ここでも，社会(地理・歴史」および「公民」)と理科の科目選択の間にかなり明瞭な関係の存在することが分かる．また，これらの関係性は図 3.1 より明確になっているように見える．実際，対応分析によって得られた特異値(相関係数)は，図 3.1 では 0.315 であるが，図 3.2 では 0.543 であり，この傾向のあることが支持される．

さらに，類似の目的をもつグラフ表現の方法であるモザイクプロット(Hartigan & Kleiner, 1984; Friendly, 1994)と，連関(association)プロット(Cohen, 1980)との比較を行う．図 3.3 はモザイクプロットによって図 3.1 と同一のデータの表示を行ったものである．各列が，理科の科目選択パターンに対応しており，各列の幅はそれらの周辺度数に比例している．さらに各列の内部が，各群内における社会の科目選択率に応じて分割されている．図 3.1 と同様に，各セルの面積は，それぞれの区分の受験者数を表している．また，(白黒表示では識別できないが)行と列の独立性の仮説からの違反の程度によって色分けがなされており，正の残差がある場合には青色が，また負の残差はオレンジ色で表示されている．図 3.3 で表現されている情報は，図 3.1 で表現されているものと同等であるが，データの持つ構造の視認性は，

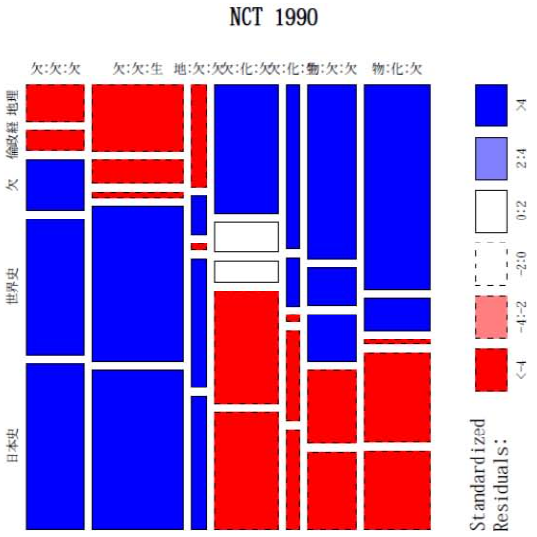

図 3.3　1990(平成 2)年センター試験
モザイクプロット

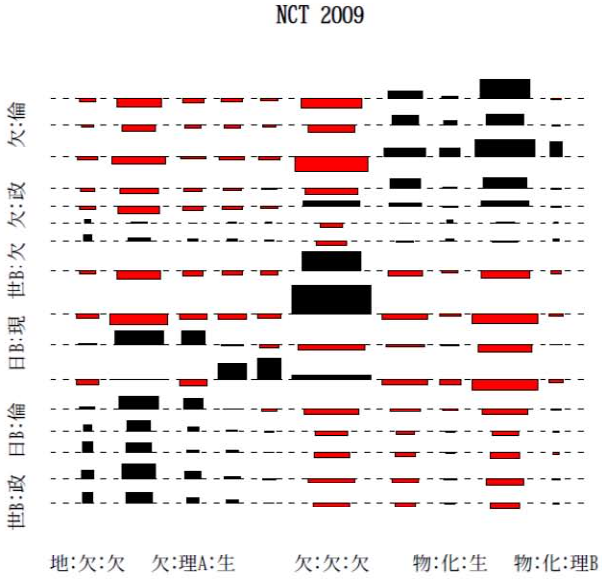

図 3.4　2009(平成 21)年センター試験
連関(association)プロット

図 3.1 のほうが高いように思える．

図 3.4 は, 図 3.2 と同一のデータを連関プロットを用いて表示したものである. こちらは, 各セルの度数そのものは表示せず，独立性の仮定からの残差のみを表示対象としている. 各セルに表示された矩形の面積が残差の大きさを表している．ここで，各セルの観測度数をn_{ij}とし，独立性の仮定の下での予測値をμ_{ij}とする．各矩形の幅は$\sqrt{\mu_{ij}}$を示し，矩形の高さは$(n_{ij}-\mu_{ij})/\sqrt{\mu_{ij}}$を示す. このため，それぞれの矩形の面積は$n_{ij}-\mu_{ij}$を表すことになる. 連関プロットによっては，独立性の仮定からの逸脱の程度がよく把握されるが，ベルタンの方法に基づく図 3.2 においても同一の情報がよく把握されており, さらにn_{ij}の大きさ自体の把握も容易である.

これらの例を検討すると，モザイクプロットおよび連関プロットのいずれと比較しても，時代的にはより古いベルタンの方法の方が優れているように思われる.

3.4 可変順序マトリックス表示プログラム

以下に，著者が作成した可変順序マトリックス表示のための R 言語による関数利用方法を示す．コードは www.rd.dnc.ac.jp/~otsu/Rcodes で公開している．以下に示したものは，可変順序マトリックス表示の機能のみを持つものである. 対応分析の計算には，R-2.11 に含まれている MASS パッケージ中の corresp 関数を利用することができる．また本文中のモザイクプロットは，R の mosaicplot 関数を用い，連関プロットには，assocplot を用いた.

機能

この関数は 2 重分割表データを入力し，Bertin(1981)の可変順序マトリックス表示を行う．

引数

```
bertin.q <- function(x    , title= deparse(substitute(x))
                          , rowlabels= dimnames(x)[[1]]
                          , columnlabels=dimnames(x)[[2]]
                          , roworder=seq(dim(x)[1])
                          , columnorder=seq(dim(x)[2])
                          , bardirection="h"
                          , rowheadermargin=0.15
                          , columnheadermargin=0.15
                          )
```

x: 表示対象の行列．行列の要素は，それぞれの区分の頻度(または相対頻度)を表す数である．

title: 主タイトル

rowlabels: 各行のラベル．x の行数と同じ長さの文字ベクトル．

columnlabels: 各列のラベル．x の列数と同じ長さの文字ベクトル．

roworder: 行の表示順．ここで指定された数値の大小順に行が描画される．下が 1 番小さい順位に対応し，上が大きい順位に対応する．デフォールトでは x の行の順．

columnorder: 列の表示順．ここで指定された数値の大小順に列が描画される．左が 1 番小さい順位に対応し，右が大きい順位に対応する．デフォールトでは x の列の順．

bardirection: 各セルに表示される小さな棒グラフの方向の指定．"h"か"v"を指定する．"h"(デフォールト)を指定すると横方向，"v"では縦方向に描画する．

rowheadmargine: 表示画面における，行ラベルの表示領域の比率．

columnheadmargine: 表示画面における，列ラベルの表示領域の比率．

利用例

先頭の>は，R システムのプロンプトを表し，利用者は入力する必要はない．次のものは Goodman(1985)に引用されている社会経済地位(SES)と精神の健康状態(health)の関係を表す 2 重分割表である．

```
> goodman.mental        (5 × 6 の行列)
                ses
health          A     B     C     D     E     F
   well        64    57    57    72    36    21
   mid         94    94   105   141    97    71
   moderate    58    54    65    77    54    54
   impaired    46    40    60    94    78    71
> bertin.q(goodman.mental)
```

引用文献

Bertin, J. (1967). *Semiologie Graphique. Les diagrammes, les reseaux, les cartes. With Marc Barbut (et al.)*. Paris : Gauthier-Villars. (English translation (1983). *Semiology of Graphics* by William J. Berg., The University of Wisconsin Press, Madison Wisconsin.)

Bertin, J. (1977). *La graphique et la traitement graphique de l'information*, Flammarion, Paris, (English translation (1981). *Graphics and graphic information-processing*, by Berg, W.J., Scott, P., Walter de Gruyter, Berlin.)

Benzecri, J. P. (1992). *Correspondence analysis handbook*, CRC Press.

Bourdieu, P. (1979). *La distinction critique sociale du jugement*, de Minuit. 石井洋二郎訳(1990). ディスタンクシオンI,II, 藤原書店.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistical-Theory and Methods*, **A9**, 1025-1041.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American statistical association*, **89**, 190-200.

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Annals of Statistics*, **13**, 10-69.

Hartigan, J. A., and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, **38**, 32-35.

Hayashi, C. (1952). On the prediction of phenomena from quantitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, **3**, 69-98.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1-27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115-129.

大津起夫 (2003). 調査データからの推論：実践的入門, 竹村彰通(他編) 「統計科学のフロンティア10　言語と心理の統計」第III 部, 岩波書店.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, **42**, 337-360.

坂元平八著, 竹内啓編 (1976). 統計学の未来-推計学とその後の発展, 東京大学出版会

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, 125-140.

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika*, **27**, 219-246.

Stouffer, S. A., Guttman, L., Suchman, E., Lazarsfeld, P. F., Star, A., Clausen, A. (1950).

Measurement and prediction, (Studies in social psychology in world war II vol.IV), Princeton university press.

Togerson,W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

第 4 章　大学入学者選抜における調査書の利用について

倉元直樹

本稿は, 平成 24(2012)年 3 月 1〜3 日に行われた「大学入試のためのスタンダードの作成」研究会において「大学入学者選抜における調査書の利用について」と題して行われたプレゼンテーションの内容をまとめたものである.

学力試験の問題点として，一発勝負で取り返しがつかないことが挙げられる．一方，調査書は受験生の日常的活動記録の集積であり，選抜資料としてより妥当性が高いような印象を与える．そのため，政策的にも調査書の利用は推奨され続けてきたが，測定尺度としては問題が多い．調査書依存の選抜方法の問題点を整理すると同時に，大学入試における適切な調査書利用の方法を考える．

なお，本稿は倉元・金澤 (2010，2009a) に修正を加えたものである．

4.1 大学入学者選抜制度の変更とグレード・インフレーション

4.1.1 問題の所在

大学入試の選抜資料として，調査書への期待が高まっている．例えば，全国高等学校長協会からの中教審大学分科会への要望書(全国高等学校長協会長，2008)でも三つの要望項目の一つに取り上げられ，中教審学士課程答申(中央教育審議会，2008)では積極的活用が提唱された．平成 23(2011)年度入試からは調査書の様式が一新された．その一方で，選抜資料としての構造的問題は手つかずのままである．

調査書に含まれている情報の中では比較的取り扱いの容易な学力指標である評定平均値についても，評定値の基準となる測度が存在しないために様々な問題が生じている．例えば，学校間格差の存在は周知の事実である．さらにそれが集積して，結果的にコースや地域の格差として現れている(倉元・西郡・石井，2010).

倉元・金澤(2009a,b)は合否入替りの考え方を援用して調査書の評定平均値を選抜資料の中に無理なく取り入れる方式を考案したが，実際には，高校教員からは調査書の積極的活用は望まれていない(倉元・當山・西郡・石井，2009).

外的教育環境の影響を受け易いことも選抜資料としての調査書の信用度の低さの一因と思われる．もしも，大学入試で利用されるとなれば，グレード・インフレーション(grade inflation)の発生は避けがたい．倉元・西郡・石井(2010)は個別大学の AO 入試が評定平均値に与える影響を見出すことを試みたが，証拠は得られなかった．実際，特定大学のみの入試制度の影響力を析出するのは技術的に極めて困難であろう.

そこで，本研究では，評定平均値の分布における経年変化を手掛かりに大学入試制度の変化が高校調査書に与える影響を分析することとした.

4.1.2 分析方法

国立 A 大学平成 m 年度と 8 年後の平成 n 年度入試において，x 学部の複数の入試区分の選抜資料として提出された調査書から「学習成績概評」欄の「成績段階別人数」および学校名，コース，卒業(または，卒業見込)年度を抽出した．その結果，930 校，1,778 種類のデータが得られた．複数年度のデータが得られたのは，そのうちの 456 校であった．なお，本研究で分析の対象とするものの中には，受験者の個人情報の類は一切含まれていない．

「概評 A」段階と評価された生徒の比率について，倉元・川又(2002)の SS 値を用いて分析した．SS 値とは，成績段階別人数分布に基づき，各高校の評定平均値の A 段階と B 段階の境界値である「4.25」が，評定平均値に関わる全データの分布に照らすと何点に相当するかを相対的に示す指標である．すなわち，A 比率が高い「甘い評価」であるほど SS 値は小さくなる．

倉元・川又(2002)にならい，同じ高校の異なるコースは独立した学校とみなした．地域区分は倉元(2007)に基づいて分類した．また，高校ランクは平成 m 年度のデータを基に，直近時期において，中村(1999，2002)に基づいて判定した．なお，一部カテゴリーを合併して分析に用いている．

4.1.3 結果

各データ値を独立とみなして分析を行った．

まず，全評定に含む「概評 A」の比率の算術平均は 24.5%であったが，最大値は 95.1%，最小値は 1.3%と大きく散らばっていた．SS 値に換算した場合，平均値は 4.28，最大値 5.10，最小値 2.94 であった．なお，最大値が 5 を超えたのは，モデルの限界である．

卒業年度は「1: m-2 年度まで」が 7.3%，「2: m-1 年度」が 20.7%，「3: m 年度」が 31.7%，「4: m+1～n-2 年度」が 2.3%，「5: n-1 年度」が 12.1%，「6: n 年度」が 26.0%であった．

高校ランクは「1: A1 以上」が 20.0% (年度を区別せずに学校を単位としたとき 13.8%，以下同じ)，「2: A2-A3」が 21.7% (16.8%)，「3: B1」が 19.9% (18.8%)，「4: B2」が 15.7% (17.3%)，「5: B3」が 17.8% (24.7%)，「6: C 以下」が 4.9% (8.5%) であった．

コースは 90.1% (87.5%) が普通科，設置者では 68.6% (67.6%) が国公立であった．以下，「卒業年度」「高校ランク」「コース」「設置者」の四つの変数を説明変数として用いることとした．

なお，地域は「1: 第 1 群」が 20.3% (15.5%)，「2: 第 2 群」が 16.7% (15.3%)，「3: 第 3 群」が 38.3% (38.9%)，「4: 第 4 群」が 24.7% (30.3%) であった．

次に，説明変数を絞りこむための事前分析として，倉元・西郡・石井(2010)によって評定平均値の分布に関係があることが判明している「コース」と設置者を用いて 2 元配置の分散分析を行った．その結果，交互作用はなく，「コース」に加えて「設置者」の主効果も辛うじて有意であった ($F[1,1751] = 164.07, p<.0001; F[1,1751] = 5.53, p<.05$) が，シェッフェ法による多重比較の結果，国公立と私立に差は確認されなかった．自由度が大きく，バラン

スが悪い(私立の理数科がわずか 16 校)ことから，設置者要因を除外し，「コース」は「普通科」に絞って分析を進めることとした．

卒業年度に関しては，「4: m+1～n-2 年度」が少ないので以後の分析からは除き，「卒業年度 (5 水準) ×高校ランク (6 水準)」の 2 元配置の分散分析を行うこととした．その結果，1,539 種類のデータ(86.6%)が以降の分析に含まれることとなった．

分析結果からは，交互作用は見られず，卒業年度，高校ランクともに主効果が有意であった (F[4,1509] = 44.71, p<.0001; F[5,1509] = 22.77, p<.0001)．シェッフェ法による多重比較の結果，卒業年度は「1」「2」「3」と「5」「6」の各対に 5%水準で有意な差がみられ，新しい年度の SS 値が小さい，すなわち，評定が甘くなっていることが判明した．

高校ランクは「1」とそれ以外，「3」と「5」の間に 5%水準で有意な差がみられた．おおむね，ランクが高くなるにつれて SS 値が小さくなっているが，「3: B1」の SS 値平均が「2: A2-A3」よりも小さく，高校ランクの順序性が崩れていることが分かった (表 4.1 参照)．

表 4.1 「卒業年度」と「高校ランク」の水準別 SS 値平均

	卒業年度					高校ランク					
水準	1	2	3	5	6	1	2	3	4	5	6
SS 値平均	4.38	4.37	4.34	4.25	4.20	4.21	4.32	***4.30***	4.33	4.36	4.37

次に数量化 I 類を用いて「卒業年度」「高校ランク」「コース」「設置者」の四つの変数を外的基準として SS 値を予測し，その残差を用いて地域要因に関する分析を行った．
予測値と実測値の重相関係数は .46，カテゴリー値は表 4.2 の通りである．卒業年度と高校ランクのカテゴリー値に関しては，概ね表 4.1 と整合的である．

表 4.2 数量化 I 類におけるカテゴリー値

	卒業年度						高校ランク						コース		設置者	
水準	1	2	3	4	5	6	1	2	3	4	5	6	普	理	公	私
値	.089	.073	.039	-.032	-.044	-.110	-.099	.004	-.011	.030	.072	.086	.023	-.222	.009	-.029

次に，残差(評定が「甘い」と「＋」の方向に寄与)について地域を要因として 1 元配置の分散分析を行った．その結果，地域の主効果は有意 (F[3,1731] = 30.72, p<.0001) で，多重比較の結果，残差が大きい順に「1」，「4」，「2」，「3」であり，「4」と「2」の間以外は全て 5%水準で有意であった．

さらに，卒業年度の「2」か「3」，および，「5」か「6」の双方にデータが存在する学校のみを抽出し，それぞれの時期の残差平均を求め，さらに度数が 5 以上存在する 25 都道府

県について平均値を求めた (図 4.1 参照). 時期による顕著な変化は見られないが，都道府県で違いがあり，特定都道府県の影響で地域差がもたらされている可能性が示唆された.

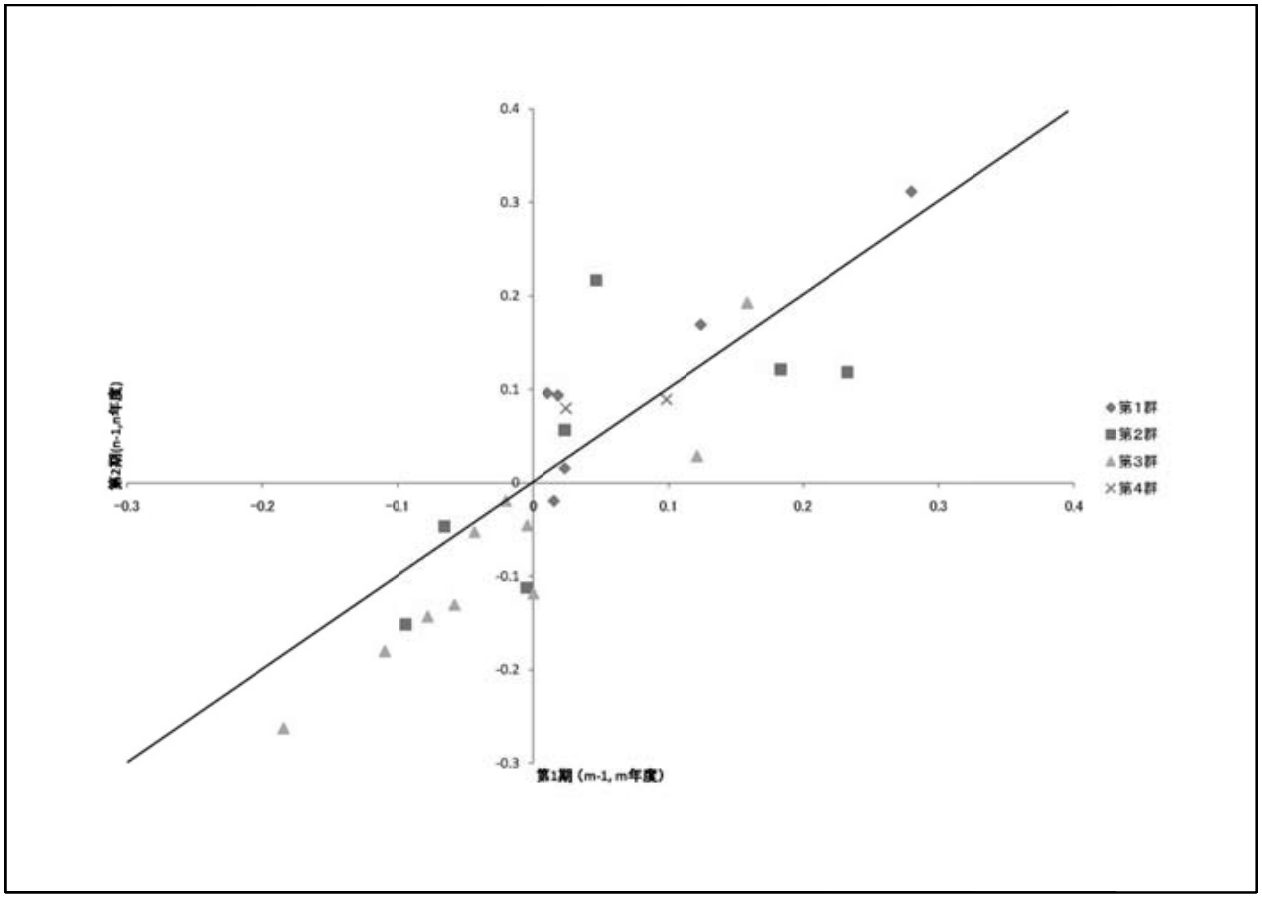

図 4.1 残差の都道府県別平均値

4.2 大学入学者選抜における調査書の活用法

4.2.1 ユニバーサル化時代における大学入学者選抜

総体として大学入試の選抜性が薄れ，制度的な学力維持装置としての機能が弱まっているという認識がある．少子化に伴い，高校から大学への門戸が全体として広くなったことは事実だが，個別大学の視点に立てば，逆に大学入試のあり方がかつてないほどに重みを増している．それは，少子化の進行と並行して入試の多様化政策が進められてきたことによる.

入試の多様化とは，見方を変えれば，高等教育分野の市場主義化，規制緩和政策と捉えられる．すなわち，大学入試における個別大学の自由裁量権が拡大したと同時に，自己責任を負わなければならなくなった.

かくして，大学入試にも教育機関としての質を維持向上するための戦略が各大学に求め

られ，入試設計と遂行能力が問われる状況となっている．わが国より先にユニバーサル段階を迎えた米国の例(木村，2008)からも分かるとおり，大学は二極分化していくだろう．その際，「受験生を育てる大学」と見られるか「学力度外視で志願者集めをする大学」と見られるか(倉元，2008)が，消費者たる志願者側から見た重要な大学の選別指標の一つとなることに疑いはない．

4.2.2 多様化政策の中での調査書の位置付けと構造的問題

多様化政策の中で重要な鍵を握るのは，「選抜方法の多様化」，「評価尺度の多元化」である．その中で調査書の活用は常に推奨され続けてきたが，評価尺度としての構造的欠陥があまりにも大きすぎて，積極的に利用されてこなかった．例えば，「評価基準の恣意性」，「学校間格差の存在」，「入学者選抜制度の波及効果 (ないしは遡及効果)」， 受験生以外の要因の介在による「測定の妥当性・信頼性の問題」等，選抜資料としての調査書の欠陥はあげつらえばきりがない．実際，受験者側の利益に立っている高校教員からも，調査書の積極的利用は望まれていない(倉元・當山・西郡・石井，2009)．大学入試を巡る歴史と現状を知って知らずか，未だに繰り返される，漫然とした調査書重視政策は改められるべきである．

その一方で，それはさておき，大学入学者選抜の現場では否応なしに調査書を使った選抜を迫られる可能性をあらかじめ考えておく必要がある．その際，調査書を具体的にどのように位置付け，どのように活用するのかは，大学の裁量権を拡大させる今の教育政策の流れにあっては，大学が問われる入試設計能力の一つである．そこで，本研究では，調査書の活用方法に関して一つの具体的なモデルを提示する．

4.2.3 調査書を用いて何を評価するのか

本研究では入試で選抜が成立している状況を前提とする．更に，選抜方法の設計上，学力検査等の調査書以外の評価尺度が主たる選抜資料として機能しているとする．また，これ以後，「学力」を主要な選抜の指標としている状況を前提として論を進める．

調査書を用いて評価することができる可能性がある情報は，二つに大別することができる．「学業成績以外」と「学業成績」である．このうち，学業成績以外の評価には必ず他者の力量が介在することになる．すなわち，端的に言えば，調査書を作成する教員の熱心さや文章能力が評価に決定的な影響を及ぼす．一方，学業成績にも様々な誤差要因が混入するが，前者よりは客観性と信頼性が担保される条件が整っている．

そこで，本研究では学業成績，具体的には「評定平均値」を利用するケースについて考える．評定平均値は「すべての教科･科目の評定の合計数」を「すべての評定数」で割って算出する(例えば，文部科学省高等教育局長，2008)ことと定められている．

個々の評定の基準について評価対象となる集団を超えた統一性がない以上，評定平均値をそのまま学力指標と考えることはできない．そこで，評定平均値が表す特性について以下のような仮定を置くこととした．

(1) 評定平均値が高いことは高校で課される全ての科目に真摯に取り組んできたことを示す
(2) 同じ程度の学力を持つ受験生であれば，教育水準の低い高校に所属する者の方が，成績が良い

そして，「学力に実質的な差がないならば，相対的に水準の低い学校でまじめに頑張ってきた受験生を優先的に合格させる」という選抜方針を立てることとする．このような選抜方針を個別大学の入試制度の中で，無理なく具体化するための方法論を提示することが本稿の目的である．

4.2.4 総合得点方式と合否ボーダー層

「合否ボーダー層」とは，選抜ラインギリギリに位置する受験生である．平・池田(1994)等が操作的定義を試みているが，本質的なものではない．そこで，本稿では，上述の選抜方針にのっとり「合否付近で学力に実質的な差がない集団」と考えることとした．

ところで，合否ボーダー層に対して主たる選抜指標とは異なる次元の尺度を用いて選抜を行う制度は，既に存在している．オーストラリア・クィーンズランド州の大学入学者選抜制度では総合指標である OP (Overall Position)で定員近くまで合格者を決定した後に, 同じ OP の志願者から専門分野と関係が深い FP (Field Position)の上位から順に定員まで合格者を出す制度を取っている(山村・平，1998)．しかし，クィーンズランド州の制度は，そのままではわが国にはなじみにくい．

わが国では複数の選抜資料を数値化して加算した総合得点に基づき，上位者から順に合格者を選ぶ制度が基本である．情報開示制度の普及を考えるとそれ以外の原理は説明責任の上で大きなリスクが伴う．総合得点方式は池田 (1992) の加算方式と同一である．池田(1992)によれば，総合得点方式はミスによる得点のロスを他で補うことが可能な「相互補完原理」に基づく測定論的に優れた選抜制度と言うことができる．

次に，総合得点方式を前提として合否ボーダー層をどのように操作的に定義するのかが問題となる．本研究では，古典的テスト理論(CTT: Classical Test Theory)の信頼性概念を適用することを提案する．具体的には，何らかの方法で信頼性係数を推定し，例えば「合格最低得点を 90%信頼区間に含む範囲」を合否ボーダー層とすることが考えられる．

4.2.5 合否入替りと配点

1990 年代後半に国立大学入学者選抜研究連絡協議会 (入研協，現在の全国大学入学者選抜研究連絡協議会の前身)で盛んに行われた「合否入替り」に関する研究は，共通 1 次と 2 次試験の二重負担論との関係といった時代限定的テーマが中心のためにその後はさほど普及しなかった．

倉元・西郡・木村・森田・鴨池(2008)は，忘れ去られようとしていたこの分析方法に着眼

し，合否入替り法を得点調整問題に援用した．この考え方は，実際の意思決定場面においては同時に加算される得点をあえて逐次的に扱って，合否決定のプロセスにおける得点の構造を見えやすくしたことに特徴がある．

本稿では，倉元他(2008)の方法に倣い，学力指標で暫定的合格者を決定してそれに調査書得点を加えて合否入替りを観察する．すなわち，本稿の方法論によれば「合否ボーダー層に対する他次元の評価尺度としての調査書を利用した選抜」は結局は「評定平均値の配点」という教育測定論関係の専門家以外の一般関係者にも十分理解できる，なじみ深い問題に帰着させることが可能となるのである．

4.2.6 数値例

国立 T 大学の H 年度入試における文系 a 学部，理系 b 学部の入試データに対して，調査書得点を加えて合否入替り法を適用した．合否ボーダー層は上述の方法で定めた．信頼性係数の推定値は教科ごとの標準化得点を元に算出した α 信頼性係数を用いた．

a 学部の実質倍率は 2.6 倍，α 信頼性係数は.78，合否ボーダー層に該当したのは合格者の 80.0%，不合格者の 52.1%であった．一方，b 学部の実質倍率は 2.4 倍，α 信頼性係数は.90，合否ボーダー層は合格者の 62.3%，不合格者の 33.3%であった．

表 4.3 は両学部の入試科目の影響力に関する分析結果である．いずれも個別試験の実質的な影響力(共分散比)が見かけ(配点比)よりも大きい．

表 4.3 T 大学 H 年度における各教科の影響力分析結果

試験		センター試験						個別試験				
教科		国語	地/公	数学	理科	英語	合計	国語	数学	理科	英語	合計
a 学部	配点比	8.7%	13.0%	8.7%	8.7%	8.7%	47.8%	17.4%	17.4%	-	17.4%	52.2%
	共分散比	4.8%	9.4%	7.9%	8.6%	7.6%	38.2%	12.3%	29.2%	-	20.3%	61.8%
b 学部	配点比	6.5%	3.2%	6.5%	6.5%	6.5%	29.0%		25.8%	25.8%	18.8%	71.0%
	共分散比	2.2%	2.0%	6.6%	5.6%	4.6%	21.0%	-	31.5%	31.2%	16.3%	79.0%

調査書の配点について，(1) 最小配点科目並，(2) センター試験 1 科目並，(3) 個別試験 1 科目並，(4) センター試験並，(5) 個別試験並，としたときの合否入替りを表 4.4 に示す．入替り率の定義は清水 (1995) にしたがった．

表 4.4 調査書得点を加算した場合の合否入替り

学部	a 学部					b 学部				
調査書配点	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
配点比	4.2%	8.0%	14.8%	32.4%	34.3%	3.1%	6.1%	16.2%	22.5%	41.5%
共分散比	0.3%	0.5%	1.0%	3.1%	3.4%	0.1%	0.2%	0.6%	0.9%	2.3%
入替り率	3.4%	6.8%	12.1%	23.7%	25.6%	1.4%	2.9%	7.2%	7.2%	20.3%
圏外からの浮上率	0.0%	0.0%	0.0%	1.0%	2.4%	0.0%	0.0%	0.0%	0.0%	1.4%

両学部とも，調査書と学力検査総合得点との相関係数は.35 程度となっている．調査書の標準偏差が相対的に小さいことを考えると，正規分布仮定の下で理論的に求められる入替り率(Kikuchi & Mayekawa, 1995)よりは大きな値が得られた印象である．すなわち，相当に配点を大きくしてもボーダー層圏外からの逆転浮上はあまり起らない．ただし，実際には調査書の影響力は実際より大きく受け止められる傾向がある(倉元・山口・川又, 2007)ため，配点の公表が志願者層に与える影響を慎重に吟味する必要があるだろう．

本研究の定義によると結果的に合否ボーダー層をかなり大きく取ることとなったが，その点に関しては検討の余地がある．さらに，実際には肝心の逆転浮上者と逆転不合格者のプロフィールについて，求める学生像や学生募集戦略の観点から詳細に検討した上で，適切な配点を決定する必要性があるだろう．

引用文献

中央教育審議会 (2008). 学士課程教育の構築に向けて (答申)，文部科学省.

池田央 (1992). テストの科学―試験にかかわるすべての人に― 日本文化科学社.

Kikuchi K., & Mayekawa, S., (1995). On the Sampling Distribution of Swap- Rate, *Behaviormetrika*, **22**, 185-204.

木村拓也 (2008). 格差を広げる入試制度はどのように始まったのか？―日本におけるオープンアドミッション・システムの淵源―，クオリティ・エデュケーション，第 1 巻特集 再チャレンジ可能な社会の条件，91-113.

倉元直樹 (2007). 東北大学入試広報戦略のための基礎研究 (1) ―過去 10 年の志願者数・合格者数等から描く「日本地図」―, 東北大学高等教育開発推進センター紀要, **2**, 9-22.

倉元直樹 (2008). 大学入試問題 安易な「再利用」に走るな, 讀賣新聞, 解説面, 論点, 2008 年 4 月 10 日.

倉元直樹・金澤悠介 (2009a). 大学入学者選抜における調査書利用の考え方―「合否入替り」法を利用して―，日本高等教育学会第 12 回大会発表要旨集録，100-101.

倉元直樹・金澤悠介 (2009b). 大学入試における「評価尺度の多元化」に則った調査書利用法に関わる一考察，日本テスト学会第 7 回大会発表論文集，150-153.

倉元直樹・金澤悠介 (2010). 大学入学者選抜における調査書利用の考え方(2) ―grade inflation の問題を中心に―，日本高等教育学会第 13 回大会発表要旨集録，88-89.

倉元直樹・川又政征 (2002). 高校調査書の研究 ―「学習成績概評Ａ」の意味―，大学入試研究ジャーナル，**12**，91-96.

倉元直樹・西郡大・石井光夫 (2010). 選抜資料としての調査書，大学入試研究ジャーナル，**20**，29-34.

倉元直樹・西郡大・木村拓也・森田康夫・鴨池治 (2008). 選抜試験における得点調整の有効性と限界について －合否入替りを用いた評価の試み－，日本テスト学会誌，**4**，

136-152.

倉元直樹・當山明華・西郡大・石井光夫 (2009). 東北大学ＡＯ入試における調査書利用の考え方と高校側の意見，東北大学高等教育開発推進センター紀要，**4**，147-159.

倉元直樹・山口正洋・川又政征 (2007). 受験生からみた東北大学工学部の AO 入試，大学入試研究ジャーナル，**17**，43-49.

文部科学省高等教育局長 (2008). 平成 21 年度大学入学者選抜実施要項 (平成 20 年 5 月 29 日 20 文科高第 140 号).

中村忠一 (1999). 全国高校格付け 2000 年版，東洋経済新報社.

中村忠一 (2002). エリートへの道は中学・高校選びで決まる，エール出版社.

清水留三郎 (1995). 入学者選抜における試験の効果の評価―合否入替り率等を中心に (第 1 報)―，大学入試研究ジャーナル，**5**，1-4.

平直樹・池田輝政 (1994). 入試科目の効果に関する新しい評価法, 大学入試研究ジャーナル, **4**，40-44.

富永倫彦 (2005). 入学者選抜における調査書利用の実態調査，大学入試研究ジャーナル, **15**, 85 -91.

山村滋・平直樹 (1998). オーストラリア・クィーンズランド州の高校成績を利用した大学入学者選抜制度，大学入試研究ジャーナル，**8**，35-40.

全国高等学校長協会長 (2008).「学士課程教育の構築に向けて（審議のまとめ）」への意見，全高長第 20 号，平成 20 年 5 月 12 日.

第 5 章　局所独立性指標による IRT 適用可能性の測定

橋本貴充

要旨

項目反応理論(IRT)を用いたテスト項目の分析がしばしば行われるが，適用対象が IRT の前提条件を満たすかどうかの検討が行われることは少ない．本報告では，IRT の前提条件の一つである局所独立性が満たされるかどうかを，局所独立性指標の一つである LCI 指標を用いて測定し，項目の再利用に役立てる方法を提案する．具体例として，平成 21〜23 年度大学入試センター試験「世界史 B」を東京都内の国立大学 1 年生が解答したデータへの適用を試みる．

5.1 はじめに

実施された大学入試の問題を分析し，その性質を明らかにすることは，大学入試の標準化に欠かせない過程の一つである．中でも，項目反応理論(Item Response Theory, 以下，“IRT”とする．Lord & Novick, 1968; 村木, 2011)を用いたテスト項目の分析を行うことが近年増加しつつあり，吉村(2005)の作成した大学入試センター試験の試験問題統計情報データベースにも，IRT のロジスティック・モデル(Brinbaum, 1968)のパラメータ推定値が変数に加えられている．ロジスティック・モデルは IRT で標準的に用いられているモデルであるが，適用の際，以下の 3 つの仮定が前提とされている．すなわち，(1) 項目の正誤が受検者の能力という潜在変数で説明・予測できること，(2) 項目に正答する確率が能力潜在変数の単調増加関数(この関数を“項目特性関数”もしくは“項目特性曲線”という．)として記述できること，(3) 能力潜在変数を所与として，各項目への正答確率が条件付き独立になること，である．特に(3) の仮定を“局所独立性”という．局所独立性の仮定が満たされないテストに IRT のロジスティック・モデルを適用すると，項目パラメータの推定値が偏る(佐野, 2009)ことが知られている．しかし，これまでの IRT を用いた項目分析で，局所独立性が検討されることはほとんど行われてこなかった．本報告では，局所独立性の指標である LCI (Latent Conditional Independence)指標(Hashimoto & Ueno,2011)を用いて項目間の局所独立性を検討し，IRT 適用の適否を判断する方法を提案する．

5.2 方法

5.2.1 従来の局所独立性測定方法の問題点

局所独立性は，能力潜在変数を所与とする条件付き独立性である．一般に，能力の高い受検者はテストのどの項目にも正答し，能力の低い受検者はどの項目にも誤答するため，テスト項目間には能力潜在変数を介した相関関係がある．したがって局所独立性を測定す

る様々な手法が提案されてきた(Chen & Thissen, 1997; Glas & Suarez Falcon, 2003; Tsai& Hsu, 2005; van den Wollengerg, 1982; Yen, 1984)．しかし，これらの手法は IRT の項目パラメータまたは能力潜在変数の推定値を必要とする．これらは局所独立性の測定対象以外の項目に局所独立性が成り立つことを暗黙に仮定して得られるため，測定対象以外の項目間に局所従属な項目があると，局所独立性を正しく測定できないという問題がある．

これに対し，Hashimoto & Ueno (2011)の LCI (Latent Conditional Independence)指標は，IRT のパラメータ推定が不要で，測定対象以外の局所従属性に頑健であるという特徴を持つ．本報告では，LCI 指標を用いて項目間の局所独立性を測定する．

5.2.2 本報告で用いる局所独立性測定方法

本報告で局所独立性を測定するために用いる LCI 指標の計算方法は以下のとおりである．

テストがp個の項目から成り立っているとする．項目$i(i = 1, \ldots, p)$に対してX_iという確率変数を仮定し，項目iへの正答という事象を$X_i = 1$，誤答という事象を$X_i = 0$ で表す．項目iと項目i'を除くp-2 個の項目の集合に対して$\boldsymbol{X}^{\neg ii'}$という確率変数を仮定する．$\boldsymbol{X}^{\neg ii'}$の観測された値の数が J 通りであるとき，そのj番目$(j = 1,\ldots, J)$の値を$\boldsymbol{x}_j^{\neg ii'}$で表し，これを“項目$i$と項目$i'$を除くすべての項目に対する$j$番目の反応パターン”とよぶ．

$X_i = x_i\ (x_i = 0,1)$，$X_{i'} = x_{i'}\ \left(x_{i'} = 0,1\right)$，かつ $\boldsymbol{X}^{\neg ii'} = \boldsymbol{x}_j^{\neg ii'}$である受検者の人数を $N_{x_i x_{i'} j}$とする．同様に，$X_i = x_i$かつ $\boldsymbol{X}^{\neg ii'} = \boldsymbol{x}_j^{\neg ii'}$である受検者の人数を$N_{x_i \bullet j}$，$X_{i'} = x_{i'}$かつ$\boldsymbol{X}^{\neg ii'} = \boldsymbol{x}_j^{\neg ii'}$である受検者の人数を$N_{\bullet x_{i'} j}$，$\boldsymbol{X}^{\neg ii'} = \boldsymbol{x}_j^{\neg ii'}$である受検者の人数を$N_{\bullet\bullet j}$とする．また，受検者の総数を$N$とする．このとき，項目$i$と項目$i'$の LCI 指標 $I_{ii'}$は次の式で計算される．

$$I_{ii'} = \frac{1}{N}\sum_{j=1}^{J}\sum_{x_i=0}^{1}\sum_{x_{i'}=0}^{1} N_{x_i x_{i'} j} \log_2 \frac{N_{x_i x_{i'} j} N_{\bullet\bullet j}}{N_{x_i \bullet j} N_{\bullet x_{i'} j}} \tag{5.1}$$

$I_{ii'}$ は項目iと項目i'が局所独立ならば 0 となり，局所従属性が強まるに従って値が大きくなる．Hashimoto & Ueno (2011)は，LCI 指標の閾値を定め，LCI 指標の値が閾値以上である項目の対は局所従属とし，そうでない項目の対は局所独立であるとする“LCI 検定”を提案した．そこでは 0.01，0.05，0.10 という閾値が用いられているが，本報告では局所従属性を可能な限り多く検出できるよう，0.01 を閾値として用いる．

5.2.3 分析対象

本報告では，大学入試センター研究開発部が毎年 1 月に実施する，大学入試センター試験モニター調査の，平成 21～23 年度(荘島 2009; 荘島 2010; 荘島 2011)のデータを利用し，各年度の世界史 B 本試験の項目の局所独立性を調査した．この調査の受検者は東京都内 5 つの国立大学の 1 年生で，世界史 B 本試験の受検者数および項目数は表 5.1 のとおりである．

表 5.1 分析に用いたデータ

年度	受検者数	項目数
平成 21 年度	160	36
平成 22 年度	121	36
平成 23 年度	119	36

表 5.2 局所従属性の観測された項目対

平成 21 年度	
項目対	LCI 指標
第 1 問 問 5 － 第 4 問 問 6	0.0125
第 2 問 問 5 － 第 3 問 問 4	0.0125
第 2 問 問 5 － 第 4 問 問 1	0.0125
平成 23 年度	
項目対	LCI 指標
第 1 問 問 7 － 第 3 問 問 9	0.0168
第 3 問 問 5 － 第 4 問 問 9	0.0168

5.3 結果

LCI 指標の値が 0.01 を上回った項目対は表 5.2 のとおりで，平成 21 年度は 3 組，平成 23 年度は 2 組であった．平成 22 年度は LCI 指標の値が 0.01 を上回った項目対がなかった．この 3 年間の世界史 B の本試験は 36 項目から成り立つため, 項目の組み合わせは 630 通りとなる．したがって，本報告の分析の限りでは，世界史 B の試験では局所独立性の仮定がほぼ成り立っていると考えられる．

他の項目との間に局所従属性が観測された項目を含めた場合と，そのような項目を除外した場合で，IRT の 2 母数ロジスティック・モデルのパラメータを推定し，推定値を比較した結果は図 5.1〜5.4 のようになった. なお, パラメータ推定には EasyEstimation(熊谷, 2009) を利用した．この結果より，局所独立性の成り立っている項目は，テスト内に局所従属性のある項目が含まれるか否かにかかわらず，パラメータ推定値はほとんど影響を受けていないことがわかる．

5.4 考察

本報告では，LCI 指標を用いて項目の局所独立性を測定し，IRT 適用の可否を判断することを試みた．例として，センター試験の世界史 B に東京都の国立大学 1 年生が解答したデータに適用したところ，ほとんどの項目で局所独立性が成り立っているという結果が得られ，そのような項目は，局所独立性の成り立たない項目がテストに含まれていても，IRT のパラメータ推定値は影響を受けないことがわかった．

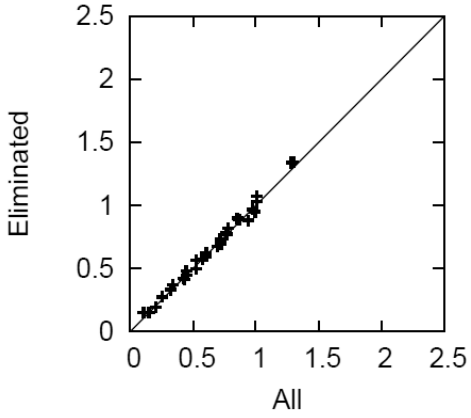

図 5.1 平成 21 年度の項目の識別力

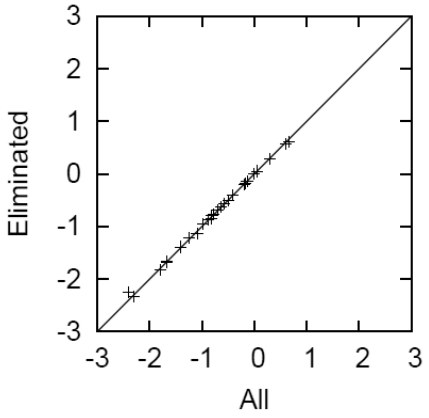

図 5.2 平成 21 年度の項目の困難度

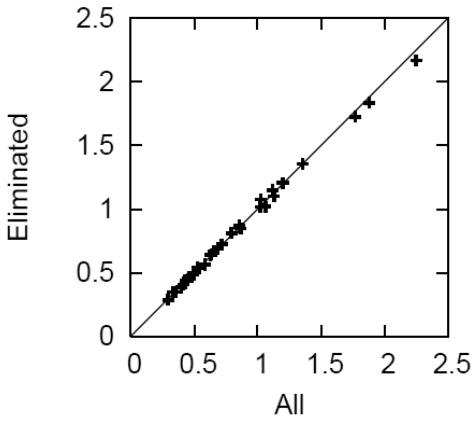

図 5.3 平成 23 年度の項目の識別力

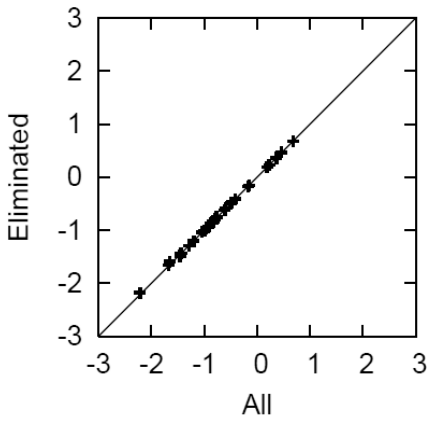

図 5.4 平成 23 年度の項目の困難度

ただし，本報告のデータでは局所従属性がほとんど観測されていないが，これは受検者数が十分でないことに起因する可能性がある．(5.1)式のとおり，LCI 指標は反応パターンごとの条件付き相互情報量の加重平均である．それぞれの反応パターンに対して，条件付き相互情報量を計算するのに十分多くの受検者がいない場合，その反応パターンの人数の分だけ LCI 指標の値は小さくなる．本報告のデータでは受検者数が 120～160 人であるため，より多くの受検者数で分析すれば，より多くの項目対で局所従属性が観測される可能性がある．

本報告で局所独立性の成り立っている項目の IRT パラメータ推定値は，他項目との間に局所従属性のある項目を含めた場合でも除いた場合でも，ほぼ同程度の値であった．このことから，他項目との間に局所従属性のある項目を含めて IRT のパラメータ推定を行っても，局所独立性の成り立っている項目のパラメータ推定値はその値を信用してよいと考えられる．ただし，これが常に成り立つかについては，実データの解析だけでなく，条件を統制した数値実験などを通じて確認することが必要である．

以上より，IRT 適用可能性の判断材料として，局所従属性のある項目およびその項目との間の LCI 指標を，各項目の統計情報として付加することが勧められる．しかし，項目バンクにこれらの情報を付加することは慎重にすべきである．なぜなら，このテストが多人数に対して行われたものならば，項目の公開による項目の品質劣化が避けられず，逆に非公開で少人数に対して行われたものならば前述のとおり LCI 検定の結果が疑問視されるからである．したがって，LCI 検定の結果は，局所独立な項目を集めた項目バンクの作成を目的とするのではなく，既に実施されたテストを評価するために用い，新規項目作成のための参考資料とするのが望ましいと考えられる．

引用文献

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores, Reading*. MA: Addison−Wesley. pp.397−424.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, **22**, 265−289.

Glas, C. A. W. & Suarez Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, **27**, 87−106.

Hashimoto, T. & Ueno, M. (2011). Latent conditional independence test using Bayesian network item response theory. *IEICE Transactions*, **E94-D**, 743−753.

橋本貴充・植野真臣(2011). 潜在変数周辺化による項目潜在構造分析　日本教育工学会論文誌, **35**, 205−215.

熊谷龍一(2009). 初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発　日本テスト学会誌, 5, 107−118.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. MA: Addison-Wesley.

村木英治(2011). 項目反応理論　朝倉書店

佐野真(2009). 相互情報量を用いた項目識別力の過大推定の検出　日本テスト学会誌, **5**, 3−21.

荘島宏二郎(研究代表者)(2009). 平成 21 年度大学入試センター試験モニター調査報告書. 大学入試センター研究開発部.

荘島宏二郎(研究代表者)(2010). 平成 22 年度大学入試センター試験モニター調査報告書. 大学入試センター研究開発部.

荘島宏二郎(研究代表者)(2011). 平成 23 年度大学入試センター試験モニター調査報告書. 大学入試センター研究開発部.

Tsai, T. H. and Hsu, Y. C. (2005). The use of information entropy as a local item dependence assessment. *Paper printed at the annual meeting of the American Educational Research*

Association, Montreal, Quebec, Canada.

van den Wollengerg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, **47**, 123−140.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, **8**, 125−145.

吉村宰(2005). 大学入試センター試験成績データからの採点及び統計量計算プログラムの開発と試験問題統計情報の整備　石塚智一(研究代表者)平成 14 年度～16 年度共同研究 I 報告書試験問題統計情報の整備に関する研究　大学入試センター研究開発部 pp.1−19.

第 6 章　入試における障害者支援と公平性・妥当性
—発達障害を中心に—

立脇洋介

6.1 はじめに

障害のある受験者にとって入試における支援は合格を左右する重要な問題である．高等学校の授業で受けている支援を入試で受けられなければ，本来持っているはずの力を発揮しにくくなる．その一方で，試験実施者や障害のない受験者の中には，「特別な支援を受けながら受験することは公平性や妥当性の点で問題がある」と考える人もいる．その結果，支援を受けること自体が合否に影響すると考え，必要な支援すらも申請しにくい受験者も出てくるであろう．このような問題は，各種障害の中でも，見えにくい障害と言われる発達障害で特に生じやすい．

日本では，平成 23 年から大学入試センター試験(以下センター試験と表記)において発達障害のある受験者も支援を受けられるようになった．しかし，各大学が実施するセンター試験以外の入試では，どのような支援が公平・妥当であるかということがほとんど議論されておらず，支援もあまり広がっていない．本章では，入試における障害者支援の現状を整理した後，公平かつ妥当な支援の在り方について考察する．

6.2 センター試験における障害者支援

国内の大学は，入試において「センター試験に準ずる」障害者支援を実施している(全国障害学生支援センター, 2007)．そこでまずは，国内の大学入試における障害者支援のスタンダードとみなせるセンター試験を取り上げる．

支援の対象は，「視覚障害」「聴覚障害」「肢体不自由」「病弱」「発達障害」「その他」である．このうちの発達障害は，平成 16 年に施行された発達障害者支援法で定義されている「自閉症」「アスペルガー症候群」「その他の広汎性発達障害」「学習障害」「注意欠陥多動性障害(ADHD)」を指す．これらの障害の主な特徴をまとめたものを 6.1 に示す．発達障害は中枢神経系の機能不全が原因として考えられている点は共通しているものの，障害の種類によって特徴が大きく異なっている．

表 6.1　発達障害の各障害の特徴

障害	特徴
自閉症	・他人との社会的関係の形成の困難さ ・言葉の発達の遅れ ・興味や関心が狭く特定のものにこだわる
学習障害 (LD)	・全般的な知的発達に遅れはない ・聞く、話す、読む、書く、計算する又は推論する能力のうち特定のものの習得と使用に著しい困難を示す。
注意欠陥／ 多動性障害 (ADHD)	・年齢あるいは発達に不釣り合いな注意力、衝動性、多動性 ・社会的な活動や学業の機能に支障をきたす

文部科学省 (1999, 2003)

次に具体的な支援の内容を説明する．表 6.2 では，各障害で受けることができる支援を，支援の種類に注目して整理した．「出題形式」とは，通常の問題冊子を読むことが困難な人への支援であり，点字や 1.4 倍に拡大した文字による出題がある．「解答形式」とは，マークシートへの記入が困難な人への支援であり，点字，文字，チェックでの解答や代筆者による記入がある．「時間延長」は，主に読みや書きが困難な人への支援であり，試験時間を 1.3 倍または 1.5 倍に延長する．「指示の伝達」とは，口頭だけでは説明が十分に伝わらない人を対象にした支援であり，文書や手話通訳士による説明が追加される．「部屋・座席」と「特別な機材の使用」は，明るい座席や照明器具の使用(視覚障害)，前列の座席や補聴器の使用(聴覚障害)など各障害の特性に応じて，必要な試験環境や機材を選択できる支援である．「リスニング」は聴覚障害の人のみが受けられる支援であり，スピーカーなどによる特殊な聴取方法や試験の免除が行われる．

表 6.2　センター試験における障害者支援

	視覚障害	聴覚障害	肢体不自由	病弱	発達障害
出題形式	点字／拡大文字	－	－	－	拡大文字
解答形式	点字／文字		チェック／代筆	－	チェック
時間延長	1.5倍／1.3倍	－	1.3倍	－	1.3倍
指示の伝達	－	手話通訳士／文書	文書	－	文書
部屋・座席	明るい座席	前列の座席	出入り口付近 トイレに近い部屋 1階の部屋	1階の部屋 別室	出入り口付近 トイレに近い部屋 1階の部屋 別室
特別な機材の使用	照明器具 拡大鏡	補聴器 （人工内耳）	特別な机・椅子 車椅子／杖	杖	－
リスニング	－	免除 スピーカー等の使用	－	－	－

独立行政法人大学入試センター (2011a) をもとに作成

発達障害の人が受けられる支援は，多岐にわたっている．これは，先述したように障害の種類や程度によって特徴が異なり，必要な支援も多様なためと考えられる．「拡大文字」

は，1.4 倍の大きさのゴシック体で印刷された問題冊子が配布される支援であり，学習障害で読みが困難な人に有効である．「チェック解答」とは，マークシートを塗る代わりにチェックで解答する支援であり，細かいマークシートを苦手とする人(例：書き困難がある学習障害の人やこだわりが強い自閉症の人)に有効である．「1.3 倍の試験時間延長」は，様々な原因で通常の時間では十分に解答できない人にとって有効である．「注意事項等の文章による伝達」では，口頭での説明に加え，説明が記載された文章が配布される．自閉症や ADHD のうち，聞き逃しの多い人にとって有効である．「別室受験」は，少人数の教室で試験を受けることができる．

続いてセンター試験における障害者支援の利用者数を概観する(大学入試センター, 2011b など)．過去 6 年間利用者数を，障害種別に集計した結果を図 6.1 にまとめた．最も新しい平成 24 年は，利用者が 1,472 人であった．これは全志願者 555,537 人の 0.26%にあたる．内訳をみると，視覚障害 58 人，聴覚障害 348 人，肢体不自由 189 人，病弱 73 人，発達障害 135 人，その他 669 人である．

6.3 国内大学の入試における障害者支援

次にセンター試験以外の国内大学の入試における障害者支援の現状についてまとめる．日本学生支援機構は毎年国内の全ての高等教育機関を対象に障害者支援の実態について調査を実施している(日本学生支援機構, 2011 など)．平成 23 年の各大学の入試で，支援を受けた受験者は 2,121 人であった．このうち合格した人は 783 人，実際に入学した人は 564 人である．入学した人の内訳は，視覚障害 68 人，聴覚障害 203 人，肢体不自由 145 人，病弱 30 人，重複障害 25 人，発達障害 14 人，その他 80 人である．障害の種類別に平成 19〜23 年の入学者数をまとめたものが図 6.2 である．センター試験の結果(図 6.1)と比較すると，以下の 3 つの特徴が見られる．第一に，どちらの試験も視覚障害，聴覚障害，肢体不自由，病弱の人数は，ここ数年ほとんど変化していない．第二に，聴覚障害の人が毎年最も多く，次いで肢体不自由の人が多い．聴覚障害の人は，口頭での説明に対する支援だけでなく，センター試験ではリスニングの免除も必要なため，人数が多いと考えられる．肢体不自由の人は，書字への支援や座席・教室の配慮などが必要となる．第三に，発達障害の人への支援は，各大学の入試に比べてセンター試験で多く行われている．センター試験での人数は，平成 23 年が 95 名，平成 24 年が 135 名である．一方，各大学の入試では平成 19 年から毎年 10 名前後で推移している．

平成 23 年の入学者の入試形態を見ると，学力入試が 52.8%に対し，特別入試(AO 入試 18.1%，推薦入試 27.1%，障害者特別選抜 2.0%)が 47.2%であった．文部科学省(2011)が実施した調査によれば, 国内の大学の全入学者では, 学力入試が 55.7%を占め, AO 入試が 8.7%, 推薦入試が 35.1%である．両者を比較すると，障害のある受験者は AO 入試で支援を受けて入学する割合が高い．面接や小論文など AO 入試で科される課題は，大学側が支援をしやすく，受験者も取り組みやすいと考えられる．

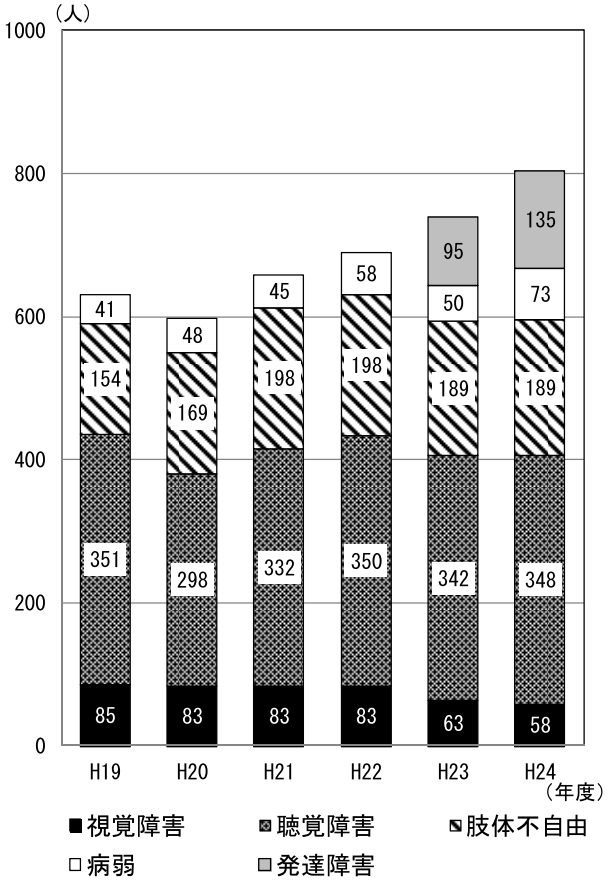

図 6.1　センター試験で支援を受けた
障害受験者数の推移

大学入試センター(2011b など) をもとに作成

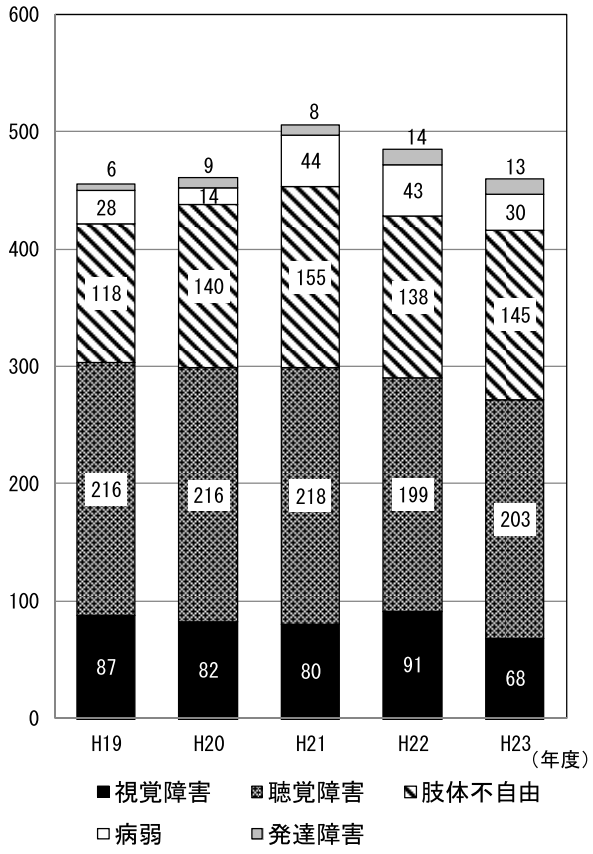

図 6.2　各大学の入試で支援を受けて
入学した障害学生数の推移

日本学生支援機構 (2011 など) をもとに作成

6.4 国外の大学入試における障害者支援

ここでは，障害者支援が進んでいるアメリカとイギリスの入試を取り上げる．日本のセンター試験と類似した試験として, アメリカでは SAT(Scholastic Assessment Test)と ACT (The American College Testing)という 2 種類の試験が実施されている．イギリスでは教育修了資格試験 GCE(General Certificate of Education)がある．これらの試験において実施されている支援のうち，センター試験で行われていないものを表 6.3 にまとめた．出題形式では，視覚障害や読字障害の人に向け，パソコンや人による「音声出題」が行われている．解答形式に関しては，「パソコン」のワープロ機能や音声入力機能が利用されている．試験時間に関しては，「2 倍以上」「希望する時間を申請する」などの支援も見られる．部屋・座席の「別会場」とは，試験会場に来ることが困難な人が病院や高等学校などで受験できる支援である．その他に，注意の持続が困難な人や受験の負担が多い人が，「試験中の休憩」や「複数日に渡る受験」などを認められる支援もある．

支援内容以外で異なる点としては，以下の 2 つが挙げられる．第一に，利用する人数で

ある．2000 年に SAT を受けた人のうち支援を利用した障害受験者は 2.0%を占めており，国内の 10 倍ほどである．第二に，支援に段階が設定されている．例えば SAT や ACT では 1.5 倍の時間延長や拡大文字などの支援は全ての試験会場で受けることができる．しかし，1.5 倍以上の時間延長，複数日受験，音声出題など，通常の会場で対応することが困難な支援については特別な会場のみで受けることができる．

表 6.3　欧米の大学入試における障害者支援

	SAT	ACT	GCE
出題形式	音声出題 （人／PC／テープ） 色つき用紙 行間や文字間の拡大	音声出題 （人／PC／テープ）	音声出題 （人／PC） 色つき用紙
解答形式	PC／拡大用紙	PC	PC
時間延長	2倍／2倍以上	1.5倍以上	2倍／2倍以上
部屋・座席	別会場	別会場	別会場
特別な機材の使用	PC	PC	PC／英語の辞典
その他	休憩／複数日受験	休憩／複数日受験	休憩

ACT (2011)，College Board (2011)，JCQ (2011)をもとに作成

6.5 試験・評価の方法と障害者支援

以上のように，国内でも大学入試における障害者支援が実施されているが，現状では支援方法や利用者数の点でアメリカやイギリスに遠く及ばない．特に，発達障害の受験者に対する支援は，開始したばかりであり，今後改善していくことが望まれる．ただし，両国と日本では「試験・評価の方法」などが大きく異なる．そのため，国外の支援をそのまま導入した場合，公平性や妥当性の点で問題が生じる可能性がある．そこで本章のまとめとして，いくつかの試験・評価の方法を取り上げ，それぞれの方法における発達障害者支援のあり方を考察する．

6.5.1 試験の方法

教科，難易度，出題・解答形式など試験の方法によって，必要な支援は異なる．例えば欧米では，読み障害の受験者がコンピュータの読み上げソフトなどを使用できる．しかし国内の入試では，国語の漢字や英語の発音など，読み自体の能力を問う問題もあり，公平性を保つためには「該当部分を読み飛ばす」「それらの問題を採点から除外する」などが必要になる．また文字を正確に書くことが採点の対象になっている試験では，スペルチェックや文字変換の機能を切ることでワープロ等の使用も可能になるであろう．

難易度に関して，Mandinach et al. (2005)の実験では，成績が平均以下の人では時間延長をしても得点が変化しないが，平均以上の人では時間延長によって得点が高くなっていた．また立脇(2012)がセンター試験を用いて実施した実験では，国語・英語・公民の得点では時

間延長の効果が見られなかった．しかし数学では，成績のよい集団で時間延長をした際，得点が高くなった．これらの結果より，難易度が易～中程度で，制限時間が得点に大きく関わる試験(スピードテスト)では，時間延長の扱いが難しいといえる．

出題形式に関して，日本の入試では，ある文章を読んで，それに関するいくつかの問題に答える「大問形式」が中心である．大問形式の試験では，説明文などがあるため，読む分量が多く，読みに困難がある人の負担が大きい．マークシートなどの「選択式」で解答する形式も，解答の選択肢を読む必要があるため，読みの負担が大きい試験である．一方，「論述式」で解答する形式の試験では，書きの負担が大きくなりやすい．さらに，面接試験の場合，質問も解答も口頭でなされるため，コミュニケーションや口頭指示の理解に困難を抱える人には文章による伝達などの支援が必要となる．試験の方法ごとの障害者支援の例をまとめたものを表 6.4 に示す．「大問形式」「選択式」であるセンター試験は，読みに関する負担が大きい試験と考えられる．

表 6.4　試験の方法ごとの障害者支援の例

試験の方法	負担の大きい人	支援の方法
面接試験	コミュニケーションや口頭指示の理解に困難を抱える人。	文章等による指示の伝達。
筆記試験		
論述式	書きに困難のある人。	ワープロの使用。時間延長。
選択式	読みに困難のある人。	拡大文字。音声出題。時間延長。
大問形式	読みに困難のある人。	拡大文字。音声出題。時間延長。
長時間の試験	注意集中に困難のある人。	休憩。別室受験。
スピードテスト	様々な理由で時間が通常以上にかかる人。	時間延長。 ※得点に大きな影響を及ぼすため、公平性を考慮する必要がある。

6.5.2 評価及び入学者決定の方法

試験の結果をどのように評価して入学者を選択するかという点について，日本と欧米では方法が大きく異なっている．日本の入試では，大学全入化時代と言われる現代でさえ，受験者を成績順に並べて上から合格としていく相対評価が中心であり，全ての受験者が同一の条件で試験を受けることが期待される．そのため障害者支援は，障害のない受験者の順位を相対的に低下させることにもつながってしまう．一方アメリカやイギリスでは，入学者を選択するにあたり，試験の結果は主に基準点に到達しているかを判断するために使用され，小論文や高校の成績などとともに，入学者を選択する資料の一つとして扱われる．また，試験が一年間に複数回実施されるため，同じ科目を何度も受験することができる．受験者によって異なる試験を受けているため，数点の得点差を比較することに向いていない．障害受験者が支援を受けたとしても，他の受験者への影響は相対評価に比べると少な

い．国内でも資格試験のように到達度を評価する目的で実施される試験では，公平性の問題が生じにくいと考えられる．

6.5.3 まとめと今後の課題

入試における障害者支援がさらに進展するために，各大学はまずアドミッション・ポリシーをもとに入試で測定したい能力を明確にし，その能力に適した試験や評価方法を選択することが必要である．本章で示したように，試験方法によって，支援の必要な人は異なる．それらの支援が，測定したい能力を直接底上げするものでない限り，支援をしていくことが望まれる．

引用文献

ACT (2011). Services for Students with Disabilities. http://www.act.org/aap/disab/index.html

College Board (2011). Services for Students with Disabilities. http://www.collegeboard.com/ssd/student/index.html

大学入試センター (2011a). 受験案内別冊. 大学入試センター.

大学入試センター (2011b). 過去のセンター試験データ. http://www.dnc.ac.jp/modules/center_exam/content0092.html

JCQ (2011). Access Arrangements, Reasonable Adjustments and Special Consideration. JCQ.

Mandinach, E.B., Bridgeman, B., Cahalan-Laitusis, C., & Trapani, C. (2005). The impact of extended time on SAT test performance. College Board Research Report No. 2005-8.

文部科学省 (1999). 学習障害児に対する指導について(報告). http://www.mext.go.jp/a_menu/shotou/tokubetu/material/002.htm

文部科学省 (2003). 今後の特別支援教育の在り方について(最終報告). http://www.mext.go.jp/b_menu/shingi/chousa/shotou/018/toushin/030301.htm

文部科学省 (2011). 平成 23 年度国公私立大学・短期大学入学者選抜実施状況の概要. http://www.mext.go.jp/b_menu/houdou/22/10/1297952_1532.html

日本学生支援機構 (2011). 平成 23 度(2011 年度)大学，短期大学及び高等専門学校における障害のある学生の修学支援に関する実態調査結果報告書. 日本学生支援機構.

立脇洋介 (2012). 時間延長がテストの得点に及ぼす影響. 大学入試研究ジャーナル, **22**, 193-198.

全国障害学生支援センター (2007). 大学案内 2008 障害者版. 全国障害学生支援センター.

第 7 章　因子数が明らかでない場合の信頼性のベイズ推定

岡田謙介

要旨

テストや尺度の分析において，信頼性の推定には通常因子分析モデルが利用される．一般に真の因子構造や因子数は分析者には未知であり，したがって信頼性の推定においてもモデルの不確実性があるが，これは通常の分析では無視されてしまっている．そこで本研究では，因子分析モデル自体の不確実性を考慮した，ベイズモデル平均化による信頼性の推定法を提案する．モンテカルロ実験により，既存の方法と対比した本手法のパフォーマンスを示す.

7.1 導入

信頼性(reliability)の推定は入試やテストの標準化や精密な評価のために不可欠である．わが国において 1979 年に導入された大学入試の共通一次学力試験も，その大きな目的のひとつに「妥当性・信頼性の高い入試の実現」があった(大学入試センター, 1979)．荒井 (2003) はこれを指して共通一次試験の「最大の眼目」と述べている．大学入学者選抜のシステムにおける試験の信頼性を保証することは，今後の大学入試選抜のあり方を考える上でも避けて通れない大きな課題である(国立大学協会, 1998).

しかし，その重要性の一方で，統計学的な観点からは信頼性の推定は決して単純な問題ではない．概念的には，信頼性は観測値の分散に占める真値の分散を表す量である．古典的テスト理論(classical test theory)にしたがいテストの観測値Yが真の潜在特性値Tと誤差Eの和として観測される，すなわち

$$Y = T + E \tag{7.1}$$

であるとする．このとき信頼性は

$$\rho = \frac{var(T)}{var(Y)} \tag{7.2}$$

によって与えられる．ただし$var(\cdot)$は分散を表す．たとえば一般的な知能検査の信頼性の推定値は 0.80〜0.90 と高い水準を満たすことが求められている(Eysenck, 2000).

古典的テスト理論の考え方に基づく信頼性の推定値としてもっとも広く使われているのは Cronbach のαであろう．この指標を提案した Cronbach (1951)の論文の引用回数は 6,500 を超え，20 世紀最大の科学的発見とも呼ばれる二重らせんの発見の論文の引用回数を上回ることが報告されている(Sijtsma, 2009)．しかし，Cronbach のαは元来信頼性係数の下界として導出された指標であり，一般の条件下では多くの場合に信頼性の真値を過小評価すること

が多くの研究で報告されている．また，そのためもありαに代わる信頼性係数に関して数多くの研究がとくに近年なされている(岡田, 2011).

中でも有望と考えられているアプローチのひとつに，構造方程式モデリング(structural equation modeling, SEM)に基づく信頼性の推定がある．これは SEM，とくに通常は因子分析モデルの最尤推定を用いて観測変数の分散を真値(因子)の分散と誤差分散とに分離し，前者が全分散に占める割合を信頼性の点推定値とする考え方である．多くの研究者によって SEM による信頼性の推定が有望視されている(Bentler, 2009; Bollen, 1989; Raykov & Shrout, 2002)．Yang and Green (2010)は様々な条件下で SEM による信頼性推定のバイアスや精度を調べるモンテカルロ実験を行った．彼らの結果によれば，モデルが正しく設定されている場合には，SEM による推定は相対的にバイアスの小さな望ましい推定ができていた．しかし一方で，モデルが誤設定された(misspecified)場合には，SEM による推定には一般に大きなバイアスが生じていた.

そこで本研究では，モデルが誤設定された場合のバイアスという SEM による信頼性推定の問題について，ベイズモデル平均化(Bayesian model averaging, BMA; Hoeting, Madigan, Raftery, & Volinsky, 1999)によって頑健な推定を行う方法を提案する．7.2 節ではこの方法論について述べ，7.3 節では既存の方法と比較した数値実験の結果を示す．7.4 節でまとめと今後の展望について議論を行う.

7.2 信頼性のベイズ推定

7.2.1 モデル

本稿では真の因子数がわからない場合に，複数の因子数のモデルをベイズモデル平均化によって統合することを考える．SEM の下位モデルである因子分析モデルは,

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{7.3}$$

$$\boldsymbol{\eta}_i \sim N_k(\mathbf{0}, \mathbf{I}_k) \tag{7.4}$$

$$\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{7.5}$$

と表現することができる．ここで$\boldsymbol{y}_i$は$(p \times 1)$の個体i についての$(i = 1, \dots, n)$の測定値ベクトル，$\boldsymbol{\Lambda}$は$(p \times k)$の因子負荷量行列，$\boldsymbol{\varepsilon}_i$は$(p \times 1)$の誤差ベクトル，$\boldsymbol{\eta}_i$は$(k \times 1)$の標準正規分布にしたがう個体iの因子得点ベクトルである$(k \ll p)$．ここで$\boldsymbol{y}_i$の周辺分布は

$$\boldsymbol{y}_i \sim N_p(\mathbf{0}, \boldsymbol{\Omega}) \tag{7.6}$$

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma} \tag{7.7}$$

となる．このモデルのもっとも標準的といえるベイズ推定は，$\boldsymbol{\Lambda}$に正規事前分布・誤差分散に逆ガンマ分布という自然共役事前分布を設定し，ギブスサンプラーを用いた母数の事後分布の推定を行うものである(Arminger, 1998; Song & Lee, 2001)．しかし，この設定の元では超母数の選択がしばしば困難であり，またその選択によってはギブスサンプラーのミキシ

ング効率が悪くなってしまうことが知られている．そこで本研究では，Ghosh & Dunson (2009)のデフォルト事前分布を用いたモデルを利用する．このモデルでは，以下の関係式で表される母数変換を利用して，通常の因子分析モデルの母数$\boldsymbol{\eta}_i, \boldsymbol{\Lambda}$を母数拡大(parameter expanded; PX)因子分析モデルの母数$\boldsymbol{\eta}_i^*, \boldsymbol{\Lambda}^*$へと変換する：

$$\lambda_{jl} = sign(\lambda_{ll}^*)\lambda_{jl}^*\psi_l^{1/2} \tag{7.8}$$

$$\eta_{il} = \text{sign}(\lambda_{ll}^*)\psi_l^{-1/2}\eta_{il}^*, \tag{7.9}$$

ただし$\text{sign}(x) = -1\ (x < 0);\ \text{sign}(x) = 1\ (x \geq 0)$である．このとき，母数変換後の PX 因子分析モデルは

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i \tag{7.10}$$

$$\boldsymbol{\eta}_i \sim N_k(\mathbf{0}, \boldsymbol{\Psi}) \tag{7.11}$$

$$\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0},\ \boldsymbol{\Sigma}) \tag{7.12}$$

となり，$\boldsymbol{\eta}_i$の分散共分散行列が対角行列とならない．つまり，$\boldsymbol{\Lambda}^*$と $\boldsymbol{\Psi}$の対角成分が冗長化されていることになる．$\boldsymbol{\eta}_i^*$ を周辺化すると，$\mathbf{y}_i \sim N_p(\mathbf{0},\ \boldsymbol{\Lambda}^*\boldsymbol{\Psi}\boldsymbol{\Lambda}^{*'} + \boldsymbol{\Sigma})$を得ることができる．

7.2.2 事前分布

ベイズ推定では未知母数に事前分布を設定する．各母数について，次の事前分布を利用する．

$$\lambda_{jl}^* \sim N(0,1), \quad j = 1, \dots, p,\ l = 1, \dots, min(j,k) \tag{7.13}$$

$$\lambda_j^* = 0, \qquad j = 1, \dots, (k-1), l = j+1, \dots, k \tag{7.14}$$

$$\psi_l^{-1} \sim Gamma(a_l, b_l), \quad l = 1, .., k, \tag{7.15}$$

$$\sigma_j^{-2} \sim Gamma\big(c_j,\ d_j\big), \quad j = 1, .., p \tag{7.16}$$

7.2.3 パスサンプリング

ここまで因子数kを固定して話を進めてきたが，本研究では因子数kが未知の場合を考えている．そこで最大の因子数をmとし，因子数kについて多項事前分布

$$Pr(k = h) = \kappa_h,\ \kappa_h = \frac{1}{m}\ ,\ \text{h} = 1, \dots, m \tag{7.17}$$

を設定する．因子数がhのとき，7.2.1 節および 7.2.2 節で$k = h$としたモデルが当該因子数の因子分析モデルとなる．このとき，因子数hのモデルの事後モデル確率は

$$\Pr(k = h|\boldsymbol{y}) = \frac{\text{O}[h:j] \times \text{BF}[h:j]}{\sum_{l=1}^{m} \text{O}[l:j] \times \text{BF}[l:j]}, \tag{7.18}$$

となる．ここで$\text{O}[h:j] = \kappa_h/\kappa_j$は事前オッズであり，これは(7.17)式の設定より定数となる．

一方$\mathrm{BF}[h:j]=\frac{\pi(\boldsymbol{y}|k=h)}{\pi(\boldsymbol{y}|k=j)\}}$はベイズファクターであり，これを数値的に推定するためパスサンプリングを利用する(Gelman & Meng, 1998; Lee & Song, 2002)．パスサンプリングでは，M_0, M_1を，それぞれ因子数 $h-1$, h のモデルとしたとき，この 2 つのモデルがパス

$$\mathrm{M}_t:\mathbf{y}_\mathrm{i}+\boldsymbol{\Lambda}_t\boldsymbol{\eta}_i+\boldsymbol{\varepsilon} \tag{7.19}$$

で結ばれると考える．ただし，ここでの$\boldsymbol{\Lambda}_t$は

$$\boldsymbol{\Lambda}_t=(\lambda_1,\lambda_2,\ldots,\lambda_{h-1},t\lambda_h) \tag{7.20}$$

という行列であり，$t=0$ のときがM_0に，$t=1$のときがM_1にそれぞれ対応する．各$t_0=0<t_1<\cdots<t_S<t_{S+1}=1$についてそれぞれでの MCMC による推定を実行し，

$$log\big(\widehat{BF}[h:(h-1)]\big)=\frac{1}{2}\sum_{s=0}^{S}(t_{s+1}-t_s)(\bar{U}_{s+1}+\bar{U}_s) \tag{7.21}$$

により因子数$\mathrm{h}-1$, hのモデルを比較するベイズファクターの推定値を数値的に得ることができる．ただし

$$U(\boldsymbol{\Lambda},\boldsymbol{\Sigma},\boldsymbol{\eta},\boldsymbol{y},t)=\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\Lambda}_t\boldsymbol{\eta}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{0}^{p\times(\mathrm{h}-1)},\boldsymbol{\lambda}_\mathrm{h})\,\boldsymbol{\eta}_i \tag{7.22}$$

であり，$\bar{\mathrm{U}}_\mathrm{s}$ は $\pi(\boldsymbol{\Lambda},\boldsymbol{\Sigma},\boldsymbol{\eta}|\boldsymbol{y},\boldsymbol{t}_{(s)})$からの$J$個の MCMC 標本 $\mathrm{U}\big(\boldsymbol{\Lambda}^{(j)},\boldsymbol{\Sigma}^{(j)},\boldsymbol{\eta}^{(j)},\boldsymbol{y},\boldsymbol{t}_{(s)}\big),\ j=1,\ldots,J$ に関する標本平均である．

7.3 数値実験

7.3.1 方法

提案手法のパフォーマンスを調べるため，数値実験を行った．真のモデルとして，1 因子モデル・2 因子モデル・3 因子モデルの 3 種類を用意した．それぞれにおける因子負荷量の真値は次の通りである．

1 因子モデル:

$$\boldsymbol{\Lambda}=(.8,.7,.6,.5,.4,.3)' \tag{7.23}$$

2 因子モデル:

$$\boldsymbol{\Lambda}=\begin{pmatrix}.8 & .7 & .6 & 0 & 0 & 0\\ 0 & 0 & 0 & .5 & .4 & .3\end{pmatrix}' \tag{7.24}$$

3 因子モデル:

$$\boldsymbol{\Lambda}=\begin{pmatrix}.8 & .7 & 0 & 0 & 0 & 0\\ 0 & 0 & .6 & .5 & 0 & 0\\ 0 & 0 & 0 & 0 & .4 & .3\end{pmatrix}' \tag{7.25}$$

各 3 通りの真のモデルにつき, 100 セットの乱数データセットを真のモデルから生成した．標本サイズは$N=300$とした．そして，提案手法と SEM による信頼性推定，および Cronbach

のαの値を比較した. 提案手法は Matlab を用いて実装し, burn-in を 1,000 回とった後の 6,000 回の MCMC 標本を推定に利用した．また，提案手法と比較するため，各 3 通りのモデルをそれぞれ真のモデル似設定した SEM による推定法 3 通り，McDonald のω_t, ω_h (McDonald, 1978, 1999), Cronbach のαの値も各データセットについてすべて計算した．SEM による推定には R 上で動作する OpenMx (Boker et al., 2011)を利用し, 応用上最もよく利用される最尤推定を行った.

提案手法を含めた上記 7 通りの手法による推定値と真値との相違を次のバイアス(bias)と精度(precision)

$$bias = \frac{(\bar{\hat{\rho}} - \rho)}{\rho} \tag{7.26}$$

$$precision = \sqrt{\frac{\sum_{i=1}^{n_{iter}} (\hat{\rho}_i - \rho)^2}{n_{iter} - 1}} \tag{7.27}$$

により評価し，各 100 セット分の平均値を算出した.

7.3.2 結果

バイアスと精度の結果について述べる前に，まず推定の収束について述べる．提案手法および Cronbach のαについては算出に問題のあったケースは見られなかったが，SEM による推定ではいくつかの標本でアルゴリズムが適切に収束しないケースが見られた．各条件における 100 回の繰り返しのうち適切に収束した割合を図 7.1 に示す．この図からわかるように，真のモデルよりも小さなモデルを推定に利用した場合に収束に問題のあるケースが見られた．また，1 因子モデルによる推定は 2 因子モデルによる推定よりも幾分頑健な傾向があった.

続いて各手法によるバイアスと精度についての結果を示すが，SEM による推定で上述の適切に収束しなかったケースの推定値を用いると大幅にバイアスや精度の値が悪化した．したがって，今回の結果では適切に収束しなかった標本は分析から除外した．この操作は Yang & Green (2010)などでも行われているものである.

各手法によるバイアスを図 7.2 に示す. バイアスは 0 に近いほど平均的に真値を正しく推定していることを表す．また，本来信頼性の低いテストを誤って信頼性が高いと評価してしまうことの方が，その逆すなわち高い信頼性のテスト誤って信頼性が低いと評価してしまうことよりも社会的な悪影響が大きいと考えられる．そのため，信頼性の真値を過大評価していることを表す正のバイアスよりは，過小評価していることを表す負のバイアスの方が望ましいとされる．図 7.2 からは，提案手法が安定的に真値を下側から推定しており，そのバイアスは各条件において真のモデルを設定した場合の SEM を除いてはもっとも小さいことがわかる．SEM で真のモデルを設定した場合のバイアスは各因子数ごとに見ると最

小であるが，その真のモデルが誤設定された場合のバイアスは非常に大きくなっている．それに対して，提案手法は真のモデルによらず安定して提案手法を下側から推定できている．なお，McDonald のω_tは上側から，ω_hは下側から真値を推定しており，そのバイアスはいずれも提案手法よりも大きい．またαも真値を下側から推定しているがやはりそのバイアスは提案手法よりも大きい．

次に，各手法の精度を図 7.3 に示す．精度は 0 以上の値をとり，0 に近いほど真値まわりでのその推定値のばらつきが小さい，よい推定量であることを表す．バイアスの場合と同じく，正しいモデルを設定した SEM が各条件ごとに見るともっとも精度の意味で望ましい推定値を与えていることがわかるが，モデルが誤設定された場合にはその精度は非常に悪化してしまうこともわかる．それに対して提案手法は真のモデルによらず，安定的に優れた精度を示しており，比較した各手法の中で最良と評価できる．

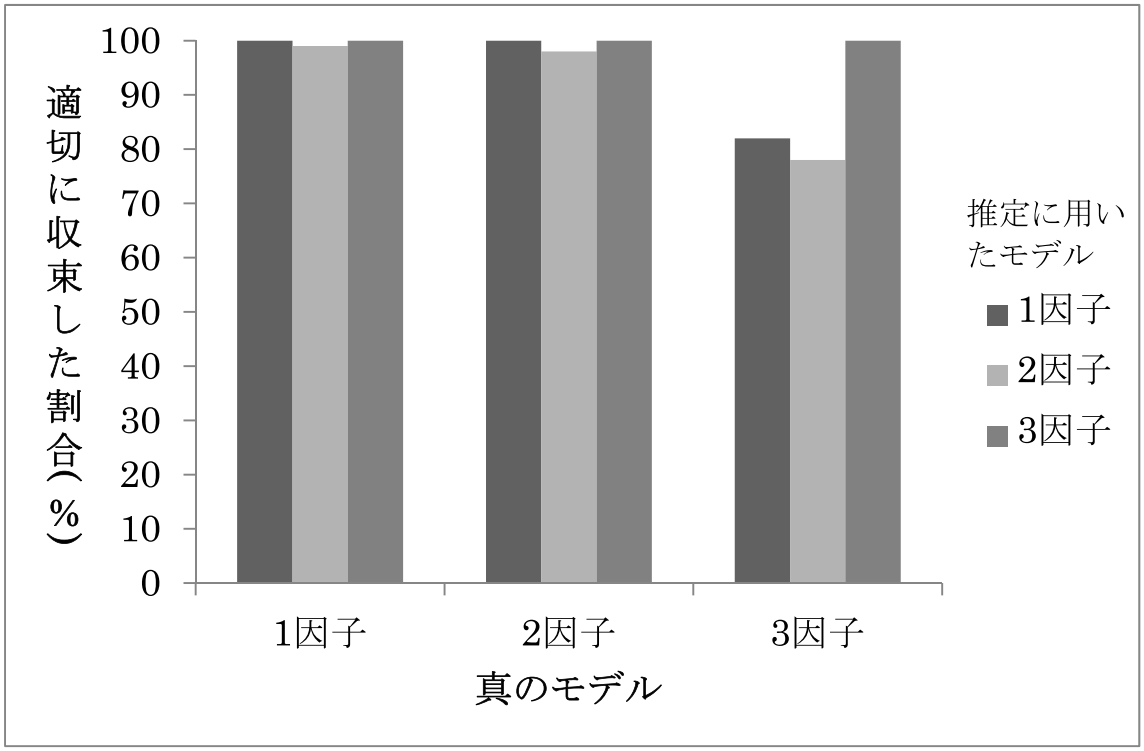

図 7.1 SEM による最尤推定が適切に収束した割合

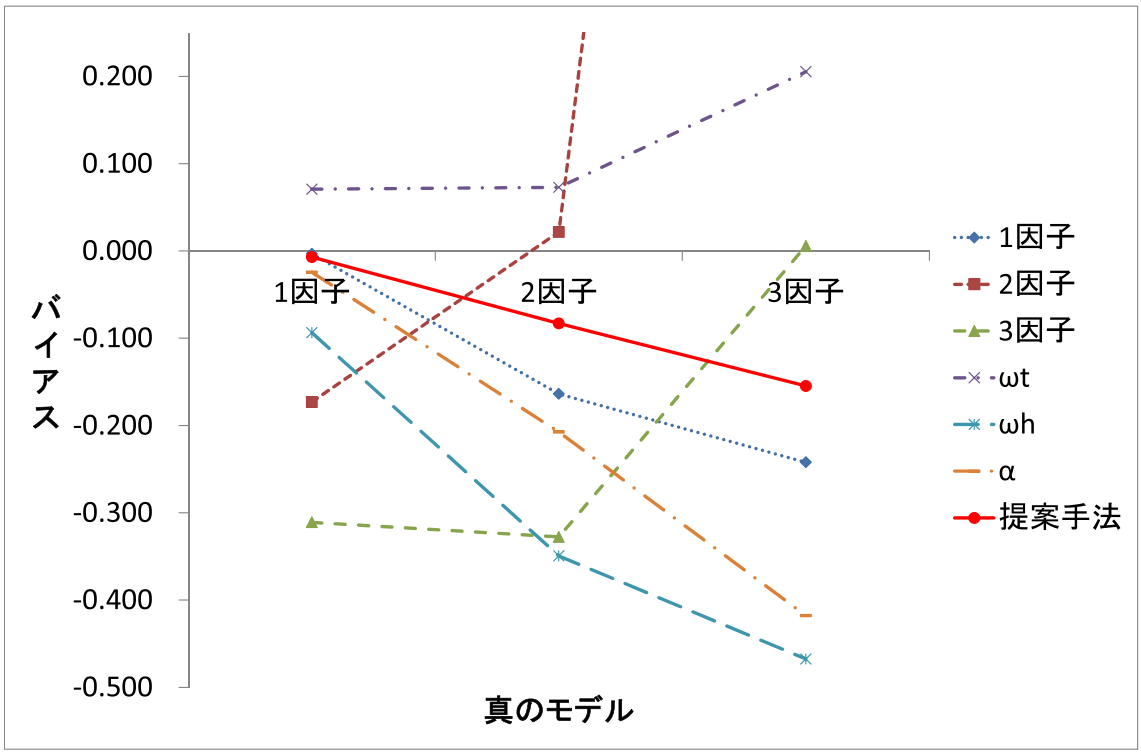

図 7.2 各種信頼性の推定法のバイアス

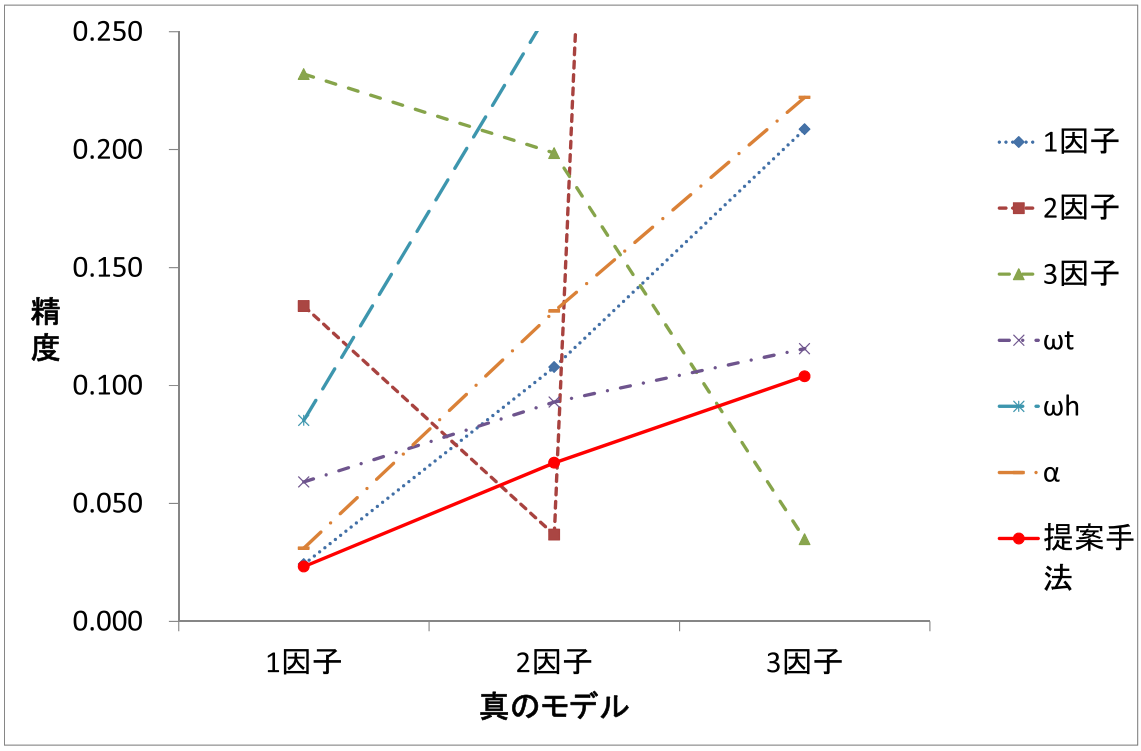

図 7.3 各種信頼性の推定法の精度

7.4 まとめと展望

本研究では，ベイズモデル平均化(BMA)に基づく新しい信頼性の推定法を提案した．数値実験からは，真のモデルによらず提案手法が頑健に真値を推定することが示された．一方，これまでに推奨されることの多かった SEM に基づく方法はモデルが正しく設定されていれば適切な推定ができるものの，モデルが誤設定された場合には非常に不適切な推定値を与えた．また Mcdonald のω_t, ω_hや Cronbach のαのいずれと比較しても提案手法のパフォーマンスは総合的に見て優れていた．

今後の研究の方向性について述べる．本研究では標本サイズは 1 条件しか設定していないため，他の条件，とくに標本サイズがより小さな条件での各手法の挙動を確認しておくことは有用であろう．また，現実のテストでは誤差相関のあるモデルがしばしば当てはまることも報告されており(Yang & Green, 2010), 真のモデルがそのような場合の各手法の挙動を調べることも重要と考えられる．

引用文献

荒井克弘. (2003). 学力評価システムの日米比較. 教育社会学研究, **72**, 37-52.

Arminger, G. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika,* **63**, 271-300. doi: 10.1007/bf02294856

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika,* **74**, 137-143.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., & Fox, J. (2011). OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika,* **76**, 306-317. doi: 10.1007/s11336-010-9200-6

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* **16**, 297-334.

大学入試センター (1979). 新しい大学入試 昭和 55 年度版　大学入試センター.

Eysenck, M. (2000). *Psychology: A Student's Handbook*. London: Psychology Press.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science,* **13**, 163-185.

Ghosh, J., & Dunson, D. B. (2009). Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis. *Journal of Computational and Graphical Statistics,* **18**, 306-320. doi: 10.1198/jcgs.2009.07145

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science,* **14**, 382-401.

国立大学協会. (1998). 大学入学者選抜の改善に向けて　国立大学協会.

Lee, S. Y., & Song, X. Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika,* **29**, 23-40.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and gereralizability. *Educational and Psychological Measurement,* **38**, 75-79. doi: 10.1177/001316447803800111

McDonald, R. P. (1999). *Test theory: a unified treatment.* . Mahwah, NJ: Lawrence Erlbaum Associates.

岡田謙介. (2011). クロンバックのαに代わる信頼性の推定法について. 日本テスト学会誌, **7**, 37-50.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: point and interval estimation using structural equation modeling approach. *Structural Equation Modeling,* **9**, 195-212.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,* **74**, 107-120.

Song, X. Y., & Lee, S. Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical & Statistical Psychology,* **54**, 237-263.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling,* **17**, 66-81.

独立行政法人大学入試センター　入学者選抜研究機構入試評価部門報告書
「大学入試の標準化、多様化、および精密化」

発行　　平成 24 年 10 月 15 日

編集・発行　独立行政法人大学入試センター入学者選抜研究機構
〒153-8501　東京都目黒区駒場 2-19-23
電話：03-3468-3311（代）

印刷　　株式会社　コームラ
