

験者でも、属する下位集団が異なると正答率が異なる、という現象が特定の項目で生じているときに、「特異項目機能が生じている」と言う。

項目応答(反応)理論の文脈では、テスト項目の項目特性曲線が受験者の下位集団間で異なる、すなわち、項目パラメタが同一項目であるにもかかわらず、下位集団が異なると異なる値を示すことになる。

高校入学時の英語学力の年次推移

一項目応答理論を用いた県規模英語学力テストの等化— 齊田 智里

1 茨城県高等学校英語学力テスト

茨城県高等学校教育研究会は、県内の高等学校に勤務する教員のほぼ全員が加入している研究団体である。このような高校教員の研究組織(高教研)は全47都道府県に存在しており、多くの高教研では数学と英語の学力テストを県規模で長年実施している。「茨城県高等学校英語学力テスト」は2種類作成され、毎年4月に受験希望校に実施される。すなわち、テストA(高校1年生対象)とテストB(高校2, 3年生対象)の2種類で、テストAは受験者数が多く、幅広い学力層が対象となり、また、長年同じ形式で実施されており、テストBは受験者数がテストAの3分の1程度で、やや高い学力層を対象としている。

こうした英語学力テストの特徴としては、1)年に一度、同一問題で、同じ時期に、学校単位で実施、2)多くの県で昭和40年代から継続実施、3)試験問題は公開、毎年問題は作り変えられる、4)分析は平均と正答率が中心、5)古典的テスト理論の枠組みでのテスト分析・開発、6)集団基準準拠テストの性格が強い、などがあげられる。いわば、典型的な日本の学力テストといえる。

2 項目応答理論による分析

項目応答理論を用いた分析を新たに実施することによって、1)項目と能力の2つの要因を分離してテスト分析が行える、2)複数のテスト結果を共通尺度上で比較可能にすること(等化)により、項目プールが作成できる、能力値の変化を明らかにできる、などの効果が得られる。今回は特に「学力の変化」という面に焦点をあてた。

2.1 尺度の等化の必要性

英語学力の年次推移を検討するためには、問題項目の異なる各年度のテストによる測定結果が同一の原点と単位をもつ共通尺度上で表示されなければならない。過年度に実施されたテストを年度ごとにIRT分析した場合、通常は年度ごとに能力値の平均を0、標準偏差を1として推定値が求められるため、そのままでは、異なる年度の結果を相互に比較することはできない。そのため、年

度間で共通の原点と単位をもつ共通尺度に各年度の特性尺度を等化する必要がある。

2.2 等化のデザイン

実際に複数の尺度を等化するためには、異なるテスト間に共通項目がある「共通項目デザイン」と、異なるテストを受験した共通の受験者がいる「共通受験者デザイン」のいずれかが用いられる。

過年度に実施された学力テストの場合、テスト間に共通項目もなければ、同一の受験者に同時に実施されたわけでもない。仮に、1年後に同じ受験者が再び受験をしたとしても、能力が変化しているため、尺度等化の情報にはならない。そこで、別途に既存のテスト項目を用いて「等化用テスト」を作成し、実施する。例えば、2つのテストから項目を半々取り出して、共通受験者デザインによる等化用テストを作成する。2つのテストの項目パラメタ値は事前に推定されているので、一人の等化用テスト受験者に対して、2つの能力パラメタ値が得られる。これらの値から等化係数を推定して、2つのテストを等化する。

2.3 茨城県英語学力テストの等化

1995年度から2002年度までの8年間にわたる延べ14万人の茨城県英語学力テストAのデータがある。問題構成やテスト形式は毎年ほぼ共通している。因子分析の結果、少数の項目を除けば各年度のテストについてIRTモデル適用の前提である一次元性の仮定が満たされることが確認された。2パラメタ・ロジスティック・モデルを適用して各年度の項目パラメタ値を推定した。しかしテスト問題は年ごとに違い、受験者も年ごとに異なっているため、これらを相互に比較することはできない。そこで、8年間のほぼ中間にあたり、テストの一因子性も強かった1999年度のテストを等化の基準となるテストとし、1999年度と1995年度、1999年度と1996年度というように、1999年度ともう一つ別の年度のテスト問題からそれぞれ項目を選び、共通被験者計画による等化用テストを全部で7種類作成した。

等化用テスト項目選択にあたっては、1)どの大問からも満遍なく項目を選択、2)極端な項目困難度パラメタ値(-1.5以下, 1.5以上)は排除、3)困難度パラメタ推定値にできるだけばらつきがでるような項目を選択、4)高い識別パラメタ推定値の項目を選択、5)テスト構成・項目数・実施時間は、もとのテストと同じにする、という基準を適用した。

等化用テストは、県下の協力校18校の高校1年生3,215名に実施された。一人の受験者に対して得られた2つの能力パラメタ推定値の情報をもとに、等化係数を求めて、もとの約14万人分の受験者パラメタ値を共通尺度

上で表わした。その結果、高校入学時の高校生の英語学力平均値が年々低下していること（8年間で偏差値換算で約3.8の低下）、特に1998年度と2002年度の落ち込みが大きいことが示された。

学力テストデータの差異項目機能（DIF）分析

孫 媛

1 DIF とは？—項目バイアスと DIF—

1960年代、1970年以降米国では「項目バイアス」研究が重視された。それは、テストが測定しようとしている能力とは別の要因によって項目正答率に集団差があることが問題となったことによる。そして、集団差を探す統計的方法の研究が活発化した。その中で「バイアス」という用語の使用については慎重であるべきで、テスト項目における集団差を指すには「DIF」という中立的な用語を用いることが望ましいという主張が出されてきた。すなわち、集団間で違いのあることが、直ちにバイアスであるとは限らないということを用語にも反映させようというものである。

比較する2つの集団について、基準となる集団を「参照集団」、DIFの影響を受ける可能性に関心のある集団を「焦点集団」と呼ぶ。例えば米国社会において、白人集団は前者、黒人・ヒスパニック集団は後者にあたることが多い。

また、DIFは「均一DIF」と「不均一DIF」に分類される。

前者は、一方の集団の項目正答率が能力尺度のどの水準においても他方の集団より高くなるようなDIFをいい、後者は、2集団の項目正答率の差がすべての尺度水準で等しくないようなDIFをいう。

2 IRTの観点から見たDIF

基本的には、1) 参照集団、焦点集団の別に項目パラメタを推定、2) 2群における θ の分布の違いについての調整（等化）を行なう、3) 2群における項目パラメタを比較する（=ICCを比較）、という手順を採る。

IRTモデルを適用して、DIF検出のための統計量を計算する方法は、1) 項目特性曲線間の面積を用いた方法、2) 項目パラメタ値の集団間の差を検定する方法に大別される。後者には、Lordの χ^2 法（Lord, 1980）、尤度比検定法（Thissen, Steinberg, & Wainer, 1993）、BILOG-MGのDIFコマンドによる方法、などがある。

なお、IRTモデルを用いることなくDIF検出のための統計量を導く、Mantel-Haenszel法、SIBTEST（Simultaneous Item Bias Test）などの方法もある。

3 学力テストデータのDIF分析

3.1 学力テストデータ

全国の公立共学高校15校の高校3年生に2001年6、7月に実施され、受験者総数は4,718名であった。

テストの構成は、国語23項目、数学17項目、英語33項目で、すべて多肢選択形式で正答か誤答かの2値で採点された。

3.2 参照集団と焦点集団

性別と学校により下位集団を設定した。性別に関しては、男性を参照集団、女性を焦点集団とした。また、学校に関しては、15校のうち、学力が平均水準に近く、受験者数も比較的多い学校Jを焦点集団として選び、その他の14校をまとめて参照集団とした。

3.3 DIF検出法

IRTに基づく方法は、1) 項目パラメタ値を正確に推定するために多くの受験者を必要とするが、今回は学校Jを焦点集団とする分析では受験者数が300名程度と少ない、2) 今回のデータは一次元性が満たされない、3) 計算手順が複雑でコストがかかる（特別なプログラムを必要とする）、などの理由から、DIF検出法にはMantel-Haenszel法を用いた。

このM-H法は、Holland & Thayer (1988) が医学・生物統計学の分野で古くから用いられてきたM-H統計量をDIF探索に適用したもので、1) 概念的に単純、2) 有意性検定ができる、3) 均一DIFに対する検出力が高い、4) 標本数が比較的少なくてもよい、5) 利用が容易（SPSSなどを使って簡単に実施でき、特別なプログラムを必要としない）、などの長所がある。

3.4 DIF分析からの示唆

各教科別に基本統計量（略）の計算を実施したのち、DIF分析を実施した。

DIF分析の結果から、1) 各科目とも性別について、数は少ないが比較的大きなDIFを示す項目が存在した。したがって、重要な決定を伴う大規模テストについては、性別DIFに関心を向けるべきである、2) 国語における「古文」「漢文」、数学における「数と式」「三角比」などは、領域として男性あるいは女性に有利になっているように見える。検証的に確認するにはSIBTESTによりDBF（Differential Bundle Functioning）を分析することが有益かと思われる、3) 学校別について、性別よりも大きなDIFが見い出される傾向がある、4) 数学や英語（特に数学）において領域により学校Jの有利・不利が目立つ、5) そのような傾向があるのならば、教育上の重点の違いなど、学校の特徴をDIF分析で見い出せる、6) カリキュラムを検討する際の重要なヒントを与えてくれる、などの示唆が得られた。