# 全国規模の学力調査における
# 重複テスト分冊法の展開可能性について

平成２３年度文部科学省委託研究
「学力調査を活用した専門的課題分析に関する調査研究」
研究成果報告書

平成２４年３月３０日

国立大学法人東北大学

# はしがき

公共財の教育施策への投資効果の検証や新しい教育施策を企画立案するための情報収集などを目的とする「調査」では，集団統計量をいかに効率的かつ精度よくとらえるかが厳格に問われる。国際的な学力調査の PISA や TIMSS，PIAAC，あるいは全米学力調査の NAEP などがその例である。しかし，いかに合理的であっても，そこで使われている新しいテスティング技術をそのままの形で我が国に導入することはきわめて難しい。我が国における「指導」重視のテスト観，いいかえれば，テストとは「指導」のためのツール，一人一人の児童・生徒の学力向上のための手がかりとして，集団統計量よりはむしろ個人スコアを得るための手段であるべきだとする伝統的な考え方がそこには厳として存在するからである。

もちろん，それ自体は我が国の教育において今後とも大切にすべき方向であることは言うまでもない。しかしながら，その一方で「調査」と「指導」は本質的に二律背反的な関係にある。たとえば，「指導」を大切にすれば，テスト問題は事後の指導に役立てるために当然公開されなければならない。また，公平性を担保するためには，参加する児童・生徒に同一問題・同一冊子で一斉に実施する必要もある。しかし，毎年問題を新しく作るため，平均点の変動が問題の難易度の変化によるものか学力の変化によるものかの区別が「調査」をしてもよくわからない，限られた時間と予算内で同一問題・同一冊子しか使えないため，本来，多角的・多面的に調べなければならない学力の総体を限定的にしか「調査」できない，などのさまざまな矛盾が生じる。

この報告は，平成 23 年度文部科学省企画公募研究「学力調査を活用した専門的な課題分析に関する調査研究」の『A．全国的な学力調査の調査手法における技術的課題に関する調査研究』に応募し審査会を経て採用された研究，「全国規模の学力調査における重複テスト分冊法の展開可能性について」の成果をまとめたものである。内容的には，昨年度の調査研究(柴山・佐藤・熊谷・佐藤，2011)で得られた成果の上に立って，本年度は，３つのサブゴール，すなわち，

１）　数学においては経年比較を可能とすること

２）　リーディング・リテラシー測定の観点から国語の同時実施

３）　数学・国語間の相関分析

を設定することで，重複テスト分冊法の全国的な学力調査における展開可能性を立体的に探った。

具体的には，１）においては，項目反応理論に基づく「尺度値」を導入することによって，たとえ学習指導要領の変更があっても学力の変化を経年的に追跡できることを，また，２）においては，国語を同時実施することで，重複テスト分冊法においても記述問題を精度よく扱えることを，さらに，３）においては，幅広い領域を調査できる重複テスト分冊法の利点を活かして，数学を基準に国語を相関分析的に検討することによってリーディング・リテラシー問題の特徴を実証的に明らかにすることをそれぞれ目指した。

この報告書が，我が国の伝統的なテスト観を大切にしながらも，それに加えて，科学的・客観的・合理的な根拠に基づく教育施策を実現するための「きめ細かな調査方式」への１つのステップになっていれば，それにすぐるものはない。

研究代表者　柴山　直

## 謝　辞

東日本大震災の深い傷が癒えない中であるにもかかわらず，この調査研究にご協力いただきました宮城県塩竈市・白石市・多賀城市・利府町・大和町の各中学校，生徒の皆様，先生方，ならびに各自治体教育委員会の諸氏，また本調査研究にご理解をいただき，厚いご支援をたまわりました宮城県教育庁教育企画室の皆様に，ここに記して深く感謝申しあげます。

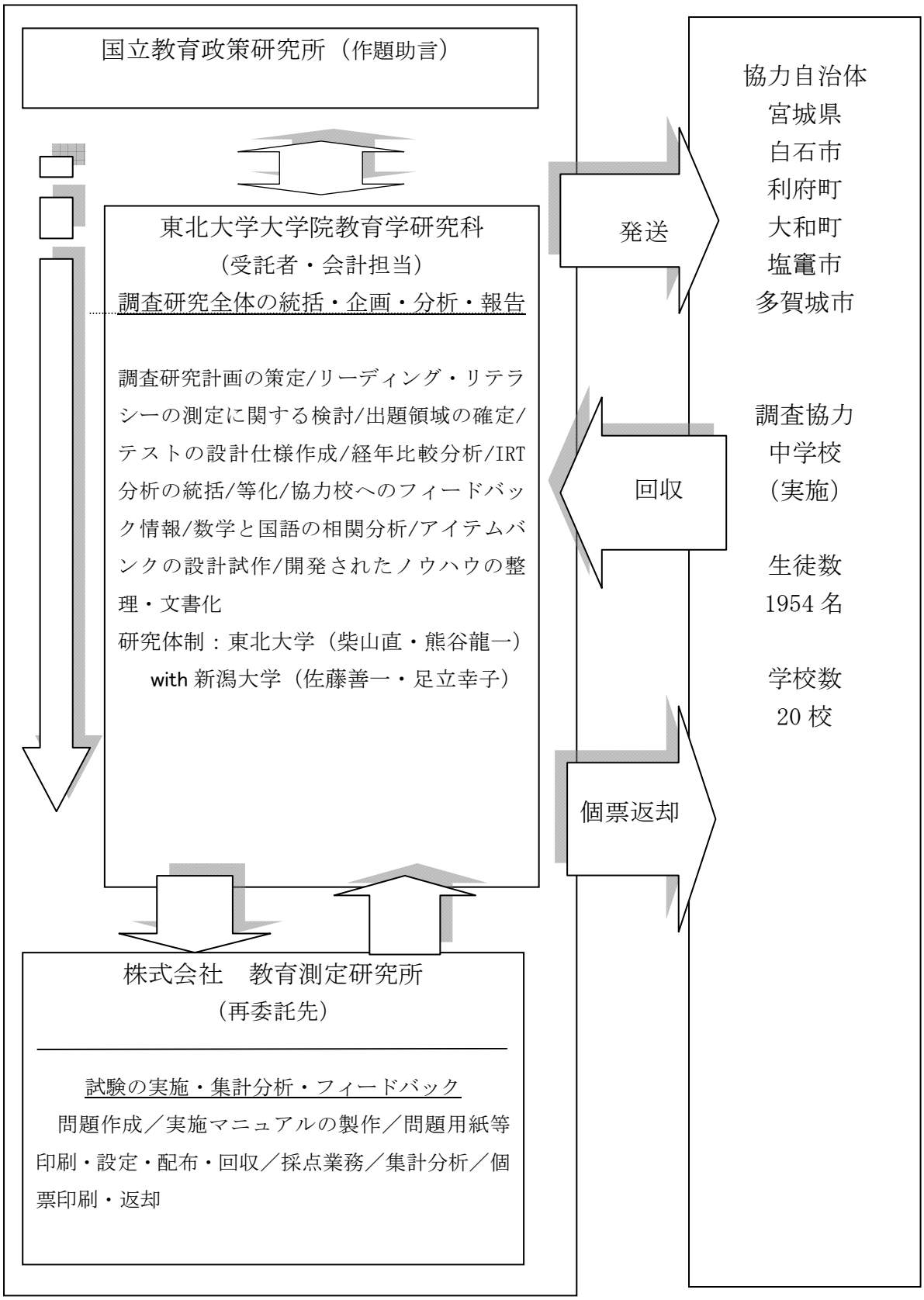
## 事業概要

| | |
|---|---|
| 事 業 名 | 学力調査を活用した専門的な課題分析に関する調査研究 |
| 事業内容 | 全国的学力調査の調査手法における技術的課題に関する調査研究 |
| 委託期間 | 平成 23 年 8 月 1 日から平成 24 年 3 月 31 日 |
| 事業者名 | 国立大学法人東北大学・大学院教育学研究科長・宮腰　英一 |
| 事 業 費 | 7,035 千円 |

## 研究組織

| | | |
|---|---|---|
| 研究代表 | 柴山　直 | 東北大学大学院教育学研究科 |
| 研究協力 | 佐藤　喜一 | 新潟大学入学センター |
| | 足立　幸子 | 新潟大学教育学部 |
| | 熊谷　龍一 | 東北大学大学院教育学研究科 |
| 研究助手 | 中野　友香子 | 東北大学大学院教育学研究科 |
| 研究補助 | 佐藤　誠子 | 東北大学大学院教育学研究科 |
| | 蛯名　正司 | 東北大学大学院教育学研究科 |
| | 宮田　佳緒里 | 東北大学大学院教育学研究科 |
| | 新国　佳祐 | 東北大学大学院教育学研究科 |
| 資料整理 | 甲斐　千晴 | 東北大学大学院教育学研究科 |
| | 千葉　陽子 | 東北大学教育学部 |
| | 坂本　佑太朗 | 東北大学教育学部 |
| 事務担当 | 紙屋　雅子 | |
| 実施集計 | 株式会社　教育測定研究所 | |
| 作題助言 | 国立教育政策研究所 | |

## 事業の実施体制図


国立教育政策研究所（作題助言）
東北大学大学院教育学研究科
（受託者・会計担当）
調査研究全体の統括・企画・分析・報告
調査研究計画の策定/リーディング・リテラシーの測定に関する検討/出題領域の確定/テストの設計仕様作成/経年比較分析/IRT分析の統括/等化/協力校へのフィードバック情報/数学と国語の相関分析/アイテムバンクの設計試作/開発されたノウハウの整理・文書化
研究体制：東北大学（柴山直・熊谷龍一）
with 新潟大学（佐藤善一・足立幸子）
株式会社　教育測定研究所
（再委託先）
試験の実施・集計分析・フィードバック
問題作成／実施マニュアルの製作／問題用紙等印刷・設定・配布・回収／採点業務／集計分析／個票印刷・返却
発送
回収
個票返却
協力自治体
宮城県
白石市
利府町
大和町
塩竈市
多賀城市
調査協力
中学校
（実施）
生徒数
1954 名
学校数
20 校

## 調査研究計画

【準備】

1. 調査デザインのための項目配置計画の策定
2. 学習指導要領移行にともなう入れ替え問題についての検討（数学）
3. リーディング・リテラシー問題についての検討（国語）
4. 出題問題の作成及び確定（数学・国語）
5. 試験実施マニュアルの作成
6. 上記事項に関する再委託先との協議
7. 協力校への依頼・実施内容の説明及び協力校の確定

【実施】

1. 調査用紙等細部点検・確認
2. 試験実施マニュアル・問題冊子・解答用紙の印刷
3. 上記事項に関する再委託先との協議
4. 協力自治体との打合せ
5. 採点ルーブリックの確定（国語）
6. 配布準備，搬送，調査実施，回収
7. 採点
8. 上記事項に関する再委託先との協議

【データ入力・集計・分析作業】

1. データ入力作業
2. 基礎統計量の集計
3. アイテムバンクの仕様設計
4. 上記事項に関する再委託先との協議
5. 協力校・協力自治体への中間報告
6. IRT 分析等データ解析作業
7. アイテムバンクの更新（数学）及び試作（国語）
8. IRT 分析等データ解析作業
9. データ解析結果の整理

【ドキュメンテーション】

1. 開発したノウハウについてのまとめ
2. 起こりうるトラブルへの対処法の文書化
3. 副次的に取得できた知見及び情報の整理
4. 報告内容に関する最終検討会
5. 最終報告書の作成

## 実施経過

2011.08.01　事業開始
株式会社教育測定研究所へ実施及び集計作業部分を再委託
2011.08.01～2011.9.12
調査デザイン・項目配置計画・入れ替え問題の検討（数学）・リーディング・リテラシー問題の検討（国語）
2011.09.13　打合せ（問題作成・実施手続き等）　於　株式会社教育測定研究所
東北大学：柴山直・熊谷龍一/新潟大学：佐藤喜一・足立幸子
2011.09.14～2011.10.31
テストの設計・調査問題の作成・採点ルーブリック（国語）の作成・検討・印刷
2011.09.22～2011.10.03
利府町・大和町・白石市・多賀城市・塩竈市の各教育委員会への調査協力依頼
2011.11.01　協力中学校宛依頼状発送
2011.11.02～2011.11.09
調査対象学級及び対象人数の確認/調査問題の配送
2011.11.10～2011.11.29
調査の実施（実施終了校から順次解答済みの調査問題の返却）
2011.11.21　打合せ（進捗状況の確認・作業工程の調整）
東北大学：柴山直　　於　株式会社教育測定研究所
2011.11.22　集計作業の開始
2011.12.08　打合せ（集計結果の中間報告及び検討）於　株式会社教育測定研究所
東北大学：柴山直・熊谷龍一・中野友香子
新潟大学：佐藤喜一・足立幸子
2011.12.15　打合せ（リーディング・リテラシー問題の質的検討）　於　東北大学
東北大学：柴山直　新潟大学：足立幸子
2012.01.11　協力校へ生徒用結果シート（個人票）ならびに学校用結果シートを返却
IRT 分析等開始
2012.02.23　打合せ（集計結果の確認・データの分析等について）
東北大学：柴山直・熊谷龍一/新潟大学：佐藤善一　於　株式会社教育測定研究所
2012.03.21　打合せ（報告書原稿の検討）於　株式会社教育測定研究所
東北大学：柴山直・熊谷龍一・中野友香子
新潟大学：佐藤喜一・足立幸子
2012.03.30　報告書提出

# 目　次

# 全国規模の学力調査における重複テスト分冊法の展開可能性について

# 1　本調査研究の概要

## 1.1　問題と目的

本調査研究は，我が国における教育施策等の検証サイクル向上のために，「平成 23 年度以降の全国的な学力調査の在り方に関する検討のまとめ」(全国的な学力調査の在り方等の検討に関する専門家会議，平成 23 年 3 月）で提案されている，

(ア) 新たな分析を可能とする調査研究を行い，

(イ) 全国的な学力調査へ適用した場合の問題点等について検討し，

(ウ) その技術的課題を明らかにすることで，

(エ) 今後の学力調査への展開可能性を探ること

の具体化をめざしたものである。そのため，昨年度において取得した重複テスト分冊法の実施ノウハウを基盤として，実際の場面になるべく近い状況を想定した上で，

１）数学においては経年比較を可能とすること，

２）主としてリーディング・リテラシー測定の観点から国語の同時実施，

３）数学・国語間の相関分析のこころみ，

の３点の技術的課題に取り組むことを本年度調査研究の目的とした。

学力測定技術としては，すでに PISA や PIAAC などの国際学力調査で採用されている項目反応理論（Item Response Theory, IRT）を引き続き採用した。IRT モデルにもとづく学力特性値（尺度値）を利用することによって，従来型のテスト得点を使う場合に対して，

①　真に測りたい学力特性の分布とテスト得点の分布とを分離して扱える，

②　項目の入れ替えがあっても，「等化」を行うことで従来の項目群と併せて，時系列的な学力特性の測定が可能となる，

③　項目の統計的性質が既知であるなら，互いに異なる項目からなる冊子で調査を実施しても，別々の冊子から得られた学力特性の値が比較可能である，

というアドバンテージが生じるのが主な理由である。

なお，IRT と組み合わせて，PISA などで実施されている調査方法は「アイテム・マトリックス・サンプリング（あるいはマトリックス・サンプリング)」と呼称されるが，そのためには項目及び受検者のサンプリングが実験計画法と標本調査法の 2 つの観点から厳密になされている必要がある。しかしながら，本研究調査の段階では，主に協力自治体や協力校への依頼の実務的な理由により厳密な実施が不可能であった。そのため，共通の項目をいくつか含みながらも構成内容が互いに異なる分冊を用いている手法上の特徴を強調して，本調査研究における実施方法の呼称としては昨年度に引き続き「重複テスト分冊法」という用語を用いることとする。厳密な実施方式に関する検討は別の機会に譲りたい。

## 1.2　数学における経年比較の試み

数学については，平成 22 年度において中学３年生を対象におこなった調査研究の際に作成済みの数学についての項目データベースを基本的に用いる。全項目中，項目分析等により品質保証の点で改善の必要なものや学習指導要領の変更にともなって差し替えるものなど計 12 項目，新たに作成するもの

20 項目により，本年度実施項目数は全部で 64 項目になる。項目はデータの信頼性を担保するため一部をのぞきすべて非公開である。ただし，内容については国立教育政策研究所の助言を受け，また，実施後得られた各統計値は昨年度報告書にすべて公開されているため，品質保証の点からも問題はない。本年度においても同様の手続きをとった。

このように問題を入れ替えると，従来の学力調査ではその得点は本質的に比較不能となる。そのため，やむをえず少数の共通項目を含ませておいてその正答率を比較するなどの方策がとられるが，教育測定論的にはわずかな数の項目別の正答率の比較だけではあまりにも誤差変動が大きく，学力の総体に対する明快な結論を導くことはきわめて困難になる。一方，問題を非公開にして毎年同一問題で実施する方式も考えられるが，我が国のようにナショナル・スタンダードとしての学習指導要領の影響がきわめて強い場合，それが改訂されれば当然問題の入れ替えが必要となり，もはや改訂前との比較は原理的に不可能になってしまう。

そこで，本調査研究では，新学習指導要領開始年度に当たっている機会を活かし，IRT モデルを利用した尺度値（学力特性値）を採用することで，これらの矛盾を乗り越えて経年比較を可能となることを実証的に示した。具体的には，昨年度の調査研究を実施した集団と本年度の実施集団とを対象にこの方法を適用することで，擬似的に年度間比較の状況を生み出し，集団統計量の推定に関するノウハウを確立した。

さらに，項目母数については昨年度データと本年度データを合併したデータを使って推定する方法や, 昨年度データによってすでに得られている項目母数を固定し本年度新作された 20 項目についてのみ推定する方法など，等化の観点からいくつかの方法を試み，実際の場面で精度を担保しつつもコスト・パフォーマンスに優れた手法を検討した。

なお，協力校へのフィードバックのため個々の生徒の学力特性値を計算し，調査実施後 6 週間程度で，それに基づく指導上有益な情報を協力校側及び生徒一人ひとりに提供した。しかしながら，本調査研究の本来的な目的からすれば，経年比較を念頭においたアセスメントとしての学力調査では個々の生徒の学力特性値よりも，例えばその年度における母集団統計量や下位集団に関する集団統計量などが重要となる。

### 1.3　リーディング・リテラシーの測定と多値 IRT モデルの必要性

全国的な学力調査において複数教科の実施は必須の前提である。そのため数学に加えて国語の実施を試みることによって，複数実施のためのノウハウの取得，記述式問題の採点ノウハウの開発，その採点結果の IRT モデルによる得点化の実証的検証が本調査研究の第 2 の目的となる。経年比較のための方法論を除けば，昨年度の調査研究ですでに算数・数学における重複テスト分冊法自体のノウハウ自体は確立できている。そのため, 本調査研究ではそのノウハウを国語へ応用展開することを試みた。

その際，知識社会到来に備え国際的にも重視されてきているリーディング・リテラシーを念頭に，平成 20 年中学校学習指導要領「国語科」と PISA 調査を視野に入れながら，本調査における独自の定義のもと，その測定を試みた。さらに，これと併行して記述式項目の採点に対応するため，多値 IRT モデルの適用可能性を探った。ここで多値 IRT モデルとは正誤のみの情報に基づく通常の IRT モデルを正誤の 2 パターン以上に一般化したものである。たとえば、採点結果を，正誤ではなく，完全に誤答，一部正答，完全に正答などのようにすれば，３とおりの値が採点結果に存在することになる。この成果は，今後，数学における証明問題や理科における実験問題などへと応用できるものである。

なお，多値 IRT モデルには，NAEP などでパフォーマンス・アセスメントのために利用されている Muraki(1982) の一般化部分採点モデル(Generalized Partial Credit model)やその基礎となった Masters(1982)の部分採点モデル(Partial Credit model)，あるいは反応カテゴリーに順序制約のない Bock(1972)の名義反応モデル(Norminal Response model)など，いくつかのバリエーションがあるが，本調査研究では順序づけられたカテゴリーへの反応確率と学力特性値 θ とを結びつける自然なモデルとして，Samejima(1969)によって提案された多段階反応モデル(Graded Response model, GR model)を採用した。

### 1.4　数学と国語との相関分析の試み

重複テスト分冊法を採用することによって，これまでの同一冊子一斉実施方式では不可能であった幅広い領域の問題を実施することができるようになる。今回，国語においてはこの利点を活かして，リーディング・リテラシーに主眼を置きつつも，いわゆる「伝統的な国語の問題」までを幅広く出題した。学習指導要領の観点から述べれば，リーディング・リテラシーはテキストを中心に種々の情報を読み取る力をさらに強調したもの,「伝統的な国語の問題」とは説明文の読解を含むことは当然として，物語文における登場人物の心情を理解できる力を主眼とするものと表現できる。

これらの力は IRT モデル上では国語の学力特性値 θ として 1 次元的に記述されるものであるが，逆に数学の学力との相関関係の文脈の中でこれらの力を分析することを試みた。その結果，リーディング・リテラシーの測定に主眼を置いた問題は，必ずしもそうでない「伝統的な国語の問題」に比べて数学の学力との相関が相対的に高いこと，リーディング・リテラシー問題における記述問題への無答率は「伝統的な国語の問題」を解く力の高い生徒群で低くなることなどが導き出せた。

以上を重複テスト分冊法が潜在的にもつ豊かな展開可能性の１つとして例示する。また，ここで示されたこの方法の展開可能性は，複数教科の組み合わせにとどまらず，たとえば，学力の形成要因としての教育社会学的な様々な指標と組み合わせることによって，的確な教育施策を施行する豊かな基礎情報をもたらすものであることもここで強調しておく。

### 1.5　本報告書の構成

本章に続く第 2 章では，本調査研究の３つの目的を実現するのに必要な理論的基礎について詳述する。問題等の入れ替えがあっても，時系列的な比較を可能とするためには，異なる年度で得られた学力特性値を交換可能なものとする必要がある。その際に使われる手続きの総称がテスト等化(test equating)である。重複テスト分冊法自体も分冊間での等化が行われている。しかし，ここでは年度を越えた時系列的な比較に重点を置いた場合の等化方法について述べる。次に，国語はいうまでもなく数学においても証明問題などでは記述式の採点が必須である。その場合には単に正誤だけではなく，段階的な採点が求められる。それに対応できる IRT モデルのバリエーションとして多段階反応モデルの数理構造を解説する。さらに，国語において主眼としたリーディング・リテラシーに関する研究の最新の成果を展望する。学力やパーソナリティ，適性などは，心理学的には構成概念とよばれ，実体としてあるのかないのかの議論には踏み込まず，それらを仮定することによって心理学的な現象を効率よく記述するためのものである。リーディング・リテラシーもまたその１つである。最新の研究成果を展望することで，その概念的定義を行うための理論的基礎を作った。

第 3 章，第 4 章ではテストの設計と開発過程，ならびに実施手続を述べた。ここで得られたノウハ

ウを文書化しておくことは，今後，新しい調査方式の実施がもし本格的に可能となった場合に役立つものとなるであろう。

第 5 章では主として分冊ごとの品質保証の観点から，伝統的な手法を用いて信頼性分析及び項目分析をおこなった。我が国の教育においては，一般にあまり認識されていないが，IRT モデルに基づいた新しい調査方式に限らず，従来型のテストであってもその品質保証のためには，このような基礎的な観点からの実証的な検証は必要不可欠のものである。

第 6 章では，本調査研究の第 1 番目の目的である学力の経年比較を具体的なデータを使いながら実施する。ここでは昨年度得られた数学の項目プール（問題項目のデータベース）の一部を，学習指導要領の変更やデータから確認された不具合のある項目の削除などの理由により差し替え，いくつかのデータ収集デザインや代表的な手法を組み合わせながら，実際に等化を試み，全国的な学力調査に適合したノウハウを取得する。

第 7 章では，国語を題材にとり，数学での検討と同様，伝統的な手法にもとづく，各分冊の品質保証，つぎに基本的な IRT 分析による項目ごとの機能の検証を行い，その上で記述問題が含まれた場合の IRT モデルの適用可能性を示す。

第 8 章では，重複テスト分冊法の展開可能性の 1 つとして複数教科間の幅広い角度からの相関分析の例をしめす。ここでは数学と古典の読解との間に相関がみられるなど興味深い結果も得られた。これらはまたリーディング･リテラシーの数学から見た場合の実証的な特徴記述になっているとともに，学力に対する複合的なアプローチの必要性をあらためて強調するものともなっている。

第 9 章では重複テスト分冊法に対する協力校の意見をまとめた。指導重視の我が国の教育現場において，本来的な意味での調査を導入する場合の現場での戸惑いや余計な負担感を回避するためのヒントとなるであろう。

資料編では本調査研究で得られた様々な情報を包括的に整理した。掲載した情報としては，数学，国語とも学習指導要領と項目の関係，各項目の正答率などの基礎統計量，点双列相関係数，因子負荷量，項目分析図，IRT 母数の推定値等，分冊及び各項目の品質を検証するために必要な情報を網羅した。また，調査にあたって協力していただいた学校等への依頼状や報告文書なども資料編としてまとめた。

なお，本研究で積み残された問題としては，1) BIB デザインの導入，2) 推算値の利用，3)マンパワーの必要性，がある。1) ，2) は技術上の問題である一方，3) に関しては，教室規模のテスト実施の延長でイメージされがちな全国的な学力調査が，その分析においても実は多大な専門的技術者集団による組織的合目的的なバックアップ体制の裏付け無しに実現不可能なことを強調しておく。

# 2　理論的準備

## 2.1　経年比較のための等化について

### 2.1.1　テスト（得点）を比較するとは

内容，問題項目が異なる 2 つ（もしくはそれ以上）のテスト得点は，そのままでは相互に比較をすることができない。受検者の学力が同一であったとしても，難しい項目群で構成されたテストではその得点は低くなり，易しい項目群で構成されたテストではテスト得点は高くなる。項目の困難度が等しいテストであっても，学力が高い受検者群が受験したテストの平均得点は高くなり，学力が低い受検者群においては平均点は低くなる。そこで，これら 2 つのテスト得点を比較可能にするための「手続き」が必要となる。“Educational Measurement”(Holland & Dorans, 2006)ではこの手続きを“linking”と呼び，さらにその目的やテストの性質により，“Predicting”，”Scale Aligning”，“Test Equating”の 3 つの下位分類を設けている(図 2.1.1)。

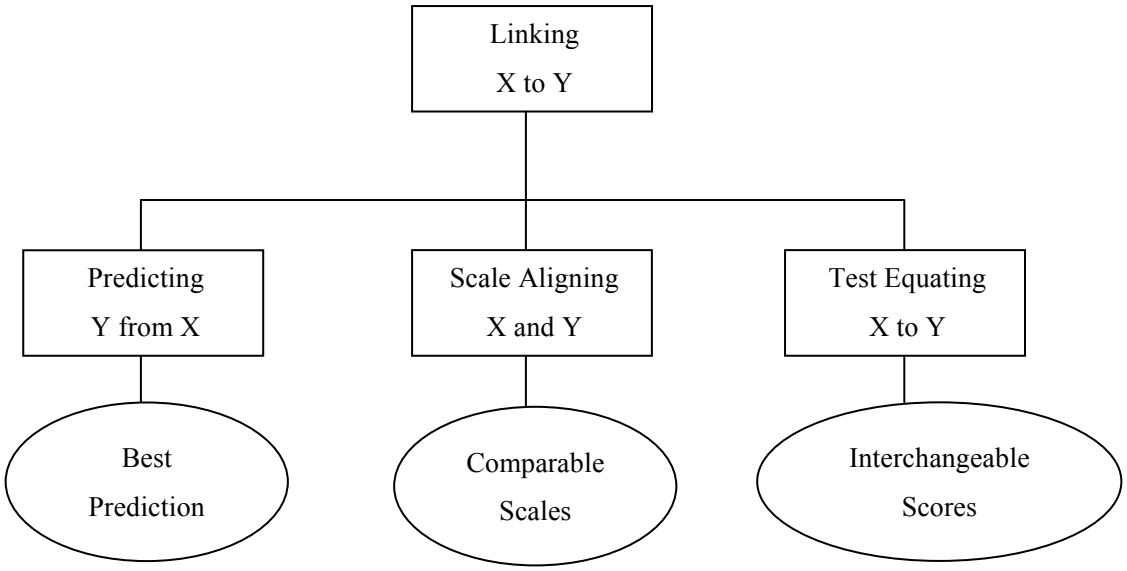



図 2.1.1　Linking の下位分類（Holland & Dorans, 2006）

ここで“Predicting”とは，名称通りテスト得点 X からテスト得点 Y を予測する関係の場合の Linking である。X から Y の方向のみを扱う「非対称性(asymmetry)」が特徴であり，具体的には入学試験の得点から，入学後の成績を予測する場合などが挙げられよう。Prediction では，2 つのテスト得点間の関係が“良い予測(Best Prediction)”となっているかどうかが問題となる。

“Scale Aligning”(Scaling と略される場合もある)では，テスト X の得点とテスト Y の得点を相互に関係づけることになる。テスト X の 60 点はテスト Y の 40 点に相当し，テスト Y の 50 点はテスト X の 75 点に相当する，というような関係づけを行うのである。これは言い換えれば，テスト X とテスト Y を“共通尺度（common scale）”上で表現しているとも言える。Scale Aligning では 2 つのテスト得点が“同等な（comparable）”な関係である。さらに Scale Aligning は，2 つのテストで測定しようとしている内容（構成概念）が同じかどうか，2 つのテストの信頼性が等しいかどうかなどにより，“Calibration”，

“Concordance”などさらなる下位分類が存在する広い概念でもある。

“Test Equating”では，2 つのテスト得点が“相互に交換可能（Interchangeable）”な関係を条件とする。“相互に交換可能”な関係について次のような例を考えてみる。たとえば，小学校 6 年生向けの算数のテストと 4 年生向けのテストを相互に関係づけたとする。このとき，小学校 6 年生向けテストの 40 点が 4 年生向けテストでは 70 点に対応するという結果が得られたとしても，実際に 4 年生向けテストで 70 点を得た 4 年生児童が，6 年生向けテストを受検したときに 40 点を得ることができるわけではない。現実には学習内容の履修・未履修などが大きく関係するが，それ以外にも，a）測定内容（構成概念）が両テスト間で異なること，b）6 年生向けテストには 4 年生向けの（易しい）問題項目がないため信頼性が低下する，などが影響するからである。このような状態は，“相互に交換可能”な関係ではないとされる。2 つのテスト得点が“相互に交換可能な”状態であるためには，2 つのテストの構成概念が同一であり，かつ両者の困難度が等しいという状況が必要となる。このように Test Equating は，Linking の 3 つの下位分類の中で最も強い制限を必要とするものである。なお，先に見た算数テストの例は，従来「垂直等化（vertical equating）」として分類されてきたが，“相互に交換可能”かどうかの観点から「垂直尺度化（vertical scaling）」として，Scale Aligning の下位分類の 1 つに区分されるようになっている。

本調査研究の数学テストは，2010 年度実施のテストと 2011 年度実施のテストにおいて，構成概念や困難度がほぼ等価なものとして設計されている。このことから，本調査研究で 2010 年度と 2011 年度の数学テストを比較する行程は「Test Equating」にあたる。以後 Test Equating を「等化」と呼ぶこととする。

### 2.1.2　IRT における等化

素点（配点に重みを設けずに正答なら 1 点，誤答なら 0 点とする）によるテストの等化手続きとしては，古典的テスト理論の枠組みの中で，線形等化法や等パーセンタイル等化法などがよく用いられてきた（これらの等化方法の詳細については，池田（1994）を参照されたい）。対して，IRT をベースとしたテスト分析が普及するに伴い，等化方法も IRT の枠組みの中で行なわれるようになってきた（村木，2011）。

IRT の枠組みにおける等化では，等化を行う 2 つのテストについてそれらを繋ぐ何らかの情報が必要となる。通常，これらは 2 つのテストを同時に受検する「共通受検者」や，2 つのテストに共通に含まれる「共通項目」などを設定することによりなされる。さらには，テストを実施する様々な状況・制限によりこれらの方法を組み合わせて用いることもある。このように，等化されるテストについて，どのようなデザインを採用するかが，非常に重要な問題となる。この等化のデザインについては，2.1.3 で詳細に見ていく。

等化のデザインが決定されると，実際にどのような計算手続きで 2 つのテストを等化するかが次の問題となる。計算手続きには，等化係数を用いる方法や，同時尺度調整法などいくつかの方法が存在するが，それらのうちいずれを選択するかは等化デザインと大きく関わっている。ただしここで気をつけなければいけないのは，等化デザインと等化の計算手続きとが必ずしも 1 対 1 で対応しているわけではないことである。ある等化デザインでテストが実施されたときに，複数の方法で等化の計算を行うことが可能な場合も多いのである。等化の計算手続きについては，2.2.4 で詳細に見ていく。

IRT における等化においては，「等化のデザイン」と「計算手続き」をどのようにするかが非常に重要である。とくに等化デザインは，受検者数，受検者への負担，テスト実施時のコスト，テストに含

まれる項目数など様々な実施上の制約の下で決定していかなければならず，また等化の成否（等化が上手くいくかどうか）だけでなく，それがどの程度（未知であり，また直接観測することのできない）真の状態を反映しているかに関する等化の精度にも大きく影響する。実際に IRT における等化を行う際には，どれだけ綿密に等化のデザインを構築できるかにその成否がかかっているといっても良い。

### 2.1.3　等化デザイン

先に述べたように，IRT における等化のためには 2 つのテストを繋ぐ何らかの情報が必要となる。通常これらを実現するために，「共通項目デザイン」，「共通受検者デザイン」，さらにはそれらを組み合わせた形の「係留テストデザイン」などが用いられる。それぞれのデザインごとに長所・短所があり，現実場面では様々な制約状況の中で，どのようなデザインを組むのか決めていくことになる。

#### （1）共通項目デザイン

共通項目デザインは，等化を行う 2 つのテストに共通の問題項目群を含めるデザインである（図 2.1.2 参照）。この共通項目群が，2 つのテストを繋ぐ情報となる。共通項目デザインは，次に述べる共通受検者デザインとは異なり，別途のテストや受検者を必要としないことから，実施面での運用のしやすさが特徴である。ただし，日本のテスト文化では，一度でもテストとして使用した問題項目はその後は利用しない，もしくは実施したテスト（問題項目）は全て公開する，という「項目使い捨て主義」ともいうべき運用がなされることも多い（例えば豊田，2002）。このような場合には，共通項目デザインをこのまま用いることはできない。またテストによっては，1 つのテスト内に含まれる項目数が非常に少ない（たとえば 10 項目など）場合もある。共通項目デザインでは，共通項目の数が多い方が等化の精度が高まるが，1 つのテストの項目数に制限がある場合など，共通項目の確保という点から共通項目デザインを用いることが難しい場合もある。

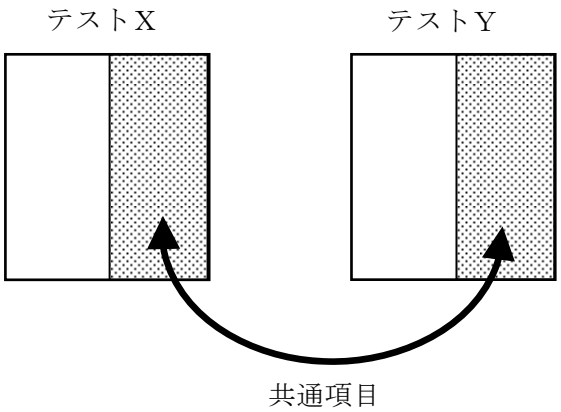



図 2.1.2　共通項目デザイン

#### （2）共通受検者デザイン

共通受検者デザインは，等化を行う 2 つのテストを同一受検者群に受検させるデザインである（図 2.1.3 参照）。このとき，受検者は 2 つのテストを「同時に」受検しなくてはいけないが，現実の実施場面では，測定目的となっている能力が変化しないと見なせる期間において両テストを実施することが多い。さらには，2 つのテストを受検する順番の効果がないように，ある受検者Aはテスト X，テス

トＹの順に，別の受検者Ｂはテスト Y，テストＸの順に受検するなどのカウンターバランスをとることなども行なわれる。

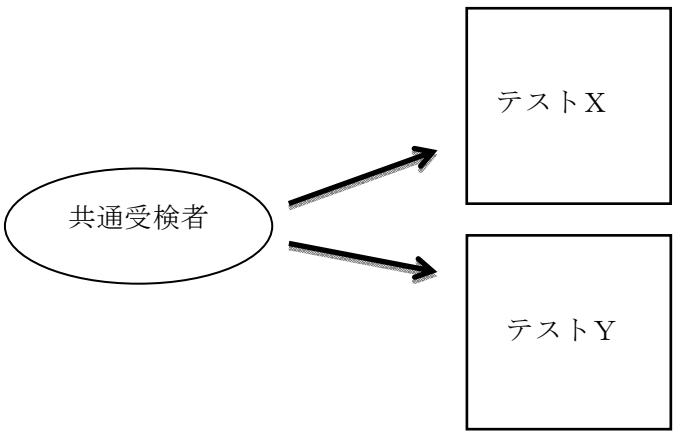



図 2.1.3　共通受検者デザイン

共通受検者デザインでは，2 つのテストに共通項目がないので，先に述べた「項目使い捨て主義」による影響を受けることがない。また，共通受検者の人数を多くすることで，等化の精度を上げることができる。共通項目デザインでは，等化の精度を上げるためには共通項目の数を多くしなければならないが，1 つのテストに含めることができる項目数などの制限を受ける。共通受検者デザインでは，（原理的には）共通受検者の人数には制限がないため安定した等化を行うことができることから，米国における全国学力調査である NAEP(National Assessment of Educational Progress)でも用いられている（村木，2006）。

一方，現実場面においては，2 つのテストをほぼ同時に受検する集団を確保することは，様々なコストの面で難しい場合も多い。

### （３）係留（アンカー）テストデザイン

係留テストデザインは，等化を行う 2 つのテストの他に，それらを繋ぐもう 1 つのテスト（係留テスト，アンカーテスト）を利用して，等化を行うデザインである（図 2.1.4 参照）。これは，共通項目デザインと共通受検者デザインを組み合わせた方法と見ることもできる。テストＸとテストＺ，テストＹとテストＺが共通受検者デザインであり，テストＺ自体が共通項目（共通テスト）となっている。

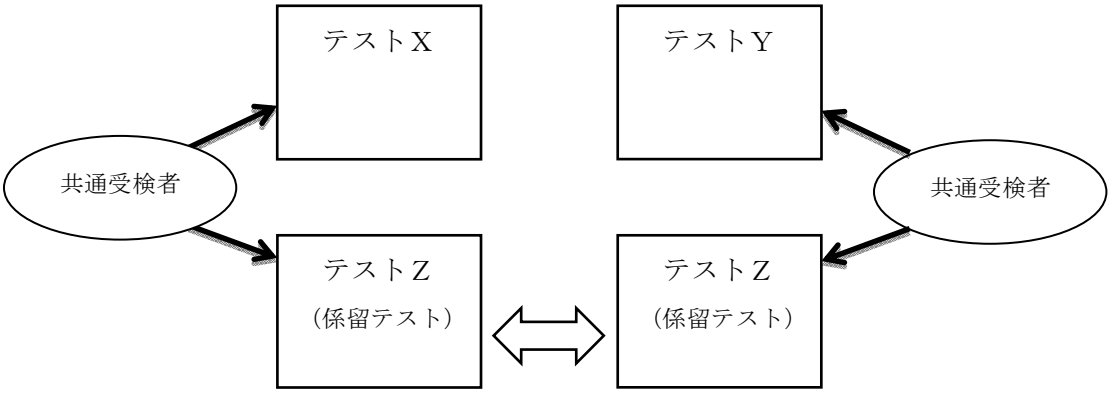



図 2.1.4　係留テストデザイン

ここで取り上げた 3 つのデザインは，2 つのテストを等化する際の最もシンプルな状態を示したものである。たとえば熊谷・山口・小林・別府・脇田・野口（2007）では，係留テストデザインを複雑に多数組み合わせ，33 ものテストを等化している。また，野口・熊谷・大隅（2007）では，テストXとテストYについて，共通項目も共通受検者も利用できない状況において，図 2.1.5 のような等化デザインを利用している。このデザインでは，テストX及びテストYから項目群を抽出し，それらを結合することで係留テストZを構築している。これにより，テストX，Yと係留テストとは共通項目デザインで繋がっている。さらに係留テストZは実際には 1 つのテスト冊子として実施されたが，これをテストXからの共通項目部分とテストYからの共通項目部分というあたかも 2 つのテストを受検したように分析を行なっている。つまり，係留テストZについては，テストX部とテストY部を共通受検者デザインで等化するのである。

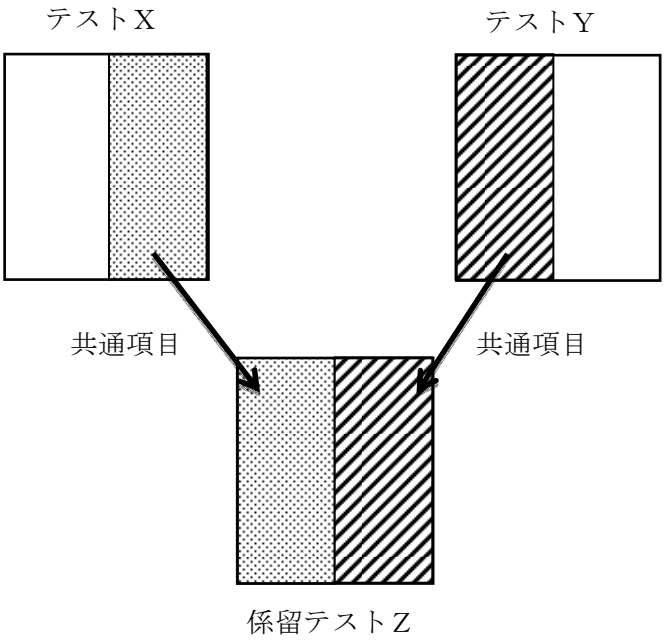



図 2.1.5　野口・熊谷・大隅（2007）による等化デザイン

繰り返しになるが，等化デザインを決定する際に考慮しなければならないのは，共通項目数や共通受検者数など，テストを繋いでいる情報をできるだけ多くすることである。それでは実際にどの程度の項目数や受検者数が必要になるかというと，これを一概に決定するのは非常に難しい。シミュレーション研究などで誤差の大きさなどを考慮した研究はいくつか行なわれているものの，実際にテストを運用する場面ではコストや実施時間など様々な制限が課された状況下において，等化デザインを構築することが迫られるからである。さらには，等化は複数年・複数回にわたり実施されることも多い。このとき，継続して実施できるようなデザイン構築が求められる。等化デザイン決定においては，このデザインを利用すればどんな状況でも対応できるというようなものは存在しないため，精度，実施可能性，コストなど様々な条件を勘案して構築することが必要不可欠なのである。

### 2.1.4　等化の計算手続

等化のデザインが決定された後，そのデザインに従ってテストデータが収集される。そして，得られたテストデータを用いて，実際に等化の計算が行なわれる。等化デザインと同様に，等化の計算手続きにおいても多種多様な方法が提案されている。以下に，その代表的なものを紹介する。

#### （1）等化係数を用いた方法

IRT においては，受検者母数といくつかの項目母数からなる項目特性関数により，項目に対する正答確率を表すことになる。IRT での代表モデルである 2 母数ロジスティックモデルでは，尺度値 θ をもつ受検者が項目 j に正答する確率 $P_j(\theta)$を

$$P_j(\theta) = \frac{\exp\left[Da_j(\theta - b_j)\right]}{1 + \exp\left[Da_j(\theta - b_j)\right]} \tag{2.1.1}$$

として表す。2 母数ロジスティックモデルの詳細については柴山・佐藤・熊谷・佐藤（2011）の第 3 章及び第 7 章を参照されたい。さて，このとき受検者母数 θ について

$$\theta^* = K\theta + L \tag{2.1.2}$$

と線形変換を施したものを $\theta^*$とする。同時に項目母数 $a_j$，$b_j$ についても

$$a^* = a/K \tag{2.1.3}$$

$$b^* = Kb + L \tag{2.1.4}$$

のように線形変換したものを $a^*$，$b^*$とすると，

$$P_j(\theta) = \frac{\exp\left[Da_j(\theta - b_j)\right]}{1 + \exp\left[Da_j(\theta - b_j)\right]} = \frac{\exp\left[Da_j^*\left(\theta^* - b_j^*\right)\right]}{1 + \exp\left[Da_j^*\left(\theta^* - b_j^*\right)\right]} = P_j(\theta^*) \tag{2.1.5}$$

のように線形変換を施す前と後とで正答確率を表す項目特性関数が等しくなることが示される。このことはすなわち，学力を表現する尺度値 θ について，K，L を用いて原点と単位とを任意に決定することができることを表している。通常の IRT 分析では，分析に用いたデータセットにおける受検者集団の母集団について，平均 0，標準偏差 1.0 となるように基準を定めて項目母数の推定などが行なわれる。前節の例で言えば，テストX，テストYはそれぞれ独立に分析した状態では，それぞれの母集団分布の平均が 0，標準偏差が 1.0 となっており，直接それらを比較することができない。そこで，（2.1.2）

式を用いて，たとえばテストＹの原点と単位をテストＸのそれに合わせるような作業を行うことで，両テストを比較可能にするのである。このときに用いている K 及び L を等化係数と呼ぶ。

等化係数を用いた等化では，いかにしてこの等化係数を定める（推定する）のかが問題となる。等化係数の推定にも様々な方法が提案されているが，なかでも最も簡便な方法として Marco(1977)による mean / sigma 法が挙げられる。この mean / sigma 法は，テストＸ及びテストＹに含まれる共通項目について，それぞれデータセットから推定された困難度母数 $b_X$，$b_Y$ の平均 $\mu_{bX}$，$\mu_{bY}$，標準偏差 $\sigma_{bX}$，$\sigma_{bY}$ を用いて，$K=(\sigma_{bY}/\sigma_{bX})$，$L=\mu_{bY}-K\,\mu_{bX}$ として等化係数の推定を行うものである。なお共通受検者デザインの場合には，2 つのテストから得られる尺度推定値 $\theta_X$，$\theta_Y$ の平均，標準偏差を用いて，同様に $K=(\sigma_{\theta Y}/\sigma_{\theta X})$，$L=\mu_{\theta Y}-K\,\mu_{\theta X}$ として mean / sigma 法を適用することが可能である。

なお，mean / sigma 法の他にも，2 つの項目特性関数の差を最小にする方法（Haebara, 1980）や，テスト特性関数の差を最小にする方法（Stocking & Load, 1983），共通受検者の項目反応パターンを利用して等化係数を最尤推定する方法（野口，1986），mean / sigma 法において誤差の影響を取り除く方法（野口・熊谷，2011）など様々な方法が提案されている。

### （２）同時尺度調整法

同時尺度調整法（concurrent calibration）は，等化デザインの下で得られたテストデータ全体を用いて，同時に項目母数の推定を行うことで等化を行う方法である。共通項目デザイン及びアンカーテストデザインにおいて，行方向を受検者，列方向を項目とした項目反応データ行列をデータ全体で示すと，それぞれ図 2.1.6 のように示される。それぞれのデータ行列において，項目が提示されない部分（未提示項目 : not presented items）が存在する。このデータ行列全体について，項目母数の推定を行うことで，データ全体が 1 つの尺度上で表現されることとなり，等化が実現される。なお，未提示項目については，項目母数推定計算時の尤度関数に含めないこととする。さらに，多母集団モデルを用いた計算ができるソフトウェア（BILOG-MG（Zimowski, Muraki, Mislevy, & Bock, 1996）や EasyEstimation（熊谷，2009））などでは，テストＸ（もしくはテストＹ）を受検した集団を基準集団（母集団分布の平均を 0，標準偏差を 1 とする）として分析を行うことができる。単一集団での分析の場合には，データ行列に含まれる集団全体の母集団分布について，平均を 0，標準偏差を 1 とすることとなる。

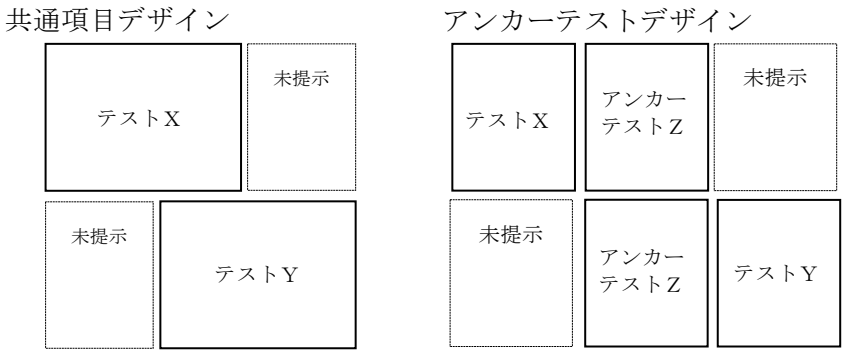



図 2.1.6　項目反応データ行列

### （３）項目固定法

項目固定法（fixed items method）は，等化したいテストの項目母数推定時に，すでに得られているテストデータから推定した項目母数を既知項目として利用し，等化を行う方法である。具体的に図 2.1.6 の共通項目デザインにおける項目固定法の手続きは以下のようになる。

１）テストＸの項目母数を推定する。

２）テストＹの項目母数を推定する際に，共通項目部については１）で得られた項目母数で「固定」して推定する。このとき，母集団分布の平均を 0，標準偏差を 1 とするような制限は必要としない。

以上の手続きにより，テストＹの項目母数は全てテストＸ上の尺度で表現されることとなる。この計算手続きは，前述の BILOG-MG や EasyEstimation を用いることができる。特に EasyEstimation では，項目固定を行う手続きについてマウス操作を前提とした GUI により，簡便に分析を行うことが可能である（図 2.1.7）。

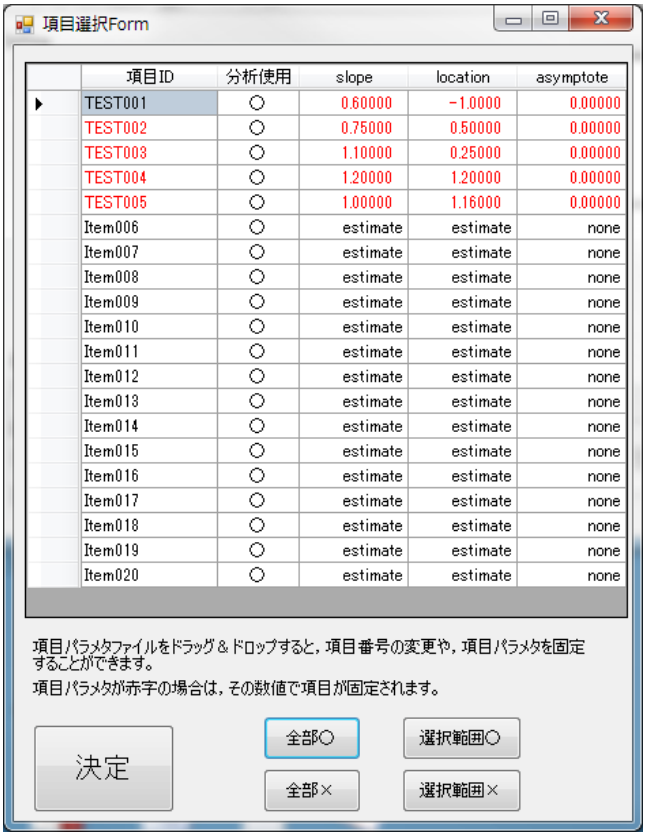



図 2.1.7　EasyEstimation における項目固定法画面

## 2.2　多値項目反応モデルの導入

### 2.2.1　項目の形式

テスト項目の形式には, 伝統的な形式だけでも, 論文形式（essay form）, 短答形式（short-answer form）, 真偽形式（true-false form），組合せ形式（matching form），配列形式（arrangement form），多肢選択形式（multiple-choice form）などが存在する（肥田野，1972）。本調査研究の国語のテストにおいては, リーディング・リテラシーの測定を目的として，多肢選択形式の項目及び短答形式の項目が出題されている。

多肢選択形式の項目は, 設問（幹, stem）とそれに対するいくつかの解答（選択肢または枝, alternatives）から構成される。選択肢には，１つの正しい解答（correct answer）あるいはもっとも適切な解答（best answer）と，いくつかの正しくない解答（まよわし，distracter）が含まれる。前者はまとめて正答（選択肢），後者は誤答（選択肢）と呼ばれる。この形式の問題では，受検者は選択肢の中から自分が正答だと考える選択肢を解答する。以下に，本調査研究の国語のテストからの出題例を 2 例だけあげておく。

問　著者は――部①「話すように書けばいい」ということをどのようにとらえていますか。次の中から適切なものを一つ選びなさい。

1　話すように書けば，だれもが苦手意識をもたずに書けるので良い。
2　話したことは流れ去ってしまうが，話したことは書くことで記録できるので，話すように書くことは良い。
3　話すことと書くことはまったく違う行為なので，話すように書くことは難しい。
4　ライブで話すときの言葉の強さは書くことでは表現できないので，話すように書くことはあきらめた方が良い。

問　書き言葉の特徴としてあげられているものを，次の中から一つ選びなさい。

1　表現力
2　定着力
3　説明力
4　生命力

多肢選択形式の利点の１つは，採点の客観性が高いことである。すなわち，採点は単純で時間もかからず，誰が採点しても同一の採点結果を得ることが可能である。それ以外にも，信頼性・妥当性の高いテストを作成する技術が発達している，かなり高度に複雑な能力を測定することも可能である，マークシートを利用した機械採点にも向いているといった利点もある。そのような利点から，多くのテスト現場では多肢選択形式がもっとも頻繁に用いられている。その反面，受検者が無作為に選択肢を選んでも「1/選択肢数」の確率で正答してしまうという，測定上，好ましくない当て推量（guessing）の問題が存在する。さらに作題者にとっては，まよわしの作成が意外にむずかしいといった欠点も指摘されている。まよわしは，幹と密接に関連し，正答と紛らわしく，もっともらしい選択肢であることが望ましい。項目の良し悪しは，まよわしの出来いかんによって決まるといっても過言ではない。

短答形式の項目には，設問中の空欄を受検者が正しいと考える字句で埋めるものや，疑問文の設問に対して受検者が自発的に解答するものがある。解答は，用語，数値，人名，年号，事実のような短い字句の場合もあるし，字数制限を含むような短い文章の場合もある。とくに，文章で答えさせる場合は，短い字句を答えさせる場合と区別して論述形式と呼ばれることがある。以下に，本調査研究の国語のテストからの出題例を 2 例だけあげておく。

問　――部「やったぜと私は思い，」とありますが，「私」が「やったぜ」と思った理由を，四十五字以上，六十字以内で書きなさい。

問　この文章では最後の一文だけ「～です」が使われていますが，それによってどのような表現効果があるか書きなさい。

短答形式には，受検者が自発的に解答するという多肢選択形式にはない特徴がある。そのため，多肢選択形式の欠点である当て推量を防ぐことができる。その反面，正答が複数ある場合や解答のつづり字に間違いがある場合は正答の判断が主観的になることがあり，採点の客観性が低くなることが少なくない。とくに複数の採点者がいる場合，あらかじめ採点者間で採点基準をそろえるといった手続きが必要になってくる。

### 2.2.2　多値モデルの必要性

多肢選択形式の項目の場合，採点結果は基本的に 2 値データ（0: 誤答，1: 正答）として表現できる。もしテストがこの形式の項目から構成されていれば，図 2.2.1（左）に示すように，テスト結果は行を受検者，列を項目とした 2 値行列として表現できる。このような行列データは，項目反応データ（item response data）あるいは項目反応パターン（item response pattern）などと呼ばれる。2 値の項目反応データを取り扱うのに適した項目反応モデル（item response model）には，Rasch モデル（Rasch model）（Rasch, 1960），1 母数ロジスティックモデル（one parameter logistic model, 1PL モデル），2 母数ロジスティックモデル（two parameter logistic model, 2PL モデル），3 母数ロジスティックモデル（three parameter logistic model, 3PL モデル）（Birnbaum, 1968; Lord, 1952; Lord & Novick, 1968）がある．このうち，Rasch モデルと 1PL モデルは数学的には同一であるものの，その発展過程が異なることから，それぞれ Rasch 系モデルと Thurstone 系モデルに属するモデルとして区別される（村木，2011）。

短答形式（論述形式）の項目の場合，採点基準によっては部分点（partial credit）を与えて多段階に採点することがある。このとき，テスト結果は 2 値データではなく多値データとして表現される。もし各項目が（0: 誤答，1: 部分点，2: 正答）と 3 段階の順序データとして採点されるなら，テスト結果として図 2.2.1（右）に示すような多値の項目反応データが得られる。項目反応モデルの中には，このような段階反応（graded response）を取り扱うことのできる多値モデルが考案されている。その代表例として，段階反応モデル（graded response model, GR モデル）（Samejima, 1969），部分採点モデル（partial credit model, PC モデル）（Masters, 1982），一般化部分採点モデル（generalized partial credit model, GPC モデル）（Muraki, 1992）があげられる。

前節で例示したように，本調査研究の国語の問題には，多肢選択形式の項目と短答形式（論述形式）の項目の両方が含まれている。本調査研究では，多肢選択形式の項目には 2 母数ロジスティックモデルを適用し，短答形式の項目には段階反応モデルを適用して IRT 分析を試みる。2 母数ロジスティックモデルの概要については，柴山・佐藤・熊谷・佐藤（2011）の第 3 章及び第 7 章を参照されたい。段階反応モデルについては，2.2.4 節で概説する。

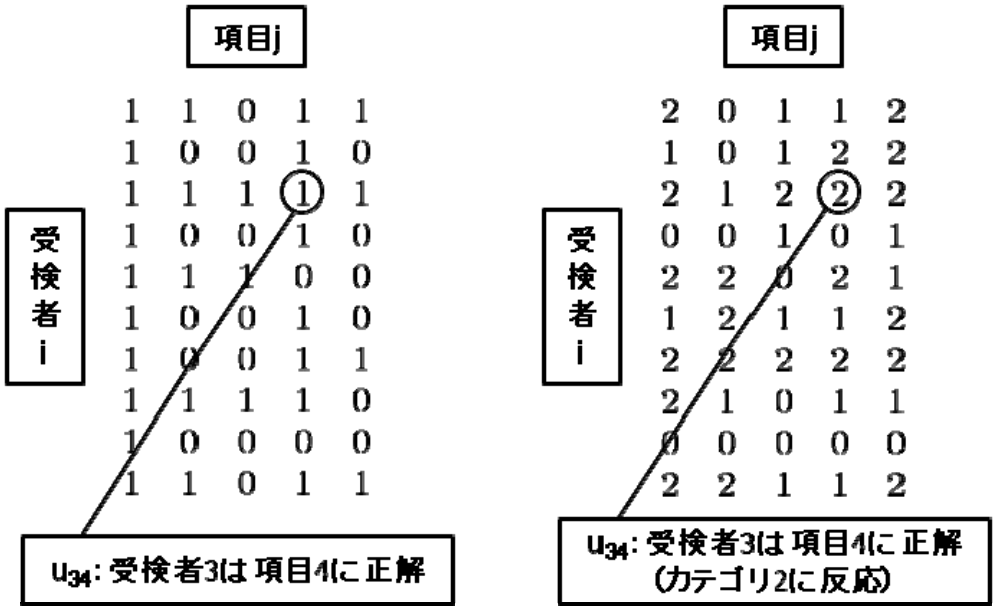



図 2.2.1　2 値の項目反応データ（左）と多値の項目反応データ（右）の例

### 2.2.3　モデル選択の問題

とくに欧米を中心に，項目反応理論（item response theory, IRT）は学力調査をはじめ資格試験や適性試験などに広く利用されている。IRT によって運用されているテストの例として，全米学力調査（National Assessment of Educational Progress, NAEP），LSAT（Law School Admission Test），PISA（Programme for International Student Assessment），TOEFL（Test of English as a Foreign Language），（医療系大学間）共用試験（医学系 CBT）などがあげられる。

IRT を利用する際には，各項目にどの項目反応モデルを適用するかを決定する必要がある。その判断基準には，項目形式，受検者数，項目数，モデルフィットの問題，結果の解釈（説明）しやすさの問題，さらには判断する人の哲学や好みなど，様々な要素が考えられる。そのため，モデルを選択する際には，テストの専門家などによる総合的な判断が求められる。

全米学力調査の場合，多肢選択形式の項目には 3PL モデル，採点結果が正答・誤答の 2 値で表現される短答形式の問題には 2PL モデル，採点結果が段階反応となる論述形式の項目には GPC モデルが採用されている。PISA では，選択肢形式，論述形式などの項目に 1PL モデルが利用されている。また，（医療系大学間）共用試験（医学系 CBT）では，多肢選択形式の項目に 2PL モデルが利用されている。

本調査研究の場合，多肢選択形式の項目には 2PL モデルを採用することにした。多肢選択形式の項目の場合，当て推量母数がモデルに含まれる 3PL モデルを採用することも考えられる。しかしながら，受検者数，結果の解釈のしやすさの問題，昨年度の調査（柴山・佐藤・熊谷・佐藤，2011）からの継続性に重点をおいて 2PL モデルを採用するに至った。また，昨年度の調査では存在しなかった論述形式の項目については，試験的に GR モデルを採用することにした。

### 2.2.4 段階反応モデル

Samejima（1969）が考案した GR モデルは，多段階で採点されるテスト結果の分析だけでなく，多くの心理検査や社会調査で用いられるリッカートタイプの質問項目（例：質問に対し，5 段階で当てはまる程度を答えさせる）などにも適用される。本節では，GR モデルを一般的な利用の文脈で説明するため，テストの文脈からはやや奇異な表現がみられる。「反応」を「採点」と置き換えて考えるなど，適宜，捕捉して理解していただけると幸いである。

項目 j は多段階に採点される項目であり，その段階反応（採点結果）は K 個（0, 1, …, K-1）のカテゴリーに分類されるとする。GR モデルでは，潜在特性値 θ をもつ受検者が項目 j にカテゴリーk と反応する確率（採点される確率）$P_{jk}(\theta)$は，

$$P(u_j = k|\theta) = P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{jk+1}(\theta) \tag{2.2.1}$$

と定義される。$u_j$ は受検者の項目 j への反応であり，右辺の $P^*_{jk}(\theta)$は潜在特性値 θ をもつ受検者が項目 j に $u_j \geqq k$ と反応する確率を表している。すなわち，$P_{jk}(\theta)$は潜在特性値 θ をもつ受検者が k 以上のカテゴリーに反応する確率と k+1 以上のカテゴリーに反応する確率の差として定義される。θ を変数として見たとき，$P_{jk}(\theta)$は項目反応カテゴリー曲線（item response category characteristic curve, IRCCC），$P^*_{jk}(\theta)$は境界特性曲線（boundary characteristic curve, BCC）と呼ばれる。

GR モデルでは，BCC を正規累積モデルあるいはロジスティックモデルによって表現する。導関数を計算しやすいなど，数学的な取り扱いが容易であるという理由から，BCC として 2PL モデルがよく利用される。カテゴリーk=1,2,…,K-1 における $P^*_{jk}(\theta)$を 2PL モデルによって表現すると，

$$P^*_{jk}(\theta) = \frac{1}{1+\exp[-Da_j(\theta - b^*_{jk})]} \tag{2.2.2}$$

となる。ただし，受検者は必ず K 個（0, 1, …, K-1）のカテゴリーのいずれかに反応するものとし，

$$P^*_{j0}(\theta) = 1 \tag{2.2.3}$$

$$P^*_{jK}(\theta) = 0 \tag{2.2.4}$$

とする。

(2.2.2)式の BCC に含まれる母数のうち，項目の特性を記述するための母数は，識別力母数（discrimination parameter）$a_j$ と BCC の位置母数（location parameter）$b^*_{jk}$ である。GR モデルでは，段階的な反応を記述するため，項目内の識別力母数はすべて等しく，BCC の位置母数の値はカテゴリーk の昇順に大きいと仮定する。それゆえ，$a_j$ の添え字にカテゴリーk は含まれず，BCC の位置母数には，

$$b^*_{j1} < b^*_{j2} < \cdots < b^*_{jk} < \cdots < b^*_{jK-1} \tag{2.2.5}$$

という大小関係がある。(2.2.2)式の D は尺度要素（定数）であり，通常，D=1 や D=1.7 が利用される。D=1.7 とすれば，(2.2.2)式は正規累積モデルの非常によい近似となる。

(2.2.1)式の IRCCC は，識別力母数 $a_j$ と IRCCC の位置母数 $b_{jk}$ によって記述される。カテゴリー0 とカテゴリーK-1 については，それぞれ $P_{j0}(\theta)=0.5$ と $P_{jK-1}(\theta)=0.5$ となる潜在特性値を IRCCC の位置母数として利用する。(2.2.2)式の 2PL モデルでは $\theta=b^*_{jk}$ のとき $P^*_{jk}(\theta)=0.5$ となることに注意すれば，(2.2.1)式，(2.2.3)式，(2.2.4)式から IRCCC の位置母数 $b_{j0}$,$b_{jK-1}$ は，

$$b_{j0} = b^*_{j1} \tag{2.2.6}$$
$$b_{jK-1} = b^*_{jK-1} \tag{2.2.7}$$

と表現される。また，カテゴリーk=1,2,…,K-2 については，IRCCC の位置母数 $b_{jk}$ として，そのカテゴリーをとる確率がもっとも高くなる潜在特性値を利用する。(2.2.1)式が(2.2.2)式の k,k+1 との差で定義されることから，カテゴリーk の IRCCC には識別力母数 $a_j$ と BCC の位置母数 $b^*_{jk}$,$b^*_{jk+1}$ が含まれる。このとき，IRCCC の位置母数 $b_{jk}$ と BCC の位置母数 $b^*_{jk}$,$b^*_{jk+1}$ との関係は，

$$b_{jk} = \frac{b^*_{jk} + b^*_{jk+1}}{2} \tag{2.2.8}$$

と表現される。

図 2.2.2（左）～図 2.2.4（左）に，3 つのカテゴリーk=0,1,2 をもつ項目の IRCCC の例を示す。各項目の識別力母数の値，IRCCC の位置母数の値は図下（注）に示すとおりである。(2.2.2)式の尺度要素は，いずれの項目も D=1 を利用している。

図 2.2.2（左）の IRCCC をみると，潜在特性値 θ に対する項目 1 の特性を把握することができる。すなわち，θ=0 付近の潜在特性値をもつ受検者はカテゴリー1 にもっとも反応しやすく，それより小さい θ をもつ受検者はカテゴリーk=0,1,2 の順に反応しやすく，それより大きい θ をもつ受検者はカテゴリーk=2,1,0 の順に反応しやすいことがわかる。また，最下位のカテゴリー0 に反応する確率は潜在特性値 θ に対して右下がりの曲線で表現され，$\theta=b_{10}=-1$ のときにその反応確率が 0.5 になっている。同様に，最上位のカテゴリー2 の場合，反応確率は右上がりの曲線で表現され，$\theta=b_{12}=1$ のときに反応確率がちょうど 0.5 になっている。一方，中間のカテゴリー1 の場合，そのカテゴリーに反応する確率は単峰形

の左右対称な曲線で表現され，$\theta=b_{11}=0$ のときに最大値をとっている。

図 2.2.2（左）と図 2.2.3（左）を比較すると，識別力の違いが IRCCC にどんな影響を与えるかを理解することができる。項目 1 と項目 2 の違いは識別力母数の値だけであり，IRCCC の位置母数は同一の値である。両者の IRCCC をみると，項目の識別力が高くなると IRCCC の傾斜が急になる傾向があり，受検者の反応したカテゴリーに応じて受検者の潜在特性値を識別しやすくなることがわかる。

図 2.2.4 は，IRCCC の位置母数の間隔が極端に狭く，中間のカテゴリーへの反応確率が非常に低い場合の例である。（0：誤答，1：部分点，2：正答）と採点する項目ならば，項目 3 は部分点をとる受検者が極端に少なく，受検者の解答が正答か誤答かのどちらかにはっきりと分かれる特性をもつ。数学の問題で例をあげるなら，2×3+5 を解くにあたり，正しく解けなかった人は 0 点，2×3 まで解けた人は 1 点，正しく解けた人は 2 点と採点することにする。通常，たし算はかけ算よりかなり易しいので，2×3 まで解ける人は最終的な正解まで答えられる可能性が高い。このような場合，部分点をとる受検者は極端に少なくなり，受検者の解答は正答か誤答かにはっきりと分かれる。

ところで，BCC に(2.2.2)式の 2PL モデルを利用する場合，2 つのカテゴリーk=0,1 をもつ GR モデルは 2PL モデルに一致する。(2.2.1)式及び $P^{*}_{j0}(\theta)=1, P^{*}_{j2}(\theta)=0$ より，

$$P_{j0}(\theta) = P^{*}_{j0}(\theta) - P^{*}_{j1}(\theta) = 1 - P^{*}_{j1}(\theta) \tag{2.2.9}$$

$$P_{j1}(\theta) = P^{*}_{j1}(\theta) - P^{*}_{j2}(\theta) = P^{*}_{j1}(\theta) \tag{2.2.10}$$

となる。2 つのカテゴリーを（0: 誤答，1: 正答）と考えれば，(2.2.9)式と(2.2.10)式は，それぞれ 2PL モデルの誤答確率と正答確率を表現している。

2PL モデルなどと同様に，GR モデルは潜在特性値 θ が 1 次元（スカラー）の項目反応モデルに属する。それゆえ，GR モデルを適用するには，対象となるテストが局所独立の仮定と 1 次元性の仮定をある程度は満たす必要がある。局所独立の仮定とは，1 つのテストにおいて，ある潜在特性値をもつ受検者がある項目に正答する確率は他の項目に正答する確率の影響を受けないという仮定である。確率論的には，ある受検者が各項目に正答するのは互いに独立な事象であるということを意味する。1 次元性の仮定とは，1 つのテストを構成する項目はただ 1 つの構成概念を測定するものでなければならないという仮定である。実際には，スクリープロットや各種の統計量を用いてモデルに必要な仮定やモデルのデータへのあてはまり具合などを確認する。

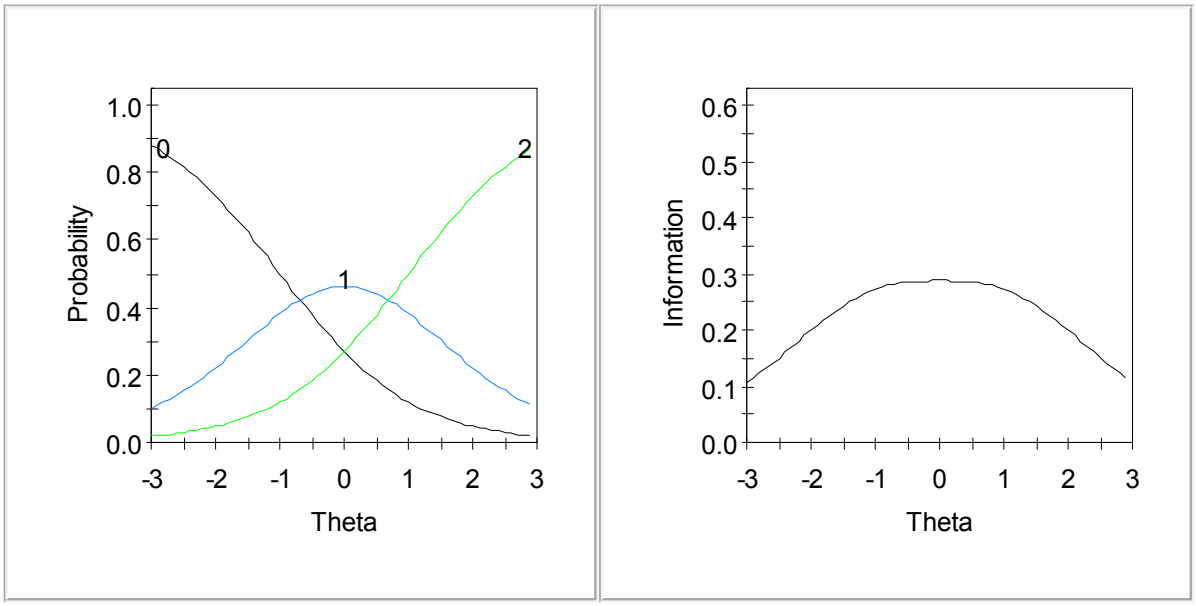



図 2.2.2　項目 1 の IRCCC（左）と項目情報量（右）

（注）$a_1$=1, $b_{10}$=-1, $b_{11}$=0, $b_{12}$=1, (D=1)

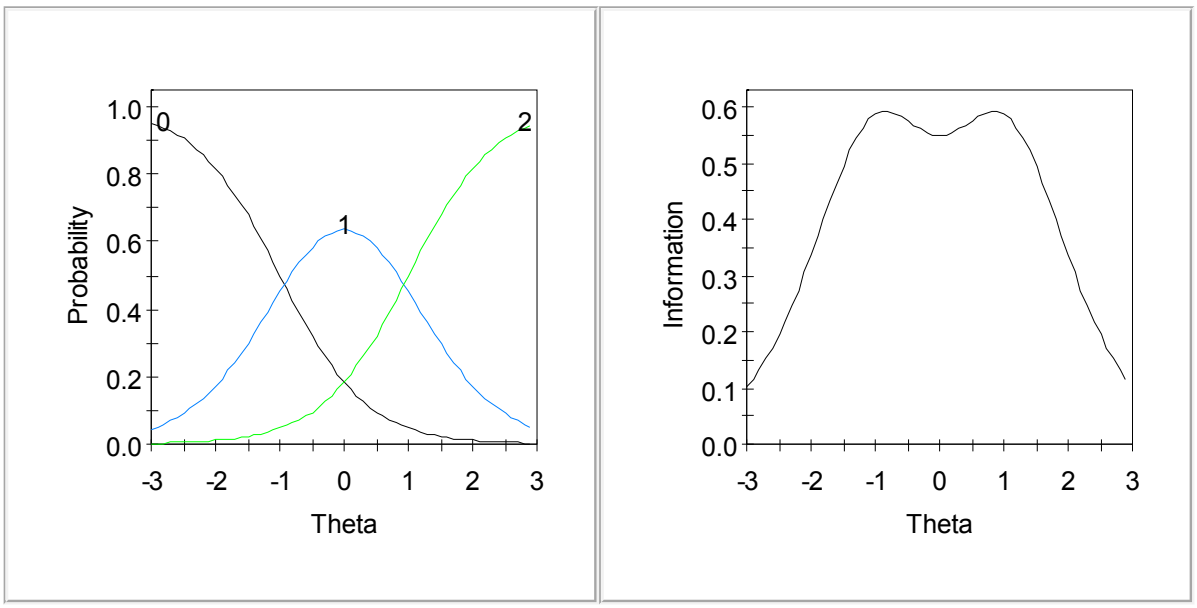



図 2.2.3　項目 2 の IRCCC（左）と項目情報量（右）

（注）$a_2$=1.5, $b_{20}$=-1, $b_{21}$=0, $b_{22}$=1, (D=1)

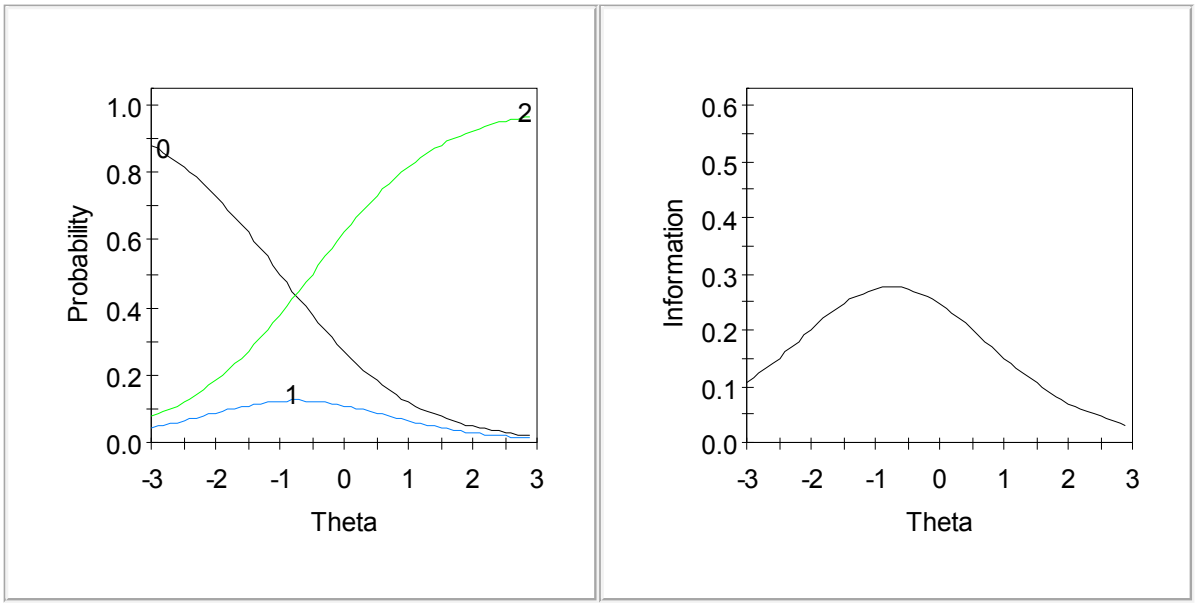



図 2.2.4　項目 3 の IRCCC（左）と項目情報量（右）

（注）$a_3$=1, $b_{30}$=-1, $b_{31}$=-0.75, $b_{32}$=-0.5, (D=1)

### 2.2.5　項目母数の推定

項目反応モデルに含まれる母数のうち，項目の特性を記述する母数をまとめて項目母数と呼ぶことがある。GR モデルの場合，識別力母数 $a_j$，BCC の位置母数 $b^*_{jk}$，IRCCC の位置母数 $b_{jk}$ が項目母数に相当する。本節では，EM アルゴリズムによる周辺最尤推定法（maximum marginal likelihood estimation method, MMLE 法）を用いて GR モデルに含まれる項目母数を推定するための一般的な方法論を紹介する。なお，項目反応モデルの母数推定については，Baker and Kim（2004）が非常に詳しい。

いま, I 名の受検者がそれぞれ K 個のカテゴリーをもつ J 項目のテストを受検したとする。このとき，潜在特性値 θ をもつ受検者が反応パターン m をとる確率は，局所独立の仮定に注意すれば，

$$P(V_m|\theta) = \prod_{j=1}^{J}\prod_{k=1}^{K} P_{jk}(\theta)^{v_{mjk}} \tag{2.2.11}$$

と表現される。ここで，反応パターン行列 $V_m$ は，観測された反応パターン m を表現するための J 行 K 列の大きさをもつ 2 値行列である。その要素は $v_{mjk}$ であり，反応パターン m において項目 j のカテゴリーk が観測されたなら $v_{mjk}$=1，それ以外は $v_{mjk}$=0 である。たとえば，図 2.2.1（右）1 行目の段階反応の場合，反応パターン行列 $V_m$ は図 2.2.5 のように表現される。なお，J 項目がそれぞれ K 個のカテゴリーをもつ状況では，反応パターンの組み合わせの数は $M=K^J$ 個になる。

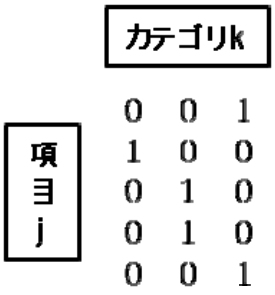



図 2.2.5　反応パターン行列 $V_m$ の例

項目母数を推定する際に局在母数となる θ を消去するため, 潜在特性値の確率密度関数 g(θ)を用いて(2.2.11)式から θ を積分消去すると，反応パターン m が観測される（周辺）確率は,

$$P(V_m) = \int_{-\infty}^{\infty} P(V_m|\theta) g(\theta) d\theta \tag{2.2.12}$$

となる。通常，潜在特性値の確率密度関数 g(θ)には標準正規分布が仮定される。

項目母数を周辺最尤推定するため，テスト結果が得られたもとでの周辺尤度関数を記述し，その尤度関数を最大化するときの項目母数の値を周辺最尤推定値とする。反応パターン m をとる受検者の数を $I_m$ とすると，多項分布(I,P($V_m$))を用いて項目母数の尤度関数は,

$$L = \frac{I!}{\prod_{m=1}^{M} I_m!} \prod_{m=1}^{M} [P(V_m)]^{I_m} \tag{2.2.13}$$

となる。尤度関数の微分の操作を容易にするため，(2.2.13)式の両辺の対数をとると，項目母数の対数周辺尤度関数は,

$$\log L = \log I! - \log \sum_{m=1}^{M} I_m! + \sum_{m=1}^{M} I_m \log P(V_m) \tag{2.2.14}$$

となる。項目母数の周辺最尤推定値は，(2.2.14)式において項目母数についての 1 次偏導関数を 0 とおいた非線形連立方程式（周辺尤度方程式）を数値的に解くことによって得られる。

項目母数を周辺最尤推定する際，(2.2.12)式の積分を計算する必要がある。通常，区分求積法の 1 つである Gauss-Hermite 求積法などを用いて近似計算する。連続値である θ 上において，H 個の離散的な求積点 $X_h$（h=1,2,…,H）とそれらの求積点に対応する重み A($X_h$)を用いて,

$$\tilde{P}(V_m) = \sum_{h=1}^{H} \prod_{j=1}^{J} \prod_{k=1}^{K} [P_{jk}(X_h)]^{v_{mjk}} A(X_h) \tag{2.2.15}$$

と近似できる。

周辺尤度方程式は，EM アルゴリズムを用いて数値的に解くことができる。周辺尤度方程式は，求積点 $X_h$ において項目 j にカテゴリーk と反応する期待頻度 $r_{jkh}$ と求積点 $X_h$ における期待人数 $f_h$ を含む形に変形できる。EM アルゴリズムでは，項目母数の更新量が基準値未満になるなどの収束条件を満たすまで E ステップと M ステップが繰り返される。E ステップにおいて仮の項目母数を定め，

$$r_{jkh} = \sum_{m=1}^{M} \frac{I_m \prod_{j=1}^{J} \prod_{k=1}^{K} [P_{jk}(X_h)]^{v_{mjk}} A(X_h) v_{mjk}}{\bar{P}(V_m)} \tag{2.2.16}$$

$$f_h = \sum_{m=1}^{M} \frac{I_m \prod_{j=1}^{J} \prod_{k=1}^{K} [P_{jk}(X_h)]^{v_{mjk}} A(X_h)}{\bar{P}(V_m)} \tag{2.2.17}$$

を計算する。M ステップでは，Newton-Raphson 法や Fisher のスコアリング法を用いて周辺尤度関数を最大化する。収束条件を満たさないならば，M ステップで更新された項目母数の推定値を仮の項目母数として E ステップに戻る。収束条件を満たしたときの計算結果が最終的な項目母数の推定値となる。

ここまで，テストを構成する J 個の項目がそれぞれ K 個という同数のカテゴリーをもつ場合について記述してきた。同様の方針により，各項目のカテゴリー数が異なる場合や 2PL モデルが混在する場合も項目母数の周辺最尤推定が可能である。各項目のカテゴリー数が異なる場合は，カテゴリー数 K を項目 j に依存する変数 $K_j$ として扱えばよい。2PL モデルが混在する場合でも，カテゴリー数が 2 つの場合の GR モデルは 2PL モデルと同等なので，$K_j$=2 と考えることによって同様の取り扱いが可能である。このとき，図 2.2.5 に示した反応パターン行列 $V_m$ の各行の列数は項目 j によって異なることになる。

なお，推定された項目母数の標準誤差は，推定が終了した時点での Fisher 情報関数行列の逆行列における対角要素の平方根として求められる。また，潜在特性値の推定については，柴山・佐藤・熊谷・佐藤（2011）の第 3 章を参考にしていただきたい。

GR モデルを含む多値 IRT モデルの母数を推定可能なソフトウェアが無料あるいは有料で提供されている。もっとも世界的に利用されているソフトウェアとして，SSI 社（Scientific Software International, Inc.）の PARSCALE 4（Muraki & Bock, 2003）がある。国内で開発されたフリーソフトウェアとしては，Easy Estimation シリーズ（熊谷，2009）がある。また，最近，SSI 社から IRTPRO 2.1（Cai, Thissen, & du Toit, 2011）がリリースされている。

### 2.2.6 GR モデルの情報関数

テストを実施するということは，テストを構成する項目を利用して受検者の潜在特性値 θ についての情報を得る行為であると解釈できる。その際，各項目を通して得られる θ に関する Fisher 情報量を

項目情報関数（item information function）と呼び，項目 j の項目情報関数 $I_j(\theta)$は，

$$I_j(\theta) = -E\left[\frac{\partial^2 \log P_{jk}(\theta)}{\partial \theta^2}\right] = \sum_{k=0}^{K-1}\left\{-\frac{\partial^2 \log P_{jk}(\theta)}{\partial \theta^2}\right\} P_{jk}(\theta) = \sum_{k=0}^{K-1} I_{jk}(\theta) P_{jk}(\theta) \tag{2.2.18}$$

と表現できる。Samejima（1969）は，(2.2.18)式の $I_{jk}(\theta)$を項目カテゴリーの情報関数，$I_{jk}(\theta)P_{jk}(\theta)$をカテゴリーk の情報量占有率と呼んだ。GR モデルでは，

$$I_{jk}(\theta) = -\frac{\partial^2 \log P_{jk}(\theta)}{\partial \theta^2} = \frac{[P'_{jk}(\theta)]^2 - P_{jk}(\theta)P''_{jk}(\theta)}{[P_{jk}(\theta)]^2} \tag{2.2.19}$$

$$I_{jk}(\theta)P_{jk}(\theta) = \frac{[P'_{jk}(\theta)]^2 - P_{jk}(\theta)P''_{jk}(\theta)}{[P_{jk}(\theta)]^2} P_{jk}(\theta) = \frac{[P'_{jk}(\theta)]^2}{P_{jk}(\theta)} - P''_{jk}(\theta) \tag{2.2.20}$$

となる。ただし，$P_{jk}'(\theta)=\partial P_{jk}(\theta)/\partial\theta$, $P_{jk}''(\theta)=\partial^2 P_{jk}(\theta)/\partial\theta^2$ である。(2.2.20)式を(2.2.18)式に代入して整理すれば，GR モデルの項目情報関数 $I_j(\theta)$は，

$$I_j(\theta) = \sum_{k=0}^{K-1}\left\{\frac{[P'_{jk}(\theta)]^2}{P_{jk}(\theta)} - P''_{jk}(\theta)\right\} \tag{2.2.21}$$

$$= \sum_{k=0}^{K-1}\frac{[P'_{jk}(\theta)]^2}{P_{jk}(\theta)} - \sum_{k=0}^{K-1} P''_{jk}(\theta) \tag{2.2.22}$$

$$= \sum_{k=0}^{K-1}\frac{[P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta)]^2}{P^*_{jk}(\theta) - P^*_{jk+1}(\theta)} - \sum_{k=0}^{K-1}[P^{*\prime\prime}_{jk}(\theta) - P^{*\prime\prime}_{jk+1}(\theta)] \tag{2.2.23}$$

$$= \sum_{k=0}^{K-1}\frac{[P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta)]^2}{P^*_{jk}(\theta) - P^*_{jk+1}(\theta)} \tag{2.2.24}$$

となる。なお，(2.2.22)式から(2.2.23)式への計算には(2.2.1)式を利用している。また，(2.2.23)式の第 2 項は，k=1,2,…,K-1 の項は消去されること及び(2.2.3)式と(2.2.4)式から 0 である。

GR モデルの BCC として(2.2.2)式の 2PL モデルを利用すれば，

$$P'_{jk}(\theta) = P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta) = Da_jP^*_{jk}(\theta)Q^*_{jk}(\theta) - Da_jP^*_{jk+1}(\theta)Q^*_{jk+1}(\theta) \tag{2.2.25}$$

と計算できる。ただし，$Q^*_{jk}(\theta)=1-P^*_{jk}(\theta)$である。このとき，項目情報関数 $I_j(\theta)$は，

$$I_j(\theta) = D^2a_j^2\sum_{k=0}^{K-1}\frac{[P^*_{jk}(\theta)Q^*_{jk}(\theta) - P^*_{jk+1}(\theta)Q^*_{jk+1}(\theta)]^2}{P_{jk}(\theta)} \tag{2.2.26}$$

となる。

図 2.2.2（右）～図 2.2.4（右）に，(2.2.26)式によって描いた項目情報関数の例を示す。例示した 3 つの項目は，2.2.3 節で利用した項目と同一である。図をみると，IRCCC の位置母数の近辺で情報量が大きいことや，その近辺での情報量は識別力母数の値が大きい項目のほうが大きいことがわかる。情報量の逆数が測定誤差の大きさと関係することから，テストを作成するときには，目的とする測定レベルに見合った難易度（位置母数）の項目を用意することや，なるべく識別力の高い項目を用意することが大切である。

項目が 2 つのカテゴリーだけをもつ場合，(2.2.26)式の項目情報関数は 2PL モデルのそれと一致する（豊田，2005，p. 84）。項目情報量の点からも，2PL モデルは 2 つのカテゴリーをもつ GR モデルと同等であることが確認できる。さらに，Samejima（1969）によれば，ある項目にカテゴリーを追加した場合，追加する以前と比べて同等かそれ以上の項目情報量が得られるということが証明されている。

Fisher 情報量の加法性から，局所独立の仮定を満たすテストの項目情報量をすべて加算するとテスト全体の情報量に相当する。項目情報量の単純和はテスト情報量と呼ばれ，潜在特性値 θ に関するテスト情報量はテスト情報関数（test information function）と呼ばれる。$P^{*}_{jk}(\theta)$として(2.2.2)式の 2PL モデルを利用すれば，テスト情報関数 I(θ)は(2.2.26)式の項目についての和となり，

$$I(\theta)=\sum_{j=1}^{J} I_j(\theta)=\sum_{j=1}^{J}\sum_{k=0}^{K-1} D^2 a_j^2 \frac{[P^{*}_{jk}(\theta)Q^{*}_{jk}(\theta)-P^{*}_{jk+1}(\theta)Q^{*}_{jk+1}(\theta)]^2}{P_{jk}(\theta)} \tag{2.2.27}$$

と表現できる。

テスト情報量は，テストの測定精度と密接な関連がある。テスト情報量を用いると，潜在特性値の最尤推定値 $\hat{\theta}$ の標準誤差を $1/\sqrt{I(\hat{\theta})}$ として見積もることができる。それゆえ，テスト情報関数をみれば，テストがどの付近の潜在特性値をどのくらい正確に測定できるのかがわかる。テスト情報量の大きい尺度値レベルが潜在特性値をより正確に測定できる部分であり，テスト情報量の小さい尺度値レベルが潜在特性値の測定精度が低くなる部分である。

図 2.2.6 に，図 2.2.2～図 2.2.4 の 3 項目からテストが構成されている場合のテスト情報関数 I(θ)と最尤推定値 $\hat{\theta}$ の標準誤差に相当する $1/\sqrt{I(\theta)}$ の曲線を示す。当該テストは，その受検者集団において平均的な尺度値レベルをもつ受検者の潜在特性値を他の尺度値より正確に測定できることが読み取れる。

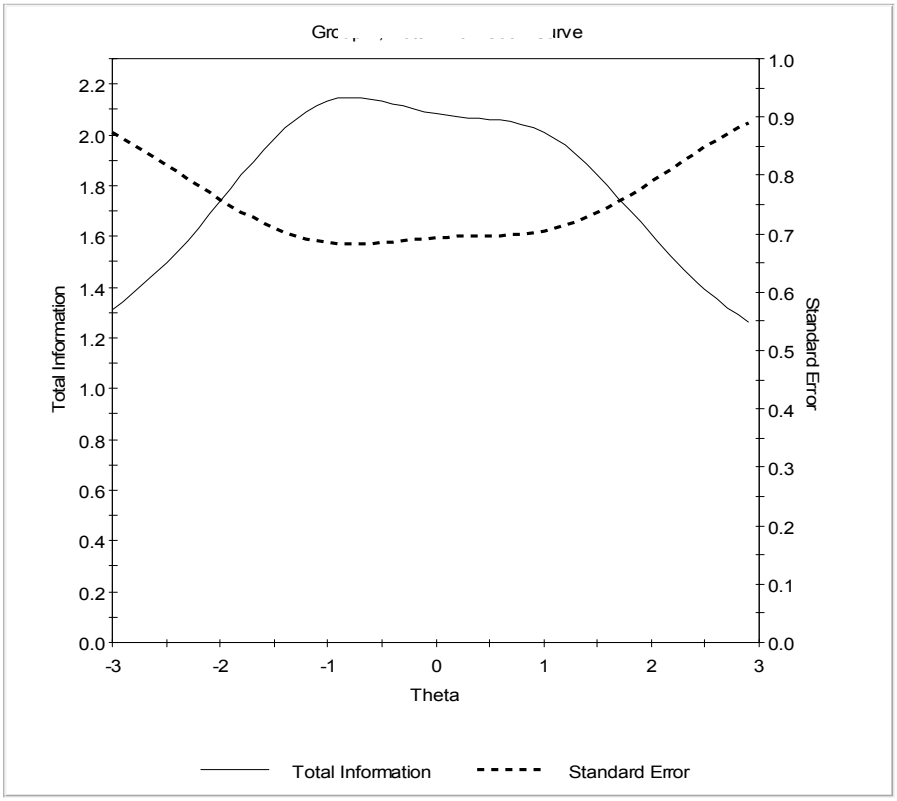



図 2.2.6 テスト情報量と標準誤差

## 2.3　リーディング・リテラシーについての展望

リーディング・リテラシー(reading literacy)という用語は，経済協力開発機構(OECD)が行っている 15 歳生徒の国際学習到達度調査 PISA(Programme for International Student Assessment)で使用されている用語で，数学的リテラシー(mathematical literacy)・科学的リテラシー(scientific literacy)とともにこの調査の 1 つの領域を指す。リテラシーとは本来「識字」あるいは「読み書き能力」のことを指すが，最近は多様に用いられている。PISA においては，リテラシーを，学校カリキュラムの習得ではなく，知識や技能を，実生活の様々な場面で直面する課題に活用する能力という意味でとらえている。リーディング・リテラシーは最初「読解力」と訳されていたが（国立教育政策研究所, 2002），これが我が国の国語科教育で従来用いられてきた「読解力」と意味範囲が異なることが明らかとなり，「リーディング・リテラシー」「読解リテラシー」「PISA 型読解力」などとして，従来の「読解力」と区別するようになってきた。

そこで本節では，このリーディング・リテラシーがどのような背景のもとに生まれてきたのか，その内容はどのようなものか，そしてこれが我が国の国語科教育にどのような影響を与えたのかを，述べることとする。

### 2.3.1　リーディング・リテラシーの背景

#### (1) 20 世紀末の状況と多様なリテラシー

20 世紀の末，交通網の発達と情報通信の発達は，コミュニケーションや社会の在り方を変えた。人々の移動や情報の伝達は盛んになり，社会のグローバル化が引き起こされた。以前は空間に制限を受けていたコミュニティーも，変化を見せるようになった。リテラシーという用語もこのことを反映して，多様な言葉と結び付けて，多様な概念内容を表現すようになった。例えば，メディア・リテラシー，コンピュータ・リテラシー，文化リテラシーなどという言葉に私たちは馴染んでいる。ここでは，特に，マルチリテラシーズ(Multiliteracies)という概念の背景を見てみる。マルチリテラシーズとは，リテラシーが多様であるという意味である。しかしながら，これは，単にリテラシーを多様に使いこなせばよいというものではない。なぜならば，１つのリテラシーは，１つの文化・社会・価値観の安定した基盤に基づいているからである。マルチリテラシーズ社会における生活は，より複雑な生活になる。表 2.3.1 は，マルチリテラシーズという概念を唱えたニューロンドングループが，マルチリテラシーズ社会の背景にある変化を，労働生活・公共生活・個人生活の 3 局面に分けてとらえたものである。ポスト・フォード主義とは，自動車のフォード社に代表されるような画一生産・大量消費の時代が変わり，生産的多様性が望まれてきていることを示している。すなわち，これまで画一的だった生産は多様化し，１つのコミュニティーに所属しているという市民概念が低下して複数のコミュニティーに属しているような複雑な形になり，これまで確保されてきた個人のプライベート空間の中に，外の世界のものが侵略してきて，多層の生活世界を生きるようになるというのである。多様な文化・社会・価値観の中で生きていかなければならない。そのための力をマルチリテラシーズと呼んだのである。

表 2.3.1　マルチリテラシーズの背景となる社会の変化

| | 変化している現実 | 社会の未来像 |
|---|---|---|
| 労働生活 | 加速する資本主義／ポスト・フォード主義 | 生産的多様性 |
| 公共生活 | 市民概念の低下 | 市民的複雑性 |
| 個人生活 | プライベート空間の侵略 | 多層の生活世界 |

### (2) キー・コンピテンシー

このように変化する社会は，経済活動のグローバル化も引き起こす。経済活動も複数の国々にまたがって行われる傾向がますます強くなってきた。そして，欧州諸国，アメリカ合衆国，日本などの 30 か国の先進工業国を中心に，経済成長，開発途上国援助及び自由かつ多角的な貿易の拡大を目的とする経済に関する国際協力機関である OECD は,経済活動の円滑な進展のためには教育の国際指標の作成が重要であることに気付いており，1988 年から教育インディケータ事業を実施してきた。PISA はそのインディケータ（指標）の１つとして開発されたもので，義務教育終了段階の生徒がそれまで身に付けてきた知識や技能を，実生活の様々な場面で直面する課題にどの程度活用できるかを国際的に測るための調査である。

PISA の 2000 年調査の後に, PISA などの調査の概念的基盤を提供するためにまとめられた DeSeCo のキー・コンピテンシーの定義は，先に述べたマルチリテラシーズの考え方と非常に類似している。DeSeCo とは，「コンピテンシーの定義と選択：その理論的・概念的基礎」(Definition & Selection of Competencies; Theoretical & Conceptual Foundations)というプロジェクトで, OECD とも関係するが，スイス連邦主導で実行されたものである。時系列としては PISA の 2000 年の調査後に出されたものではあるが，PISA の背景にある考え方をよく表現していると考えられているので，リーディング・リテラシーの背景として，ここで検討してみたい。キー・コンピテンシーとは，直訳すると「重要な能力」という意味であり，単なる知識や技能ではなく，学習への意欲や行動・行為に至るまでの幅広く深い能力のことである。DeSeCo は，キー・コンピンテンシーとして，次の９つのコンピテンシーを設定し，それをそれぞれ３つずつにカテゴリー化した。

カテゴリー１　相互作用的に道具を用いる
　コンピテンシー１Ａ：言語，シンボル，テクストを相互作用的に用いる能力
　コンピテンシー１Ｂ：知識や情報を相互作用的に用いる能力
　コンピテンシー１Ｃ：技術を相互作用的に用いる能力
カテゴリー２　異質な集団で交流する
　コンピテンシー２Ａ：他人といい関係を作る能力
　コンピテンシー２Ｂ：協力する能力
　コンピテンシー２Ｃ：争いを処理し，解決する能力
カテゴリー３　自律的に活動する
　コンピテンシー３Ａ：大きな展望の中で活動する能力
　コンピテンシー３Ｂ：人生計画や個人的プロジェクトを設計し実行する能力
　コンピテンシー３Ｃ：自らの権利，利害，限界やニーズを表明する能力

カテゴリー１の「道具」とは，コンピュータのような物理的な道具だけでなく，言語，情報，知識などの相互作用のための社会文化的な道具をも含む。特にコンピテンシー１Ａの「言語，シンボル，テクストを相互作用的に用いる能力」とは,「様々な状況において，話して書くといった言語的なスキル，コンピュータまたは図表を用いるといった他の数学的なスキルを有効に利用するものである」(Rychen & Salganik, 邦訳 2006, p.211)とされており，PISA のリーディング・リテラシー，数学的リテラシーは，このコンピテンシーとの関係が強い。

カテゴリー２の「異質な集団」は，先のマルチリテラシーズ社会の変化としてみた「市民的複雑性」と類似している。近代国家は，等質の集団でコミュニティーを作り上げたものであった。しかし，グローバル化が進む今日では，様々な人々が移動し，交錯する。異質な人々ともコミュニケーションを図り，コミュニティーを作っていかなければならない。時には，文化的摩擦があるかも知れない。その中で，コンピテンシー２Ｂ「協力する能力」やコンピテンシー２Ｃの中の「争いを処理し，解決する能力」は，結局異質な人々とのコミュニケーションの能力とも言えるであろう。

カテゴリー３「自律的に活動する」が必要な理由の１つとして,「複雑な社会で自分のアイデンティティを実現し，目標を設定する」が挙げられている。やはりマルチリテラシーズの背景としてみた「市民的複雑性」に加えて「多層の生活世界」との類似性を指摘することができる。カテゴリー３の各能力を見ると，異質な集団である複雑な社会に身を投じながら，見通しを持ち自律的に活動し，必要なことをそのコミュニティーの人々に表明していくという人間像が見えてくる。

PISA のリーディング・リテラシーは，このような 20 世紀の末の，複雑化する社会背景のもとに作られた調査の領域である。その内容については後述するが，ここでは先回りして，我が国の国語科教育における従来の「読解力」との違いを，キー・コンピテンシーに即して３点指摘しておく。１点目は，読まれるテキストについてである。リーディング・リテラシーでは,「非連続型テキスト」をはじめとして，多様なジャンルや形式のテキスト使用して調査している。これは，このような複雑な社会に参加して生きていく時に読む必要がある多様なテキストを読む時の力を,「テクストを相互作用的に用いる」能力として，測定しようとしているからである。２点目は，PISA のリーディング・リテラシーの調査問題のいくつかは，比較的長く記述することを含んでいることである。単に書かれたことを自分が理解できればよいということではなくて，自分はこのように理解したということを，異質な他者に向かってしっかりと表明しなければならない。これは，カテゴリー２の異質な集団やコンピテンシー３Ｃの「表明する」が関係していることが推測できる。３点目は，内容を理解することだけでなく,「熟考・評価」と言って，そのテキストが自分にとってどのような意味があるかを，自分の先行知識との関係から評価・判断させるような側面を調査していることである。これは，カテゴリー１に示されているように言語やテキストを「道具」としてとらえ，それを「相互作用的」に用いることや，コンピテンシー２Ｂの「争いを処理し，解決する」などとの関連が推測できる。PISA のリーディング・リテラシーが，文学作品の登場人物に感情移入して読んだり，説明的文章の内容を正確に理解したりすることにとどまらないのは，このような背景による。

### 2.3.2　リーディング・リテラシーの内容

PISA は，義務教育が終了してこれから社会に参加していくと想定される 15 歳の生徒達に対して行われる国際学習到達度調査で，2000 年，2003 年，2006 年，2009 年と 3 年ごとに調査が行われてい

る。リーディング・リテラシー，数学的リテラシー，科学的リテラシーの 3 領域が毎回調査されるが，毎回 1 つの領域を中心領域と決めて調査や分析を詳しくしている。リーディング・リテラシーが中心領域だったのは，2000 年と 2009 年である。したがって，2000 年の時に作られたリーディング・リテラシーの枠組みや調査問題は，そのまま 2003 年，2006 年の調査でも使用された。次に 2009 年にふたたび中心領域になった時に，その枠組みや調査問題が改訂された。したがって，リーディング・リテラシーの内容というと 2000 年のものと 2009 年のものでは，多少異なる点がある。ここでは，我が国の国語科教育への影響の観点から，2000 年の内容を詳しく示し，2009 年については，その変更点を指摘するにとどめる。

### (1) 定義

2000 年のリーディング・リテラシーの定義は，「読解力とは，自らの目標を達成し，自らの知識と可能性を発達させ，効果的に社会に参加するために，書かれたテキストを理解し，利用し，熟考する能力である。」である。「自らの目標を達成し」「社会に参加する」などの表現が，キー・コンピテンシーに生かされていることが分かる。

### (2) 状況

リーディング・リテラシーでは，変化する複雑な社会に参加していく時に必要な読むことを調査している。したがって，どのような状況で読むのかということが，重要になってくる。具体的には，私的，公的，教育的，職業的の 4 つの状況が設定された。これらは，テキストの用途に直結している。私的な用途のテキストには，余暇や気晴らしのために読む小説や伝記，手紙や私的な電子メール，日記形式のブログなどが含まれる。公的な用途のテキストには，公的行事に関する情報や公的文書，さらには，報道のウェブサイトなどが含まれる。教育的な用途に分類されるテキストには，教科書や教材のように，教育目的で設計されているものが含まれる。15 歳は義務教育を終了しているが，実際には高校などの教育機関で教育を受けている者が多いからである。職業的な用途には，履歴書，求人広告，職場の指示に従うことなどが含まれる。

### (3) テキスト

PISA 調査で使用されるテキストには，文章や段落から構成されている連続型テキスト(Continuous texts)と非連続型テキスト(Non-continuous texts)の 2 つがある。

連続型テキストには，物語(Narration)，解説(Exposition)，記述(Description)，議論(Argumentation)，指示(Instruction)，文書または記録(Documents or records)，ハイパーテキスト(Hypertext)がある。

非連続型テキストには，図・グラフ(Charts and graphs)，表(Tables)，図(Diagrams)，地図(Maps)，書式(Forms)，情報シート(Information sheets)，宣伝・広告(Calls and advertisements)，バウチャー(Vauchers)，証明書(Certificates)がある。

このテキストの種類は，ある程度用途と結びつくものと，様々な用途で広く使用されているものの両方がある。いずれにしても，かなり多様なテキストを含んでいることが分かる。

### (4) 側面

PISA 調査では，読む状況やテキストの種類に限らず，リーディング・リテラシーを 5 つの側面と

してとらえている。その５つとは，

１情報の取り出し
２幅広い一般的な理解の形成
３解釈の展開
４テキストの内容の熟考・評価
５テキストの形式の熟考・評価

である。１・２・３は「基本的にテキスト内部の情報を利用する」ものである。設問に解答する際に，テキストの中に解答が書かれているというものである。解答が，一文以内の比較的短い箇所で独立して取り出せるものは「１情報の取り出し」であり，テキストの全体を見なければならないような場合が「２幅広い一般的な理解」，部分的に複数の箇所と関係させなければならないような場合が「３解釈の展開」である。４・５は「外部の知識を引き出す」ものである。外部の知識とは，テキストの内部にないことであり，調査に取り組んでいる生徒にとってみれば自分の既有知識や考え方や価値観などである。つまり，テキストの中に書いてあることだけでは解答できず，自分の既有知識や考え方や価値観などに照らし合わせて解答していくことになる。つまり，中身の理解だけでなく，自分なりの判断や評価といったものが行われることになる。「４テキストの内容の熟考・評価」は，その意見についてどう思うかといったテキストの内容面に焦点をあてるもの，「５テキストの形式の熟考・評価」はその表現の仕方（形式）はよいかどうかといった，形式面に焦点をあてるものである。分析時には，１を単独で「情報の取り出し」，２と３をまとめて「解釈」，４と５をまとめて「熟考・評価」と，大きく３つに分類した。本稿では，これをリーディング・リテラシーの３側面と呼ぶことにする。このことが，後で，我が国の国語科教育に大きく影響することになる。

### (5) 質問紙調査

いわゆるテストにあたる部分のほかに，生徒質問紙と学校質問紙の 2 種類の質問紙調査が行われた。生徒質問紙では，中心領域にあたっている 2000 年・2009 年は，リーディング・リテラシーに関係すると考えられる学習の環境や，読書の状態，国語の授業など多様な質問項目があり，生徒のこれらの環境と読解力との関係を調査された。

2.3.2.6　2009 年調査での変更点

PISA 調査は OECD の教育インディケータ事業であるので，指標の経年比較という意味でも根本的な変更というものは認められない。したがって，PISA2009 年のリーディング・リテラシーの枠組みも基本的には PISA2000 年の枠組みを踏襲している。しかし，10 年近い経過を経て，2009 年の調査では，次の 2 点の変更点があった。

１点目は「読みの取り組み」(reading engagement)が強調されていることである。読む力をつけるにあたって，実際に読むことに取り組むことが重要である。このためには，読書意欲を持ち，自分の読書行動についてメタ認知ができる能力を持っていて，自分の読書行動を考えながら読むことに取り組む必要がある。このため 2009 年の定義は，「読解力とは，自らの目標を達成し，自らの知識と可能性を発達させ，効果的に社会に参加するために，書かれたテキストを理解し，利用し，熟考し，これに取り組む能力である。」（下線引用者）となった。下線の部分が PISA2000 年調査の定義と PISA2009

年の調査の定義が異なる部分であり，これが「読みの取り組み」にあたる。つまり，PISA2009 年の定義では，前の定義とほとんど同じであるが「読みの取り組み」が付け加えられたということである。これにともなって，質問紙調査では，読書の実際の取り組みを把握できるように，メタ認知に関する項目が加えられた。

2 点目は，電子テキストの読解（デジタル読解力）を取り入れた点である。現在は， 2000 年の調査時に比べて益々，コンピュータなどのメディアを通してウェブサイトなどを読むことが増えてきた。そこで，PISA2009 年調査では，紙のテストの他に，コンピュータを使用したテストが実施された。この変更に伴って，テキストも連続型テキスト，非連続型テキストという分類のほかに，混声型テキスト，複合型テキストが加えられた。リーディング・リテラシーの３側面については，「情報の取り出し」が「情報へのアクセス・取り出し」，「解釈」は「解釈・統合」に文言が修正されたが，「熟考・評価」はそのまま「熟考・評価」となっている。

### 2.3.3　PISA 調査の結果と我が国の国語科教育への影響

#### (1) PISA 調査の結果と国語科教育での取り上げられ方

PISA 調査における我が国の生徒のリーディング・リテラシーの結果は，2000 年が 8 位（31 か国中），2003 年が 14 位（41 か国・地域中），2006 年が 15 位(57 か国・地域中)，2009 年が 8 位（65 か国・地域中）あった。毎回参加国は増加しており，単純な比較はできないが，転機になったのは，2003 年調査であった。なぜなら，2000 年の結果は 8 位とは言え，統計的には 2 位と同じグループにいるとされていたのに対し，2003 年調査は，14 位であったからである。マスコミなどでも「学力の低下」として取り上げられた。このことが，我が国の教育界へ多大な影響を与えた。

国語科教育から見れば，PISA は初めて経験する国際学力調査であり，初めて外国と比較した我が国の生徒たちの読むことに関する能力の特徴が明らかになったわけである。主に注目されたのは次の４点である。１点目は，３側面のうち「熟考・評価」の成績の低さである。２点目は無答率の高さである。実は，１点目と２点目は大いに関係しており，３側面のうち「熟考・評価」の無答率が高かった。３点目は，PISA の非連続型テキストを中心としたテキストの多様さに比べた，我が国の国語科教育で扱われる教材の性質である。これまでは文学的文章と説明的文章といった文章のみが指導で扱われていた。４点目は，質問紙の回答で，読書に取り組む生徒の少なさである。特に「趣味で読書をすることはない」と回答した生徒の数が多かった。

これらの４点が国語科教育の中でどのように取り上げられたかについて示す。まず，１点目と２点目については，我が国の国語科の中では，テキストの内容に書かれている事柄の理解に焦点があたっており，「熟考・評価」が指導されてこなかったということに議論が集まった。生徒自身も「熟考・評価」の設問に慣れておらず，これが無答率の高さにつながったという考え方である。そのため，「熟考・評価」を行わせる授業が「PISA 型読解力に対応した授業」としてよく実践され，国語科教育に関係する商業誌もこのような「PISA 型読解力」を特集することが多くなった。３点目については，非連続型テキストを取り入れることが議論された。これまで国語科では，図やグラフの読み取りは数学教育の問題と考えて，扱ってこなかった。ところが，PISA2003 の結果発表後，国語教科書にも全国学力学習状況調査にも，非連続型テキストが取り入れられるようになった。４点目については，読書の奨励である。国語科の時間だけでなく，朝読書や読書旬間の実施など，様々な機会に読書が奨励され

た。

### (2) 平成 20 年中学校学習指導要領「国語」

2003 年の PISA 調査結果発表（2004 年 12 月）を受け，2005 年に文部科学省は「『読解力』向上に関する指導資料を出した。そこでは，次のような「基本的な考え方」が述べられている。

ア　PISA 調査のねらいとするところは，現行学習指導要領で子どもに身に付けさせたいと考えている資質・能力と相通じるものであることから，学習指導要領のねらいとするところの徹底が重要である。

イ　PISA 調査の結果から明らかになったことと，教育課程実施状況調査の結果とには共通点があることから，教育課程実施状況調査の結果を受けた改善の提言も併せて指導の改善に生かすことが重要である。

ウ　読解力は，国語だけではなく，各教科，総合的な学習の時間など学校の教育活動全体で身に付けていくべきものであり，教科等の枠を超えた共通理解と取組の推進が重要である。

そして，2007 年より，全国学力・学習状況調査が開始されることになる。

2008 年（平成 20 年）3 月に告示された中学校学習指導要領は，上記のような基本方針を踏まえていると考えられる。学習指導要領は，我が国の教育の様々なことを考慮して作成されているが，そのうち次の点については，PISA の影響があったと言える。

① 言語活動の充実（国語科に限らず他の教科でも言語活動の充実が強調された。）

② ［伝統的言語文化と国語の特質に関する事項］の新設（これまで「言語事項」と呼ばれていた事項に古典などの言語文化を加えた事項。小学校から我が国独自の言語文化に積極的に触れさせるようになった。リーディング・リテラシーのように国際的な力をつけていくことも重要であるが，我が国独自の言語文化を継承していくことも我が国の生徒には必要な国語科学習の内容である。）

③ 内容「C 読むこと」の枠組みの明示（「語句の意味の理解」「文章の解釈」「自分の考えの形成」「読書と情報活用」。これは，前者３つがリーディング・リテラシーの３側面「１情報の取り出し」「２解釈」「３熟考・評価」と関係していると考えられる。）

④ 読書活動の充実（特に，これまで内容には入っていなかった読書に関する項目が，C 読むことの「読書と情報活用」の項目として取り上げられた。）

### (3) リーディング・リテラシーと本調査

以上述べてきたようなリーディング・リテラシーは，社会のグローバル化や情報化が進展する中で，益々重要になってくると考えられる。そこで，本研究「学力調査を活用した専門的課題分析に関する調査研究」は，このことを重視し，リーディング・リテラシーを意識した調査を行っている。本調査の定義については，「定義１情報の取り出し」をリーディング・リテラシーの３側面の「１情報の取り出し」に，「定義２解釈」を３側面の「２解釈」に，「定義３情報の編集・統合」「定義４判断」を３側面の「３熟考・評価」にそれぞれ関連させてある。一方で，本調査は，国語科の学習内容としての力を調査するので，平成 20 年の中学校学習指導要領（国語）の内容における項目との対応関係も明示している。さらに，調査問題には，生徒が社会に参加する時に読むようなテキストを使用するようにした。