資料シリーズ
No.104

NIVR

# 職業リハビリテーションのアセスメントにおける現代テスト理論の応用可能性に関する基礎的研究

2022 年 3 月

独立行政法人高齢・障害・求職者雇用支援機構
障害者職業総合センター
NATIONAL INSTITUTE OF VOCATIONAL REHABILITATION

# まえがき

障害者職業総合センターでは、「障害者の雇用の促進等に関する法律」に基づき、我が国における職業リハビリテーションの中核機関として、職業リハビリテーションに関する調査・研究をはじめとして、様々な業務に取り組んでいます。

この資料シリーズは、当センターの研究部門が 2021 年度に実施した「職業リハビリテーションのアセスメントにおける現代テスト理論の応用可能性に関する基礎的研究」の結果をまとめたものです。

本調査研究では、職業リハビリテーションにおける職業評価や障害特性の評価に用いられるアセスメントツールを概観するとともに、多様化する障害者のアセスメントに効果的に利用できるツールの設計・開発に必要な基礎的情報と、今後の研究開発に向け、現代テスト理論の応用方法について提示することを目的としました。

本書が、我が国における職業リハビリテーションを更に前進させるための一助になれば幸いです。

最後に、本調査研究の実施に当たり、専門家の先生方にご協力を賜りました。ここに、厚く御礼申し上げます。

2022 年３月

独立行政法人高齢・障害・求職者雇用支援機構
障害者職業総合センター
研究主幹　成田　裕紀

執筆担当者（執筆順）

| | | | |
|---|---|---|---|
| 知名　青子 | 障害者職業総合センター | 研究員 | 概要、序章、第１章、第２章第３節、第３章第１節～第２節、おわりに |
| 國東　菜美野 | 元障害者職業総合センター | 研究員 | 第２章第１節～第２節 |
| 清水　求 | 障害者職業総合センター | 研究協力員 | 第２章第２節、第３章第１節、資料 |

謝　辞

本調査研究の実施に当たり、専門家の先生方（筑波大学 八重田淳先生、元九州看護福祉大学 吉光清先生、佐賀大学 西郡大先生、九州大学 木村拓也先生、東北大学 熊谷龍一先生）にヒアリング調査へのご協力をいただき、興味深い資料やご助言を得ることができました。ご協力いただいた皆様に、心より御礼申し上げます。

＜研究担当者＞

本調査研究は、令和３年度に障害者職業総合センター障害者支援部門で担当した。
研究担当者、研究担当時の職名は下記のとおりである。

| | | |
|---|---|---|
| 山科　正寿 | 障害者職業総合センター | 主任研究員 |
| 知名　青子 | 障害者職業総合センター | 研究員 |
| 國東　菜美野 | 元障害者職業総合センター | 研究員 |
| 清水　求 | 障害者職業総合センター | 研究協力員 |

# 目　次

# 概　要

昨今の職業リハビリテーションサービスの対象者における状態像の多様化に伴い、職業評価やアセスメントの難易度は着実に高まっている。

本調査研究では、障害者の支援に携わる者が利用するアセスメントツールに関する現状の問題を取り上げ、将来的なツール開発に当たっての検討事項を提示した。また、テスト理論の考え方を解説した上で、現代テスト理論の職業リハビリテーション領域におけるアセスメントツールへの応用に向け、ワークサンプル幕張版の新規課題である『社内郵便物仕分』課題を参考とした探索的なデータシミュレーションと解析を行い、その応用方法を提示した。

第１章では、職業リハビリテーションにおけるアセスメントツールの現状の課題を提示した。障害を評価するための尺度や検査は多数存在するが、職業リハビリテーションの領域では、特に作業遂行面や機能評価のために神経心理学的検査やワークサンプルなどが積極的に利用されている。特に評価の難しい認知特性や対人態度等の行動特徴に関しては、ワークサンプル幕張版や模擬的就業場面等における環境条件を含めた評価方法が主となっていることを取り上げた。また、障害の多様化・複雑化に伴うアセスメント自体の難化に対して、ツールの精度向上の必要性について指摘した。加えて、アセスメントツールを開発・維持・運用していく上で議論すべき点をあげた。

職業リハビリテーションの領域では、より効率的、効果的な支援技法の開発が求められるところだが、最近では、計算機の発達により、膨大なデータを利用して AI 等に将来や予後を推測させる取組や、適応型テストのような本人の能力に応じた出題をするといった技術が実装され始めている。このような大規模なデータを扱う統計的な手法は、情報処理の分野を筆頭に、心理測定や行動計量の分野において採用されている。本調査研究ではそのような統計的手法の中でも、現代テスト理論における項目反応理論を用いて、職業リハビリテーション領域の評価方法について検討が必要であることを指摘した。

第２章においては、テスト理論そのものの基本的な理解を目指し、古典的テスト理論に対して、現代的テスト理論がどのような考え方をもつ統計理論であるのかについて、初学者でもわかりやすいように説明した。

また、第３章ではワークサンプル幕張版のうち、実務課題として提供されている『社内郵便物仕分（簡易版)』を対象として、一般参考値に関するシミュレーションデータを作成し、そのデータを利用した項目反応理論の実施による項目分析のプロセスについて提示した。これにより、職業リハビリテーションの領域で利用するアセスメントツールやテストの品質向上のための現代テスト理論の利用可能性と、利用に当たっての留意点や課題を提示した。

# 序章（研究の問題意識）

## １　最近の職業リハビリテーションの利用者像

厚生労働省による障害者雇用実態調査（2018 年（平成 30 年）６月）の結果によると、精神障害者の推計雇用数は約 20 万人とされ、その人数は増加傾向にあることがうかがえる。また、同調査では発達障害者が新たに把握されており、推計人数は約３万９千人である。

就職活動を行う精神障害者と発達障害者の各々の実態については、障害者職業総合センター（2010, 2011, 2021）が全国のハローワークの専門援助窓口に対して実施した実態調査の結果にその詳細をみることができる。これまでの調査を概観すると、直近は 2018 年調査（対象 417 所、調査期間:2018 年６月１日から 30 日の１ヶ月間）であり、当時、新規求職申込みを行った障害者は 4,962 人であり、このうち、精神障害者は 2,352 人、発達障害者は 641 人と報告されていた。精神障害者については、同様の調査枠組みで 2008 年調査（対象 110 所、調査期間：2008 年７月１日～10 月 31 日の４か月間)が実施されており、当時 1,808 人の新規求職申込みがあったことが報告されていた。また、発達障害者についても同様の調査枠組みにおいて 2011 年調査（対象 109 所、調査期間：2009 年４月１日～2010 年１月 31 日の 10 か月間）が実施されており、538 人の新規求職申込みが報告されていた。

これら調査結果から１所１ヶ月あたりの新規求職申込み人数を算出すると、精神障害者では 2008 年の 4.1 人から 2018 年の 5.6 人へ、発達障害者では 2009－2010 年の 0.5 人から 2020 年の 1.5 人へといずれも増加している。10 年での増加率としては、それぞれ精神障害者で 36%、発達障害者で 200%である。この状況に伴って、地域障害者職業センター（以下「地域センター」という。）や障害者就業・生活支援センター等の職業リハビリテーションサービスや、就労移行支援事業所等の障害者就労支援機関における精神障害者、発達障害者の利用ニーズは着実な高まりを見せていると考えられる。

一方で、精神障害者、発達障害者の就職・復職を目指す上では、様々な課題が指摘されてきた。障害特性や病状の把握において、的確なアセスメントが求められるだけでなく、しばしば日常生活の安定に向け、関係機関や家族も取り込んでの連携といった、ソーシャルワークの要素を含んだ対応が必要となる場合もある。個人要因以外へのアプローチも必須であり、例えば、事業主に対する理解の促進や、実際の職場における環境調整が定着における重要な要因となる。精神障害者や発達障害者の就労支援を講じるべき範囲は広い。

したがって、障害者の個別性に適した支援の提供が一機関だけで叶わないことも稀ではない。地域の就労支援機関（障害者就業・生活支援センター、就労移行支援事業所、ハローワーク）等において十分な対応が図れなかった精神障害者、発達障害者のケースについて、地域センターに専門的支援を要請されるケースがあることが障害者職業総合センター（2019）の調査で指摘されている。

同調査からは、支援の困難性が認められたケースについて、困難要因が分類されており、具体的には「情報処理面（心身機能)」、「体調等の身体面（心身機能)」、「仕事理解面（心身機能)」、「自己理解面（心身機能)」、「感情・思考面（心身機能)」、「業務遂行面（活動・参加)」、「対人行動面（活

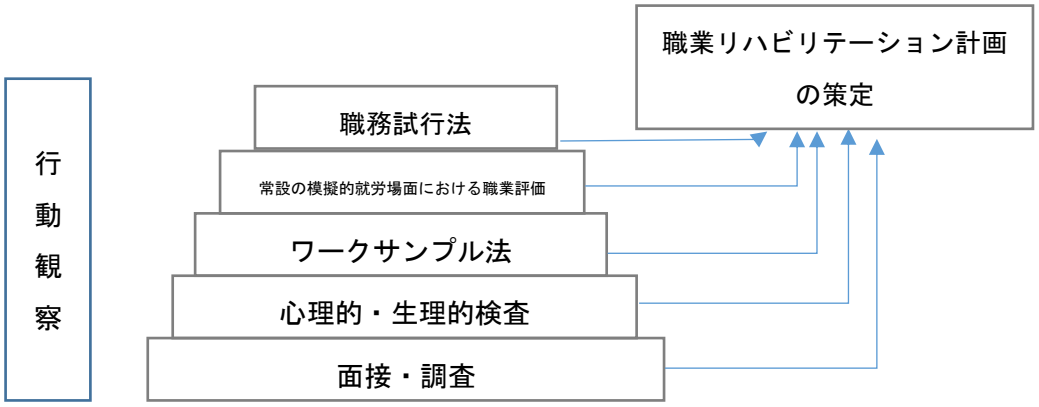
動・参加)」、「生活面（活動・参加)」が提示された。困難要因を背景とする多様な問題に対処する上では、実践場面で活用可能な支援技法や、支援の困難性に関連する特性を的確にアセスメントするための方法論を開発・検討することが最優先の課題であると考えられる。

## ２　障害者の職業評価における課題

就労支援に当たっては、障害が就労に当たってどのような影響を及ぼす可能性があるか、対象者の状態像を把握するために、職業的な評価を実施することが必要である。松為（2016）は職業リハビリテーションについて「生物・心理・社会的な障害のある人が、主体的に選択した仕事役割の継続を通して生活の質が向上するように、発達過程の全体を通して多面的に支援し、それにより社会への統合又は再統合を促進する総合的な活動」と定義している。職業リハビリテーションのサービスによる支援全体は、障害のある人の職業的な自立を支えるための様々な支援手法・支援技法が用いられるダイナミックな活動であり、その職業リハビリテーションの中でも、具体的な支援を計画していくために実施される職業評価が、重要な機能を果たしていると考えられる。

地域センターが実施する「障害者の雇用の促進等に関する法律」によって定義される「職業評価」（以下『職業評価』として記載し、他の職業評価と区別する。）とは、職業リハビリテーション計画の策定に向けて、面接や諸心理検査の実施や作業評価、模擬的就労場面における行動観察等、職務試行法などの結果と、これらの実施に伴う行動観察のほか、他機関から得られた情報なども総括して展開される一連の評価活動を示すものとされている（図序－１)。根拠に基づいた支援を実施する上で、『職業評価』は欠かせないものとなっている。

この『職業評価』について、厚生労働省職業安定局高齢・障害者雇用対策部（2003）では、地域センターの個別業務として、次のように解説している（表序－１)。




**図序－１　地域センターにおける職業能力・適性等の評価法の体系**

（「就業支援ハンドブック（高齢・障害・求職者雇用支援機構，2021)」より再構成）

**表序-１　『職業評価』についての解説（厚生労働省，2003）**

障害者の職業能力、適性等を評価し、及び必要な職業リハビリテーションの措置を判定することとされており、

①生理機能検査、心理検査等を用いて身体的側面、心理的側面等の障害者の諸特性を把握すること

②職業適性検査、職業興味検査等の各種検査、ワークサンプル法及び職務試行法を用いて障害者の職業能力を把握すること

③これらの方法を通じて、障害者の職業能力・適性に関する現状と将来性についての知見と見通しを得て、労働市場の状況及び障害者の周辺の諸環境等を総合的に勘案し、障害者が最も適した職業領域において職業的自立を図るために受けるべき職業リハビリテーションの措置を明らかにする職業リハビリテーション計画を策定すること等を内容とするものである。

以上の記述からは、『職業評価』で実際に行われることとは、客観的な評価指標を用いた特性把握のほか、相談場面、評価場面や実際の職業的な場面での観察等を通じた、職業能力や適性の把握であることがわかる。これらの活動を通じて職業リハビリテーション計画が策定されることになる。すなわち、『職業評価』とは、職業リハビリテーション計画を立てるための根拠を収集して整理し、対象者について多角的な視点で把握するプロセスをもつ活動であるといえるだろう。

昨今の職業リハビリテーションの利用者の増加や多様化に対応するためには、職業評価の活動全体のみならず、それに用いられる検査器具や評価ツールの質的な向上も求められると考えられる。ところが、アセスメントツールの新規開発や品質向上のための方法については、議論が尽くされたとは言い難い現状がある。

## ３　研究目的

そこで、本調査研究では、職業リハビリテーションにおけるアセスメントツールを概観し、検査や評価ツールの開発プロセスや、客観的評価の視点に必要となる統計的知識の整理、加えて、支援の現場でアセスメント等に利用されているツール（ワークサンプル幕張版『社内郵便物仕分（簡易版）』）を用いたシミュレーションと分析等を通して、将来的なアセスメントツールの発展に資する基礎的資料を整備することを目的とする。

**【文献】**

独立行政法人高齢・障害・求職者雇用支援機構，2021 年，令和３年度版　就業支援ハンドブック.

厚生労働省職業安定局障害者雇用対策室地域就労支援室，令和元年６月，平成 30 年度障害者雇用実態調査結果.

厚生労働省職業安定局高齢・障害者雇用対策部編著，2003 年，二　個別業務の解説　（１）職業評価（一⑤イ関係），障害者雇用促進法の逐条解説，日刊労働通信社．52－53.

松為信雄，2016 年，独立行政法人高齢・障害・求職者雇用支援機構 障害者職業カウンセラー厚生労働大臣指摘講習テキスト第３版，第１章【総論】職業リハビリテーション、第１節第３項　職業リ

ハビリテーションの定義と対象.
障害者職業総合センター　調査研究報告書 No.95（2010 年）「精神障害者の雇用促進のための就業状況等に関する調査研究」
障害者職業総合センター　調査研究報告書 No.99（2011 年）「高次脳機能障害・発達障害のある者の職業生活における支援の必要性に応じた障害認定の在り方に関する基礎的研究」
障害者職業総合センター　調査研究報告書 No.153（2021 年）「障害のある求職者の実態等に関する調査研究」
障害者職業総合センター　調査研究報告書 No.144（2019 年）「支援困難と判断された精神障害及び発達障害者に対する支援の実態に関する調査　-地域の支援機関から地域障害者職業センターに支援要請のあった事例について-」

# 第１章

# 職業リハビリテーションにおける職業評価

# 第１章　職業リハビリテーションにおける職業評価

## 第１節　職業評価に着目する必要性

障害がある人の QOL を向上させるために展開される実践といえば、医学的な治療、教育による指導・訓練・支援、福祉サービスによる相談や日常生活の援助、そして労働参加を可能とするための相談・支援・訓練といったように、その方法や場面は多様である。しかし、いずれにも共通していえることは、評価[1]が適切になされることをもって、望ましい介入（相談・治療・リハビリテーション・訓練・援助・介助・指導・教育等）を企てることが可能となる点である。特に、就労支援において実施される職業的な評価の結果は、障害者の進路やキャリアの決定のみならず、その後の長いライフスパンにも影響を及ぼすと考えられる。

知的障害、精神障害、発達障害、高次脳機能障害、難病など、一見しただけでは把握することのできないような障害・疾患の場合は、なおさら的確な評価を講じることが求められる。

障害者の就労支援における職業的な評価とは、具体的にどのような取組を指すかをみていきたい。

「障害者の雇用の促進等に関する法律」では、『職業評価』について「障害者の職業能力、適性等を評価し、及び必要な職業リハビリテーションの措置を判定すること」と定義し、障害者職業総合センター、広域障害者職業センター、地域センターによって実施される職業リハビリテーションサービスの一部としている。

さらに視点を拡大すると、国際労働機関：ILO（1985）では「職業リハビリテーションの基本的原則」の書籍の「今日の職業リハビリテーション」の章において職業評価について触れている。そこでは、職業リハビリテーションの勧告（第 99 号）で定義される職業リハビリテーションのサービスに含まれるものとして６つの項目を挙げ、そのうちの一つが、「障害者の身体的・精神的・職業的な能力と可能性について、明確な実態を把握すること（職業評価）」とされている。さらに、同書第２章「Vocatinal assesement and work conditioning」においては、職業評価の意味として、「障害者が、雇用のための適切な準備ができたり（例えば、職業訓練等による）、再就職ができる（例えば、職業あっせんや追指導（フォローアップ）などによる）ためには、事前に、その人のキャパシティ（力量）、アビリティ（能力）、ポテンシャリティ（可能性）、エンプロイアビリティ（雇用の可能性）等に関して、評価されなければならない」ことが示されており、障害者の就労可能性を高める上で、職業評価は継続的な活動であるとともに、対象者のポジティブな側面を積極的に評価する必要があることを明記している。

他にも、Stanford E.R ら（2016）は、「障害者のリハビリテーションのプロセスは、評価から始まり、計画、治療、終了（配置）までの４段階のシーケンスとして最もよく説明される」とし、職業リハビリテーションの４段階のうち、“評価段階”の目的について、「障害のある人が（ａ）現在及び潜在的な職業的機能と興味の範囲をよりよく理解することを援助すること、（ｂ）互換性のある潜在的な仕事の機会に気付くのを助けること、そしてそのような機能的能力に関心を持って（ｃ）

---

[1] 本文では評価という用語をアセスメント（assessment）と職業評価（evaluation）に区別して用いることとする。前者は、ツール等を用いた客観的評価（定量的・定性的）を意味するものとする。後者は、アセスメント結果等を含む多様な情報を統合して、対象者の職業適性等を総括的に判断することを意味するものとする。

リハビリテーションサービスとその機能を最適化するために必要なサポートについて学ぶこと」と示している。

各々の職業評価の定義からは、職業リハビリテーションにおける職業的な評価活動全般が、必ずしも対象者の一次的な機能障害のみに着目するものではないことが明確である。むしろ、職業評価とは、障害を含めた個人の状態と職場等の環境条件下における相互作用の結果も含めて評価する活動であり、更には個人の潜在的な能力を見出すこと、本人の興味や希望も含めた上でどのような仕事に向いているかを検討し、必要に応じて職業準備性を最適化するための一連のサポートであることとして理解できる。このような職業評価の理念と、昨今の職業リハビリテーションサービスの利用者の障害の多様化、重度化の傾向を踏まえて、職業評価には的確さが期待される。そのためには、客観性が高く、効果的に利用できるアセスメントツールが必須になるといえる。

障害者職業総合センター（2004a）は、「効果的なカウンセリングのためには、評価が適正であることが必要不可欠である」と、介入前評価によって得られる情報そのものに妥当性[2]・信頼性[3]が求められることを指摘している。この指摘は、実践家としての支援者の立場においては「どのような対象者に、どのアセスメントツールを用いて評価することが妥当か、そして、そのツールの適切な使い方はどのようなものか」といった問題として捉えられるだろう。支援者による評価の信頼性・妥当性とは、いかに的確に評価全体を実施するかを意味するといえる。

支援者による評価が適正であるためには、用いられるアセスメントツールの品質についても議論する必要がある。例えば、昨今の障害者の多様化の状況から、既存のツールでは個々の状態を十分に捉えられない問題をどうすればよいのだろうか。支援者の評価の力量に頼るばかりではなく、用いられるアセスメントツールの妥当性・信頼性については別途担保していくことが求められるだろう。

村上ら（2013）は、「はっきりと目に見えない、知的・社会的ハンディキャップを何とか可視化し客観化することの必要性が、多くの人々に痛感されるのは事実であろう。可視化の手段として、何らかの数値的評価の道具であるアセスメント・ツールを開発・改善することは、喫緊の課題であると思われる」とし、障害の評価に用いるためのツールそのものの開発・改善の重要性を指摘している。

そこで本調査研究では職業リハビリテーションの領域において、職業評価等に用いられるアセスメントツールの客観性を高める取組を積極的に行う必要があるとの立場から、特に、アセスメントツールの測定精度をいかに改善させるのかという点に着目する。

まずは次節において、職業リハビリテーションの領域やその周辺領域で用いられる、各種検査や尺度を概観した上で、本領域で利用されるアセスメントツールにおける課題やニーズについて整理する。

---

[2] 測定したいと考える概念を、その課題やテストによって実際に測定している程度・度合いのこと。
[3] ある課題を繰り返し実施したときに、一貫して同一の得点が得られる程度・度合のこと。

## 第２節　職業リハビリテーションと周辺領域における多様なアセスメントツール

障害者の職業的課題のアセスメントを行う上では、多様な客観的指標を用いることが多い。アセスメントの実施のために複数の検査を組み合わせることで、障害の個別特性を効果的に把握することにつながる。

では、どのようなアセスメントツールが実際に利用されているかを見ておくこととしたい。障害者職業総合センター（2011）による調査結果からは、高次脳機能障害者及び発達障害者の職業上の問題を把握するための効果的な検査が明らかとなっている（図１－１、図１－２）。

各グラフは、診断・治療を行う医療機関と、職業リハビリテーション機関である地域センターによる回答結果で、複数回答からなっている。検査の組合せは不明だが、どの検査の利用頻度が高いかの結果としてみることができる。

多様な検査の中でも、「WAIS-Ⅲ」は医療機関、地域センターのいずれの機関においても、また、いずれの障害種別においても有効と回答された割合が高い。さらに、地域センターでは、いずれの障害種別にも共通して「GATB（一般職業適性検査）」、「ワークサンプル（MWS；ワークサンプル幕張版）、マイクロタワー日本版、その他含む」、「模擬的な就労場面を活用した職業評価」の回答率が高かった。

なお、医療機関においては、有効とされる検査としては神経心理学的検査に属する検査の方が多数あげられており、特に、神経・認知面の特性に対するアセスメントが重視されていると見られた。

一方の地域センター（職業リハビリテーションの場面）においては、神経心理学的な所見は参考にしつつも、実際の作業的場面あるいは職業的な場面において、具体的に表出する行動特徴を観察したり、アセスメントすることが重要視されていると見ることができた。

このような結果からは、医学的な所見に必要な情報と、就労支援に当たって把握すべき情報の重きが異なっていることが指摘された。

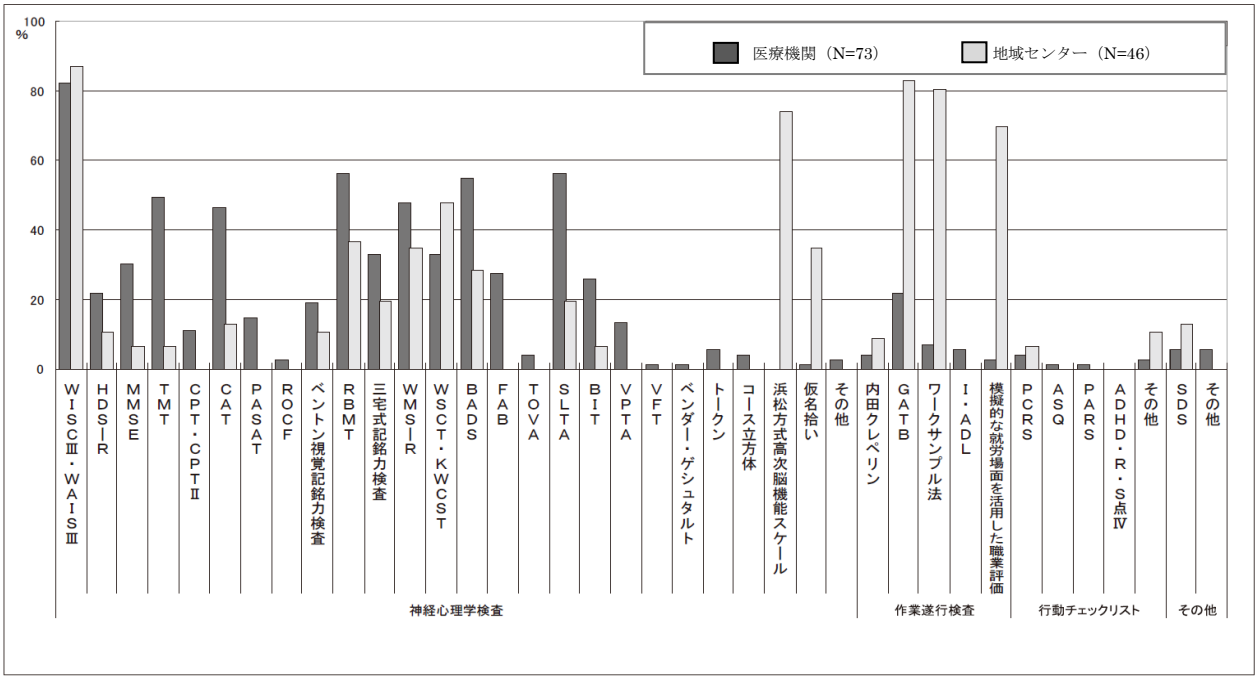



図１－１　職業上の困難度を把握するために有効と思われる検査（高次脳機能障害：複数回答）

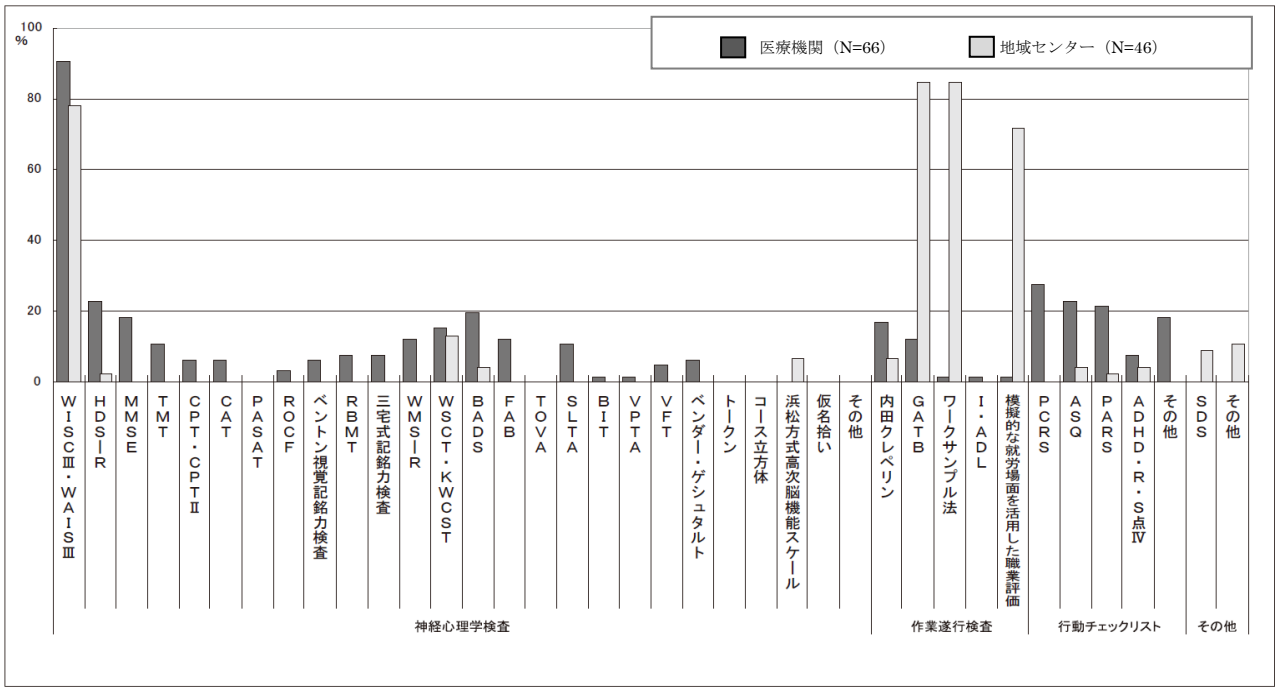



図１－２　職業上の困難度を把握するために有効と思われる検査（発達障害：複数回答）

神経心理学検査の略称…WISC-Ⅲ･WAIS-Ⅲ：ウェクスラー知能検査／ HDS-R：長谷川式簡易知能評価スケール／MMSE：ミニメンタルテスト／ TMT：トレイルメイキング検査／ CPT・CPT Ⅱ：持続性遂行テスト／ CAT：標準注意検査法／ PASAT：定速聴覚連続付加検査／ ROCF：Rey-Osterrieth の複雑図形／ RBMT：リバーミード行動記憶検査／ WMS-R：ウェクスラー記憶検査／ WCST・KWCST：ウィスコンシンカード分類検査／ BADS：遂行機能評価／ FAB：前頭葉機能検査／ TOVA：注意変動検査／ SLTA：標準失語症検査／ BIT：行動性無視検査／ VPTA：標準高次視知覚検査／ VFT・WFT：語の流暢性テスト

行動チェックリストの略称…… PCRS：Patient Competency Rating Scale ／ ASQ：自閉症スクリーニング質問紙／PARS：思春期成人期尺度

発達障害者支援でよく使用されるアセスメントツールについては、アスペ・エルデの会（2013）の調査が詳しい（表１－１）。発達障害においては、知的水準の幅が広いことから、知能検査が広く実施されている。しかし、知的水準の高低が必ずしも適応状態と関連しないことがわかっており、主障害の状態や行動特徴も含めて的確に評価することができるツールが多数紹介されている。どのツールを用いるかについては、支援者が対象者の状態に応じて検討することになるが、「支援のために非常に役に立つ情報が得られる（明翫，2013)」ことが重要である。

なお、標準化されたこれらのアセスメントツールは、それぞれに決められた所定の実施手続を厳密に守って遂行することが求められる。したがって、支援者においては、検査実施のためのトレーニングや研修受講を通じて、検査を的確に運用するための経験やスキル、場合によっては資格さえも求められるなど、専門性に基づく利用・運用が必須となることも稀ではない。

**表１－１　発達障害領域でよく使用されるアセスメントツール**

知能検査・発達検査系

ウェクスラー式知能検査　田中ビネー知能検査　新版Ｋ式発達検査
日本版 Bayley-Ⅲ乳幼児発達検査　KABC-Ⅱ　ASQ-３（Ages and Stages Questionnaire, Third Edition）

適応行動（生活能力）のアセスメント

日本版バインランド適応行動尺度Ⅱ　新版 S-M 社会生活能力尺度Ⅱ　ASA 旭出式社会適応スキル検査

情緒と行動の問題のアセスメント

CBCL･TRF　SDQ(Strengths and Difficulties Questionnaire)
異常行動チェックリスト日本語版（ABC-J）
日本版感覚プロフィール（Japanese Version of the Sensory Profile:SP-J）
RBS-R（アールビーエスアール）日本版
不適応行動のアセスメント：強度行動障害

ASD のアセスメント

M-CHAT（エムチャット）日本版　PARS　小児自閉症評定尺度：CARS　日本語版と CARS2-HF
PEP-3（心理教育プロフィール３）　TTAP（TEACCH 移行アセスメント）
DISCO-11（The Diagnostic Interview for Social and Communication Disorders-11）
ADOS（エードス）　AQ（エーキュー）日本語版（自閉症スペクトラム指数 Autism-spectrum quotient）

ADHD･LD･DCD その他のアセスメント

ADHD の評価尺度
CAARS（カーズ）日本版
LDI-R（エルディーアイ・アール：LD 判断のための調査票）
教研式標準学力検査 NRT（集団基準準拠検査）
音読検査
言語系のアセスメント：ITPA や PVT
書字のアセスメント
計算のアセスメント
協調運動機能のアセスメント：DCDQ-R、Movement-ABC2（M-ABC2）
感覚と運動のアセスメント　－JMAP と JPAN－

次に、精神障害に利用されているアセスメントツールを見ておきたい。ここでは、精神科において治療対象となるうつ病患者に限定して、その評価方法を取り上げた。

Cheung ら（2012）は、14 のうつ評価尺度をレビューし、内容にギャップがあることに触れている（表１－２）。それに対して、渡邊（2020）は、「Recovery の時代において、現在の尺度は、当事者、治療者そして時代のニーズに合っていないと言わざるを得ない。どの尺度も、患者にまつわるすべての要素を扱っているわけではなく、我々治療者は、当事者のありのままを描写するのに、自殺念慮のより細やかな評価、さらには当事者の満足度や認知機能、社会機能、QOL 等に焦点を当てた尺度をその都度組み合わせて評価するほかないのである。また、客観的評価尺度においてはたとえ評価尺度が精微になったとしても、評価者による恣意的評価は問題となり続ける。…（中略）…評価者に依拠するが故のバイアスをいかに最小化するかが求められるだろう。」としている。

うつ等の精神症状を評価する尺度は複数あっても、測定内容には違いがあることや、治療者の判断で組み合わせて利用する必要があること、尺度が精緻であっても、使用する側に恣意的な評価が生じうることに注意すべき等の指摘は、アセスメントにおいて共通する留意事項であるといえる。

**表１－２　うつ病評価尺度の比較**

| | 評価尺度 | 評価される症状のカテゴリー | | | |
|---|---|---|---|---|---|
| | | 認知面 | 身体面 | 対人関係面 | 感情面 |
| 客観的 | HAM-D | 4 | 9 | 0 | 4 |
| | GDS | 18 | 3 | 1 | 8 |
| | MADRS | 4 | 3 | 1 | 3 |
| 自己記入式 | ＢＤＩ－Ⅱ | 11 | 4 | 0 | 4 |
| | HAD Scale(Depression scale) | 5 | 1 | 0 | 8 |
| | Zung SD Scale | 6 | 8 | 1 | 3 |
| | PHQ-9 | 4 | 4 | 0 | 5 |
| | DASS | 8 | 3 | 1 | 1 |
| | CDS | 20 | 21 | 2 | 5 |
| | EPDS | 4 | 0 | 0 | 1 |
| | CES-D | 8 | 5 | 2 | 8 |
| | GHQ-28(Depression scale) | 6 | 1 | 0 | 9 |
| | CSRS | 10 | 11 | 3 | 6 |
| | QIDS-SR | 4 | 11 | 0 | 6 |

※（Cheung,H.N　et al.,（2021）による渡邊（2020）の改編を引用）

BDI-Ⅱ=ベック抑うつ質問票. HAD Scale=Hospital Anxiety and Depression Scale. HAM-D=ハミルトンうつ病評価尺度. Zung SD Scale=Zung の自己評価式抑うつ尺度. PHQ-9＝こころとからだの質問票、DASS＝Depression Anciety Stress Scales. CDS=Caroll Depression Scale. GDS=老年期うつ病評価尺度. EPDS＝エジンバラ産後うつ病自己評価票. CES-D＝うつ病自己評価尺度. GHQ-28＝精神健康調査票. CDRS=Cornell Dysthymia Rating Scale. MADRS=モンゴメリーアスバーグうつ病評価尺度. QIDS-SR=うつ度チェック簡易抑うつ病状尺度.

以上、障害ごとに利用されているアセスメントツールを概観した。使用機関・使用者や対象によって利用されるツールは様々であること、ツールを利用する上では専門性が求められること、また、専門性をもって使用する上でも一つのツールで明らかになることは限られており、複数のツールを組み合わせるなどの工夫を要するが明らかとなった。これらは全てアセスメントツールを利用する側に課せられた課題であった。

本調査研究では、上記の課題を達成するための、アセスメントツールそのものの品質の問題を取り上げることを目的としている。そもそも尺度が精緻でなければ、前述において留意事項としてあがった“恣意性”や“バイアス”といった問題は言及されるまでもない。

例えば、支援者のＡさんは検査マニュアルに沿って 10 分を計測する際、愛用の機械式時計を使うのだが、その日は時計に不具合があり、運悪く 10 分のところ 12 分も経過していた（この種の問題は第２章で“偶然誤差”として扱われる)。このように、測定器具の不備などがあると、得られる計測結果に度々ばらつきが生じ、実態が適切に反映されないデータばかりが蓄積してしまう。

アセスメントツールを利用しようとする者は、そのツールが正当に機能するものであるかどうかについて細心の注意を払う必要がある。さらにそれに加えて、適切な利用を図ることが、的確なアセスメントにつながることを教訓としなければならないだろう。

さて、アセスメントの評価対象の範囲（「神経・認知面」、「社会的行動面・対人面」、「心理・情緒・感情面」、「学習面」、「運動面」）が幅広いことは本節の多様な検査からも明らかである。また、障害を把握する上では広い視点で適切なツールを選ぶことが重要であることも指摘されていた。だが、多角的な視点や多数のツールで評価しようとすればするほど、相応のコストが必要となる。そこで次に、丁寧なアセスメントを実施しようとする際の実施者側におけるコストの問題について触れることとしたい。

## 第３節　職業的な困難を評価する上でのツール開発の課題

### （１）アセスメントのコスト面の課題

職業リハビリテーションサービスの利用者は増加傾向にあるが、特に精神障害や発達障害等の目に見えづらい機能障害においては、障害特性の個別性が高く、より丁寧な支援を要する。しかし、その“丁寧さ”とは、裏を返せば相応のコスト（＝人的な労力、時間的要素、経済的負担等）を要するということである。

例えば、職業リハビリテーションの支援プロセスは、アセスメントに基づいた支援計画における目標・目的に沿って、日々の本人の体調やペースといった可変動の状態像を、継続的かつ感度よく観測しながら行われる。同時に、これらの観測結果に基づいて支援者は臨機応変に介入方法の最適解を導いている。日頃の支援活動には、常に支援者の見立てというアセスメントの視点が機能している。このような活動の質の維持は、支援者に蓄積された経験と専門性によるといえる。しかし、支援現場は人的資源や時間的制約を抱えており、インテーク段階から相談・支援活動のプロセス全体を通した効率性が求められる。効率性が高いということは、仮に介入・改善すべき案件が幅広くあっても、優先的な課題への対応が先行し、対応すべき課題について取捨選択する場面も出てくる

ということである。

実際に、支援者はアセスメントにどの程度のコストを掛けられるのだろうか。専門性の高い支援者であれば、アセスメントを効率的に実施できる可能性は高い。一方で、アセスメントをどれだけ丁寧に行っても、それ自体は直接的に就職や復職や職場定着に結び付くものではない。

無用にアセスメントや職業評価のフェーズのみにコストを割くことができないのが支援の現場のジレンマであるといえる。したがって、アセスメントの諸活動には客観性という要素のみならず、効率性という要素こそ欠かせないことを念頭に、品質向上について検討することとしたい。

**（２）アセスメントツールの品質向上のためのプロセス**

職業評価の活動においてアセスメントの客観性と効率性を高めるためには、高品質なアセスメントツールの利用が欠かせない。逆に、アセスメントツールの品質を高めることは、アセスメントの客観性や効率性を高めることに繋がるといえる。

品質の高いアセスメントツールが世の中に普及するまでには、ツールの開発段階、開発後の利用普及の段階、時代の変遷による改修の段階といったサイクルがあり、各々にどのような取組があるかを本項で概観することとしたい。

心理学辞典（2002）によれば、一般に広く利用される知能検査、学力検査、性格検査などは、「検査実施時の受験者に対する“教示”のやり方、問題項目の呈示法、解答法の指示、検査時間などの検査実施法、および各項目に対する受験者の反応の採点法が厳密に規定されており、しかも受験者個人の結果は準拠集団の得点分布にも基づいて作成された集団規準に照らして得点が解釈できるように作成されている」とのことである。検査や評価尺度等は、一般に実施方法が決まっており、得られた結果を基準値と照らすことで評価することが可能となるものである。では、これらはどのような過程で開発、実装されるのだろうか。

図１－３には標準テストの作成フローチャート（吉森，1978）を、また表１－３にはテストの開発及び頒布にかかわる重要事項（日本テスト学会，2007）を示した。標準値や基準値を持つ検査や評価尺度は、これらに示された手続を経ることで検査の信頼性・妥当性を向上させ、品質保証が図られる。

例えば、進路先を決定したい、職業適性を判断したい、入学試験として合否を判定したい、といったように、テストを受ける者の人生に深く関わるような判断がテストの結果からなされる場合、そのようなテストは「ハイステークスなテスト」（光永，2021）といわれ、評価結果は十分に信用に足るものでなければならず、テスト（評価尺度）には高い品質が求められる。その開発にはかなりの資源が割かれ、開発後も定期的な改修がなされる。

一方で、参加者アンケートや確認事項のチェックリスト、日々の豆テスト等は、その目的が単に意見を広く収集することや、行動確認、反復学習等であり、被検者の人生を大きく左右するほどのテストではないために、設定に高い品質は求められない。このようなテストは「ロウステークステスト」とよばれるが、図１－３に示したような手続は要せず、尺度としての信頼性・妥当性は必ずしも期待されない。

しかし、障害のある人を支援するために行われる評価の場合には、ある特性や状態、症状を的確

に捉えたいと考えるのが当然である。何らかの支援やサービスを処遇する上では、評価結果が根拠（エビデンス）となるためである。したがって、利用する評価尺度やアセスメントツールの信頼性・妥当性が十分に担保される必要があるだろう。もしツールに品質が保証されていなければ、アセスメントしたいと考えている現象を的確に捉えられない可能性が大きい（例えば、明らかにしたい現象が見えにくい、障害特性がうまく捉えられない、評価者によって評価結果が大きく異なる等）。このため、多数存在する様々なアセスメントツールや評価尺度の品質について、それを利用する者はよく理解することが求められよう。この問題は、第２節でも言及した、アセスメントツールの機能についての問題である。

また、アセスメントツールや評価尺度の開発者の責任は、利用者に向けて、その品質、利用目的、利用方法、検査結果の意味や解釈の限界（信頼性・妥当性の検証結果）等も含め、報告書やマニュアル、関連文献等で言及することである。

さて、アセスメントツールの開発は、その領域の研究者が中心となって進められることが一般的である。標準化された検査等が開発された後、市販を目的とした製造（作成）、販売、研修等の頒布活動は開発（研究）とは別に役割分業が図られることが多い。ツールの利用の場や対象規模によって、販売・利用に当たってのサポートを継続的に進めていくために特化した人材や体制整備が必要となることもある。

表１－３は日本テスト学会が詳細に示した「テストの開発および頒布にかかわる重要事項」であるが、図１－３の標準テストの作成フローチャートに比較して、テストの品質保持のプロセスに着目していることがわかる。神経心理学的検査のように個別に実施するテストの改訂は頻繁には生じない。だが、大学入試や資格試験など、テスト受験者が大規模で、短い期間に高頻度で実施されるテストの場合には、大量の質が担保された問題（通称“項目プール”といわれる。）が必要となるためである。

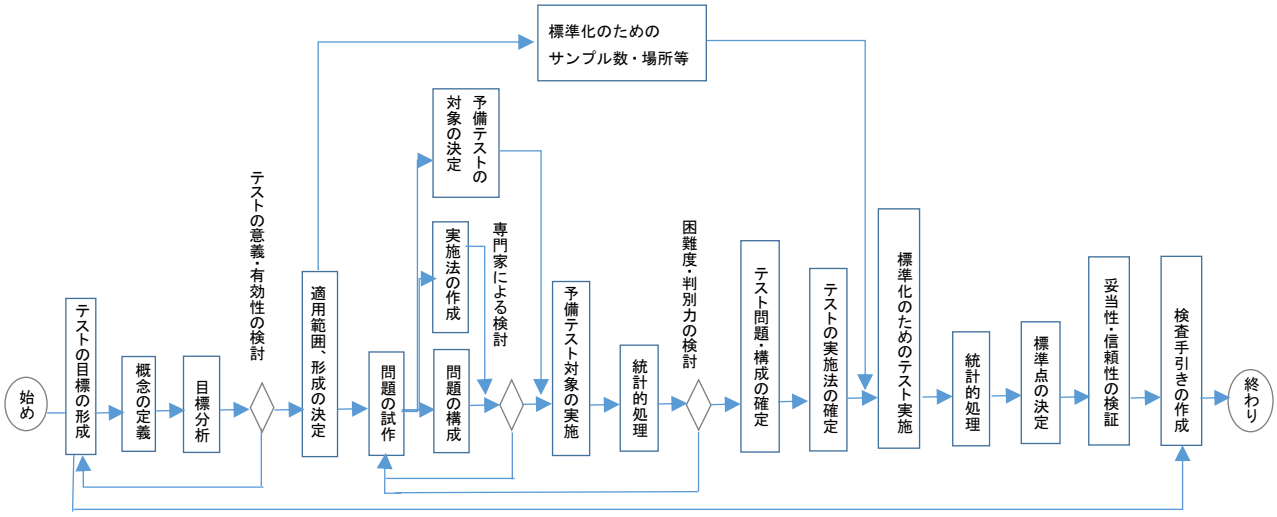



**図１－３　標準テストの作成フローチャート（吉森，1978）**

**表１－３　テストの開発および頒布にかかわる重要事項（「日本テスト学会，2007」から引用改編）**

| |
|---|
| 1　テストの基本設計<br>1　テストの目的　2　測定対象と測定内容　3　受験者層の想定　4　測定方式<br>5　テストシステムとしての配慮　6　ミスの防止　7　結果の利用方法 |
| 1．2　測定内容の定義と構造化<br>1　特性の構造　2　尺度構成と尺度得点　3　測定対象と特性値の１対１対応<br>4　尺度の性質　5　下位尺度の互換性　6　測定内容の組み換え<br>7　テストの内容にかかわる情報の公開 |
| 1．3　質問項目の設計<br>1　質問項目と作成者　2　質問項目作成の手引き　3　質問項目作成者の選定とトレーニング<br>4　テストに用いる測定形式　5　誤解の防止　6　質問項目の評価 |
| 1．4　回答方法の設計<br>1　回答形式の種類　2　回答形式の設計 |
| 1．5　採点手続きの設計<br>1　採点手引による採点　2　客観式テストにおける採点　3　客観式テスト以外の採点 |
| 1．6　尺度化の方法<br>1　質問項目に対する回答と尺度得点　2　尺度の水準　3　尺度化の方法の具体例 |
| 1．7　尺度の標準化<br>1　標準化の必要性　2　標準化の手続き　3　尺度の再構成とテストの改訂 |
| 1．8　複数の尺度得点の比較<br>1　尺度得点の利用<br>2　すべての受験者が複数個のテストのすべての質問項目に回答する場合<br>3　受験者が同一であっても回答する質問項目が異なる場合<br>4　同じ受験者が同一の質問項目を回答しても実施時期が異なる場合<br>5　共通受験者を用いた共通尺度化<br>6　共通尺度化の手続きの記録 |
| 1．9　複数のテストの共通尺度化<br>1　想定されている状況<br>2　受験者もテストも異なるが共通尺度が必要な場合<br>3　同一受験者の特性の変化量を測定する場合<br>4　発達、変化量測定の難しさ<br>5　開発者、頒布者および利用者の責務 |
| 1．１0　尺度得点の確からしさの推定と公開<br>1　尺度得点の安定性　2　信頼性係数の定義　3信頼性係数の推定方法<br>4　信頼性係数の推定にかかわる留意点　5　テストの信頼性に対する認識の重要性 |
| 1．１1　尺度得点の適切さの確認<br>1　尺度得点の適切さ　2　尺度得点の適切さの検討の重要性　3　質問項目の内容的な検討<br>4　統計的分析による妥当性の検討　5　多面的な検討の大切さ |
| 1．１2　テストの改訂<br>1　改訂と再標準化　2　改訂や再標準化の必要性<br>3　改訂および再標準化後のテストの実施と結果の利用 |
| 1．１3　質問項目の内容開示の是非<br>1　質問項目の開示についての検討　2　質問項目の内容開示が不適切な状況<br>3　質問項目の内容開示が可能な状況　4　構成なテストを受ける権利と質問項目の開示<br>5　大規模な項目プールが用意されている場合 |
| 1．１4　テストの多言語への翻訳<br>1　多言語への翻訳における検討事項<br>2　多言語に翻訳されたテストの再標準化 |
| 1．１5　手引き、解説書の内容<br>1　手引き、解説書作成の基本姿勢<br>2　手引き、解説書に掲載する内容<br>3　質問項目の作成および回答の採点に関わる手引、解説書 |

職業リハビリテーションの分野では、支援の諸活動を効果的に行うために、これまでにも様々なツールが開発されてきた。中でも障害者職業総合センターの研究部門では障害者に対する評価・支

援技法としての「職場適応促進のためのトータルパッケージ（以下「トータルパッケージ」という。）」（障害者職業総合センター，2004b）が開発され、実践現場で広く活用されている。

トータルパッケージは開発以降、職業リハビリテーションのサービスを利用する障害者の実態に合わせて、内容を改訂する活動が度々行われてきた。ツールの改修には多大なコストを要するが、今後、継続的にツールの維持・運営を行っていく上では検討すべき課題も山積している。

以下には、検査・尺度などのツール開発に当たっての現在の主な課題を示した。ここではトータルパッケージを想定しているが、その他のアセスメントツールも共通の課題を抱えているといえる。

**表１－４　検査・尺度等のツールの開発プロセスの課題**

① 開発の基本プロセスを経ること
　ツールのコンセプト、開発方法、課題の構成、基準値等の作成などは、標準テストや尺度開発を行うことと同様の基本的プロセスを経ることが必要である。

② 開発に係る手続を公開すること
　開発プロセスを記述、開示した上で、他の研究者が自由に検証・追試できるようにしなければならない。

③ ツールの維持・運営の体制を検討すること
　ツールを開発後、継続的にツールの改修・維持を行う体制を別途検討しなければならない。

④ ツールの利用者に対する研修機会を提供すること
　ツールの利用者に対して、実施方法を教育する機会を提供することが求められる。ツールの利用が拡大すれば、これらの活動にあてるコストは更に拡大する。

上記の課題について詳細に触れてみたい。

①「開発の基本プロセスを経ること」

客観的なアセスメントを達成するための基本的条件は、道具の開発において適切な手続を取ることである。第一に、これからアセスメントしたいと考えている現象や事柄について、明確に定義することである。心理領域では、“構成概念”といわれることが多い。例えば、適応行動、不適応行動、自閉症スペクトラムの特性、不安傾向など、測定対象となる状態や現象の幅は広い。既存の尺度やツールを利用して事足りる場合には、新たに作成する必要はないだろう。既存のツールでは明らかにしにくい現象や事象、状態があれば、開発が必要となる。そのためには、アセスメントしたい対象を関係者が十分に議論し、関連の先行研究を整理した上で、その現象や事象、状態を操作的に定義しなければならない。

第二に、ある状態について評価・測定する上で、想定されるツールはどのような道具であるかを決める必要がある。質問紙調査か、行動観察表か、構造化面接か、道具を使った行動測定なのか、測定のための方法・道具を決定する必要がある。その際、検査目的や利用現場のニーズを踏まえ最も的確・効果的に捉えられる方法が選択されるべきである。

第三に、アセスメントの対象とする現象や事象等に応じて、どのような課題の構造であれば的確に測定可能となるかについての検討が必要となる。想定され得る最も合理的な構成が求められる。

第四に、その構造によってアセスメントできる対象の幅を特定することである。年齢、性別といった属性、障害種別、どのような対象者に利用できるツールなのかを明確にする必要がある。さらに、想定モデルがアセスメントを受けた際、参考とすることが可能な標準値を装備することが客観性を担保することになる。特に客観性をアセスメントに持たせたい場合には、ある基準に照らして対象者がどの位置に位置づけられるか相対的に把握できるほうがよい。

これらの開発プロセスを経て、信頼性・妥当性を確保することが求められる。しかし、第四に示した標準値は、多くの開発協力者を要するため必ずしも整備できない場合もある。標準値の有無等については、マニュアル類に明示しなければならない。

②「開発に係る手続を公開すること」

①に示した開発プロセスについては、通常、開発者（研究者）が行うことが多いため、報告書や関連の科学雑誌への掲載などによって一般に公開・開示され、他の研究者の目にさらされることとなる。科学的観点からは、開発プロセスや得られたデータは公開される必要がある。情報が公開されることで、ツールの信頼性・妥当性について説明がなされ、ユーザーのツールに対する理解や認知が向上する。また、科学的な発展のために他の研究者による自由な検証・追試できるような形での公開が求められよう。

③「ツールの維持・運営の体制を維持すること」

ツールは開発後に公開され、広く利用されることとなるが、時間・時代とともにユーザーや被検者のニーズが変化する。この過程で、あまり使われないツールは淘汰されていくこととなる。逆に、頻繁な利用のあるツールは、継続的な改修・維持活動が求められることとなる。その場合、利用規模やツールの構成にもよるが、開発者（研究者）のみで改修・維持活動を行うことは難しく、ツールの改修・維持等の体制を別途構築していく必要がある。

④「ツールの利用者に対する研修機会を提供すること」

開発されたツールを的確に利用・運用できる人材を育成する必要がある。誰が評価し、誰が結果を作成し、誰が結果を利用して支援をするのか等も含め、そのツールにおける考え方を開発段階で想定した上で、利用マニュアルなどに明記する必要がある。また、ツールの利用方法などについて知識と方法を普及する活動が必要となる。

以上のように、アセスメントツールの成功とは、開発、利用・普及、改善といった各プロセスを適切に実施することにあるといえる。しかし、各フェーズは一朝一夕に成立するものではないことは火を見るよりも明らかである。アセスメントツールの開発には長期的な視野をもつことが望まれる。

**（３）基準値整備に当たっての課題　…ワークサンプル（幕張版）（MWS）を例に…**

職業リハビリテーションで利用されるトータルパッケージの構成要素の一つであるワークサンプル（幕張版）（以下「MWS」という。）は、地域センターをはじめ、広く就労支援機関で利用されている。MWS は、現在、13 の比較的簡単な課題と三つの新たに開発された難易度の高い課題から構成されている。

ワークサンプルとは、ある実在の仕事（作業）を切り出して、課題として構成されるものである。

Stanford ら（2016）は、ワークサンプルについて「実際の作業に関係する道具や材料を手順に従って使うように設計されている。加えて、職業評価の専門家によれば、ワークサンプルは、職業適性、労働者の気質、職業上の興味、手の器用さ、立ったり座ったりすることへの耐性、作業の習慣と行動、学習スタイル、及び書面と口頭での指示などを評価するために使用されているだろう」としている。障害者個々に現れる個別性を定性的に評価する上では、効果的なアセスメント方法であるとされる。

しかし、ワークサンプルを用いた場合であっても、個人の実施結果を評価するためには、成績を照合するための標準値を整備することが必要となる。MWS の開発においては、健常者に実施して収集したデータに基づく基準表の作成が企画された。その経過において、先行研究では基準値について「ワークサンプル法に限らず、障害者のための評価を目的に、障害者のデータに基づいて作成する上で隘路となるのが、『母集団が規定しにくい』『正規分布が想定できない』という点である（障害者職業総合センター，2004b）」と指摘している。そのため、ワークサンプル法による評価基準の開発にあたって多く採用されてきたのは、集団内における順位という統計量に基づいて作成するパーセンタイル基準であるとされている。

また、パーセンタイル順位は、任意の集団を対象としているので母集団の分布には関係なく、簡単に作成できるという利点がある一方で、新たなデータが加わると基準全体が変動しやすく、安定性が低いとされる。この点について障害者職業総合センター（2004b）は、MWS の基準値における信頼性と妥当性の点から「ワークサンプル法は心理検査と異なり、統計的な検証を行うためのデータ収集が難しい点もあり、また、「信頼性」と「妥当性」の検証を明確な形で行い、その有効性を向上させる努力を払うことが必要ではないかと考える」としている。この指摘は、開発時点で提供したパーセンタイル順位のもつ限界点と、基準値としての品質の向上について引き続き検討が必要であることを示唆している。

また、新たに開発された難易度の高い三つの課題についても、一般参考値（パーセンタイル順位）について次のような見解が示されている（表１－５）。

**表１－５　新規課題における一般参考値について**（障害者職業総合センター，2019）

| |
|---|
| 標準化に当たっては、相当な数の偏りのないデータを集めた上での集計と分析が前提となる。しかし、本研究において実施した一般成人に対するデータ収集については、派遣労働者（及び被験者紹介サービスの登録者）という一般成人を構成する一部の集団からしかデータを取得していない。また、MWS における基準値の在り方を「一般企業において当該作業に従事する人を準拠集団とした作業能力の評価基準」とするのであれば、一般労働者全般を対象にデータ収集を行う必要があり、派遣労働者のみで準拠集団とするには無理がある。しかし、一般労働者を対象としたデータ収集は極めてコストが高く、実現可能性が乏しい現状にあっては、派遣労働者から取得したデータに基づいて作成した統計値も、準拠集団を明確に意識して使用する限り、作業成績の解釈に資する参考値として機能しうると考えた。そこで、本研究においては、「標準化と基準値の作成」という表現を「一般成人を対象に作成をした一般参考値」という言葉に変えることとする。 |

ある作業に専門に従事する労働者集団（ピッキング作業に従事する労働者、郵便物の取扱いに従事する労働者、給与計算の業務に従事する労働者など）による作業データを標準値にすることができれば、測定結果がどのような位置に置かれるかを妥当に評価することが可能となるだろう。通常の評価尺度や検査の標準値は、そのテストの対象と想定される人と同一の属性にある集団によるデータを標準値に用いることが多い（例えば、体脂肪率などはその代表例である。体脂肪率は、性別と年代で基準値が分類されている。仮に 40 代女性の筆者が、誤って 20 代男性の基準値を参考にすると、完全な悲劇が起こるだろう）。

しかし、ワークサンプルにおいては、作業課題に密接な仕事に実際に従事している人々からデータを大量に収集することは、かなり困難である。これらのことから、現状の MWS の基準値整備に係る課題は次の点に集約される。

イ.「一般的な標準値に相当するデータをワークサンプルにおいて整備することは極めて困難である。」

ロ.「一般的な標準値に準ずるものとして扱っている“パーセンタイル順位による参考値”であっても、信頼性・妥当性の向上を検討することは必要である。」

MWS のアセスメントツールとしての品質を向上させる上では、標準値の信頼性・妥当性の向上の方法について検討を進めることが求められよう。

そして、MWS の試行結果に対する統計的な検討を行う上では、収集された集団の特性によって結果が変動してしまうといった問題と、MWS の各課題における難易度の両者を同時に考慮できる統計手法を用いることが望ましいだろう。その際、現代テスト理論における項目反応理論が役立つ可能性がある。

そこで、次の第２章では、項目反応理論を含むテスト理論全般について具体的な解説を行うとともに、第３章では実際に MWS の『社内郵便物仕分』を参考にして探索的にデータ分析を行い、アセスメントツールの精度向上のプロセスについて検討を進める。

**【文献】**

国際労働機関 ILO 駐日事務所ホームページ，1983 年の職業リハビリテーション及び雇用（障害者）条約（第 159 号），https://www.ilo.org/tokyo/standards/list-of-conventions/WCMS_238077/lang--ja/index.htm （2021. 10. 21 閲覧）

光永悠彦（2021），テストは何を測るのか　項目反応理論の考え方，ナカニシヤ出版.

明翫光宜（2013），第２節　心理アセスメントを活用することの有効性：心理アセスメントとは？，厚生労働省　平成 24 年度障害者総合福祉推進事業，発達障害児者支援とアセスメントに関するガイドライン.

村上隆・伊藤大幸・行廣隆次・谷伊織・平島太郎・安永和央（2013），アセスメントツールを用いることの重要性（１）：数値化することの意味，厚生労働省　平成 24 年度障害者総合福祉推進事業，発達障害児者支援とアセスメントに関するガイドライン.

日本テスト学会（2007），テスト・スタンダート　日本のテストの将来に向けて，金子書房.

野口裕之（2002）.標準化，心理学辞典（中島義明・安藤清志・子安増生・坂野雄二・繁枡算男・立花政夫・箱田雄司編集），有斐閣社，729.

International Labor Office Geneva （1985）.Vocational rehabilitation today, *Basic principles of vocational rehabilitation of the disabled, Third (Revised) Edition*.4-5.International Labor Organisation.

Stanford, E., Rubin, Richard, T. and Roessler, Phillip, D.Rumrill, Jr. （2016）.The Vocational Rehabilitation Process：Evaluation Phase. *Foundations of the Vocational Rehabiliation Process, Seventh Edition*. 267-305.PRO-ED.Inc.

障害者職業総合センター　調査研究報告書№56（2004 年 a）「「学習障害」を主訴とする者の就労支援の課題に関する研究（その２）－青年期における状態像の詳細区分に基づく検討―」，第３章　青年期における再評価

障害者職業総合センター　調査研究報告書№57（2004 年 b）「精神障害者等を中心とする職業リハビリテーション技法に関する総合的研究（最終報告書）」，第２章第４節５　MWS 実施のための標準化と基準表の作成，p172－p178

障害者職業総合センター　調査研究報告書№99（2011 年）「高次脳機能障害・発達障害のある者の職業生活における支援の必要性に応じた障害認定のあり方に関する基礎的研究」

障害者職業総合センター　調査研究報告書№145（2019 年）「障害の多様化に対応した職業リは離テーション支援ツールの開発（その２）－ワークサンプル幕張版（MWS）新規課題の開発－」

特定非営利活動法人アスペ・エルデの会（2013），発達障害児者支援とアセスメントに関するガイドライン，厚生労働省　平成 25 年度障害者総合福祉推進事業.

依田麻子・杉若弘子（2001），心理アセスメント序説，心理テストの条件，心理アセスメントハンドブック（監修上里一郎），西村書店.

渡邊衡一郎（2020），精神症状を理想的に測定するにはどうすればよいか―いまだに評価尺度で測定する必要性、その問題点、今後の方向性―　臨床精神薬理，23，451－458.

# 第２章

# テスト理論とは何か

# 第２章　テスト理論とは何か

本章では、テスト理論について解説する。テスト理論は、アセスメントやテスト開発の基礎を成す一連の数学的・統計的な理論である。テスト理論は古典的テスト理論と現代的テスト理論とに分かれるため、本章ではまず古典的テスト理論を概観し、その後、現代的テスト理論を概観する。なお、本章は数学や統計に詳しくない読者にもわかりやすくするために数式を用いず、また、読者に楽しんでテスト理論を理解してもらえるよう意図して執筆した。本章ではテスト理論の基本的な原則のみを扱っているが、テスト理論についてより詳しく知りたい読者は本章で引用した文献を参照されたい。

## 第１節　古典的テスト理論（Classical Test Theory）の考え方[1]

### １　古典的テスト理論の背景と特徴

1904 年にアメリカの心理学研究者、エドワード・リー・ソーンダイク（Edward Lee Thorndike）が初めて（と考えられている）テスト理論に関する本を出版すると、その理論は多くの研究者たちに歓迎され、心理学や教育学の大学院教育でテスト理論に関するカリキュラムが必修となるまでに時間はかからなかった。その後、1960 年代にかけて多くの研究者たちが当初ソーンダイクの提唱したテスト理論に新たな見解を加えたり、理論をより精緻にしたりすることで、テスト理論は発展していった。その後、1980 年代に入ると、テスト理論は急速に進歩することになる。これにはコンピュータ技術の発達（これによって、それまでは難しかった高度な計算ができるようになった）、心理測定を専門とする研究者の増加、客観的な指標へのニーズ（公費で運営されているサービスや教育プログラムのクライアント／学生がそれらサービスなどからどのような利益を受けているのかを客観的にはかる必要があった）が関係している。テスト理論において重要な転換点となった著作には、1952 年の Load によるものと 1960 年の Rasch によるものとがある。彼らの著作をきっかけに、それまでのテスト理論では扱うことのできなかった問題を新たな理論を用いて扱うことができるようになった。大まかに分けて、彼らの新たな理論提唱以前のテスト理論を古典的テスト理論と称し、それ以降のテスト理論を現代的テスト理論と称する。

古典的テスト理論は一つの理論を指すのではなく、複数の理論の集合体である。古典的テスト理論では、たいていの場合以下を前提としている。

≪X さんがあるテストを一度だけ受けた場合≫

（式１）X さんの今回の得点＝X さんの本当の得点+偶然誤差

---

[1] 本節は Allen, M.J., & Yen, W.M. （1979）; Crocker, L., & Algina, J. （1986）; Kline, J.B.T. （2005）に基づいて執筆した。

上の式の「X さんの本当の得点」とはこの場合（あるいは理論上）、X さんが永遠にこのテストを繰り返し受け続けた場合のテストの平均点を指す。しかしここで一つの疑問が生じる。永遠にテストを受け続けるのであれば、平均点を算出することは不可能なのではないだろうか。不可能である。X さんがテストに解答し続けている限り平均点は変わっていく。固定的な平均点を算出するためには X さんにテストをやめてもらわなければならない。しかしこの「永遠にテストを受け続ける」というのは先にも触れたとおり、あくまでも理論上の話であるので、読者には「永遠にテストを受け続ける」≒「干からびるほどテストを受け続ける」ということだと解釈していただきたい。つまり、仮に X さんが 123,000 回テストを受けてその平均を取った場合、その平均点を「本当の得点」とみなすことができる（言い忘れていたが X さんは超人なので 123,000 回テストを受けるぐらいは何でもない。しかし普通の人々にとって 123,000 回テストを受け続けるのは「干からびるほどテストを受け続ける」に等しいと思われる）。

続いて上の式の「偶然誤差」とは、偶発的にテストの結果に影響を与える要素である。仮に X さんが 123,000 回テストを受け続けたとしよう。57,002 回目にテストを受けた時には、X さんは風邪をひいていたかもしれない（超人でも風邪をひくのだ）。そのせいで 57,002 回目の点数はそれまでの回よりも低かったのかもしれない。古典的テスト理論では、上の式における「今回の得点」が含む偶然誤差にうまく対処することが重視される。偶然誤差が少ないほど、今回の得点は本当の得点により近くなるからだ。なお、偶然誤差に関する前提として、以下の４点がある。

**表２－１　偶然誤差の四原則**

| |
|---|
| ① 偶然誤差は正規分布する<br>X さんが 123,000 回テストを受けた場合、偶然誤差が各回のテスト得点の何点分であるかはその時々によって違うが、最終的に偶然誤差の散らばり具合を図示してみると、大半が平均付近に集まる。 |
| ② 偶然誤差の期待値（一種の平均値）はゼロ<br>X さんが 123,000 回テストを受け、エラーの分布の平均を算出するとゼロになる。 |
| ③ 偶然誤差は互いに関連していない<br>X さんのテスト得点は毎回違うが、違いが生じる体系的な原因はない。 |
| ④ 偶然誤差は本当の得点にも関連していない<br>X さんの得点が高いか低いかと、本当の得点との間には体系的な関係はない。つまり、「得点の高い時には必ず誰かが X さんの耳元で正解を囁いている」というような固定的な関係はない。 |

上記の偶然誤差の四原則は、古典的テスト理論の基礎を成すものである。また、偶然誤差の分布の標準偏差は測定の標準誤差と呼ばれ、数値が小さいほど偶然誤差から本当の得点までの距離が短くなる。つまり、測定の標準誤差は、あるテストが受験者の能力を識別するに際してどの程度役に立つのかを知るための一つの指標である。

しかし、ここで一つの問題が生じる。式１は個人を想定しているため、測定の標準誤差を算出するためには、一人の人に複数回同じテストを受験してもらわなければならない（標準誤差の算出には複数の数値が必要だからだ)。かくして我々はＸさんのような 123,000 回の連続受験をものともしない超人を連れてきて、テストの受験をお願いすることになる。だが世界はＸさんのような超人で満ちているわけではない（というよりもどちらかといえば非超人で満ちている)。非超人にとっては 50 回であっても同じテストを繰り返し受験するのは骨が折れる。そして式１が個人を想定していることで生じる問題がもう一つある。この式から算出される本当の得点や偶然誤差（そして測定の標準誤差）は、あくまでもある特定の個人のものである。それでは個人ではなく集団における測定の標準誤差を知りたい場合にはどうすればよいのだろうか……というような問題は、1970 年代に心理測定を専門とする研究者たちによって解消された。というのも、これらの研究者たちが式１は個人だけでなく、複数の人々から成る集団にも応用できることを数学的に証明したからである。ここに至って我々はめでたく複数の人々から成る集団にテストを一度だけ受験してもらうことで、その集団の本当の得点や偶然誤差と測定の標準誤差を算出することができるようになった。そして（一人の人による複数回のテスト受験ではなく）複数の人による一度のテスト受験の場合でも、上に挙げた偶然誤差の四原則は当てはまる。以下に、複数の人による一度のテスト受験の場合の古典的テスト理論における原則を記す。

**表２－２　複数人受験の場合の原則**

| |
|---|
| ① あるテストの測定の標準誤差は、テスト受験者全体に共通している<br>個々人によって標準誤差が異なるということはなく、「今回テストを受けた人々」の標準誤差は一般化できるため、「テストを（実際には受けていないが）受けるかもしれない人々」に対しても応用できる（ただし一般化するためには受験者が母集団から無作為に選ばれているなど、適切な手続が必要)。 |
| ② テスト項目が多いほどテストの信頼性が高くなる<br>受験者の数学的能力を知りたい場合、3 問の問題だけを解いてもらうテストよりも、50 問の問題を解いてもらうテストの方がよりよく受験者の数学的能力を測ることができる。 |
| ③ あるテスト受験者集団における本当の得点は正規分布する<br>受験者の得点の散らばり具合を図示すると、たいていの受験者の得点は平均点付近に集まる。 |

## ２　古典的テスト理論におけるテスト項目分析

古典的テスト理論を用いた分析によって、あるテストの項目（群）が受験者の能力や性格などを識別できるかどうかに加えて、ほかの項目（群）との関係を見極めるための結果を得ることができる。下記に、具体的な分析方法の例をいくつか挙げる。

**表２－３　各テスト項目の平均と分散の確認**

① ある項目の分散の値が低ければ、受験者間で得点にほとんど差がないということになる。つまりこの項目は使えない項目かもしれない。

② 一般的に言って、テスト項目の分散の値が高いほど、また、平均値が得点分布図の中心に近いほど、テスト項目の識別能力は高い。

**表２－４　難易度**

① 能力を測るテストでは、その項目を正解した人の割合によって各項目の難易度を示す。例）100 人中 65 人が正解した項目の難易度は 0.65 である。この指標では数値が高いほど問題が簡単であり、数値が低いほど問題が難しいことを意味する。難易度 0.5 の項目（半数が正解する項目）は最も受験者の識別能力が高い。反対に、難易度が 0（正解者ゼロ）、または１（全員正解）の項目は識別力がないということになり、このような項目を作成するのは純粋な時間と労力の無駄である。

② すべての項目の難易度が 0.5 である場合、それぞれの項目の識別力が高いというよりは項目同士が似ている可能性が高い（つまり同じような問題ばかり出題されている＝それならば問題を一問のみにすればよいと考えられる）。そこでテスト作成者は、全項目の難易度の平均は 0.5 になるように、しかし各項目の難易度は異なるようにテストを設計する。

**表２－５　項目得点と総得点の関連**

① ある一つの項目の得点がテストの総得点とどのように関連しているかを調べるためによく使われる統計的指標として、(a）積率相関係数、(b）点双列相関係数、(c）双列相関係数がある。

② 上記三つの指標を算出する際には、総得点から関連性を調べようとする項目の得点を除外した値（調整済総得点）を用いる。というのも、その項目の正誤に応じて得られた得点とテストの総得点は必然的に関連するものであるから、項目の得点を除外せずに総得点との相関係数を求めると、関連性がむやみに高く算出されてしまうという問題があるためである。

**表２－５　続き**

③ 上記三つの指標にはそれぞれ使うのに適切な場合とそうでない場合とがある。

(a) 積率相関係数は、調べたい項目と総得点の両方が連続変数の場合に使用する。連続変数とは、数えるのに永遠の時間を要する数値をいう。例）年齢を正確に数えようとすると、１歳と１日と１時間と１分と１秒と 0.1 秒と……というように永遠に終わらない。積率相関係数を用いる具体的な一例としては、年齢という項目と総得点の関連が挙げられる。

(b) 点双列相関係数は、項目への回答が「はい／いいえ」「反対／賛成」のように二つのうちいずれかに分類される場合に使用する。

(c) 双列相関係数は、項目への回答が人為的に二つに分類された場合に使用する。例）テストを実施した人々がある項目について「賛成／どちらかといえば賛成／どちらかといえば反対／反対」の四つの選択肢を「賛成（賛成とどちらかといえば賛成の総計)」と「反対（反対とどちらかといえば反対の総計)」に分類した場合[2]。

④ 上記三つの指標では、ある項目と総得点の関連が-1.00 から+1.00 の数値で表される。数値がマイナスの場合は、その項目はテストのほかの項目群と負の関連性があるということを意味する（つまり、例えばその項目に正解した受験者たちはほかの項目群は不正解である傾向がある)。ある項目と総得点の関連を調べる際に負の関連性が検出されるのは、たいていの場合あまりよろしくないものとみなされる。というのも、テストは普通、受験者の一つの能力や傾向をはかろうとするものだが、項目と項目の間で回答に反対の傾向があると「一つの能力や傾向をはかる」という大前提に疑問符がついてしまうからだ。また、関連性の数値が低い場合は、その項目がほかの項目群とあまり関連性がないということを意味する。さらに、項目―総得点間の関連性はその数値が 0.50 以上がよい（＝その項目はほかの項目群と適度に関連していて、しかも関連の方向が正であるので回答における逆の傾向もない）と考えられている。

## ３　古典的テスト理論の問題

以上でおおまかに古典的テスト理論の特徴と分析について述べたが、なぜテスト理論は古典的テスト理論で終わらなかったのだろうか。なぜ古典的テスト理論は現代的テスト理論という自身を凌駕する一連の理論群の出現を許したのか。これには古典的テスト理論が持ついくつかの致命的な弱点が関係している。一例として、エラーに関する弱点がある。第１節において、古典的テスト理論の特徴の一つとして偶然誤差について述べた。偶然誤差は確かにテストの得点に影響を与える。しかし、エラーは偶発的にしか起こらないのだろうか。実はそうではない。エラーには偶然誤差に加えて、体系的なエラー（以下「系統誤差」という。）も存在する。

系統誤差とは、テストの得点に影響を与える要素であり、なおかつテストを受ける人の性格や能

2　(c) の双列相関係数を使うべき場合に (b) の点双列相関係数を使ってしまうと、ある項目と総得点の関連が本来より低く算出されてしまうので注意。

力といった、簡単には変わらない要素を指す。例えば 123,000 回のテスト受験を難なくこなしてしまう X さんは、もちろん超人であるので並外れた忍耐力と集中力をもってテストにのぞむわけだが、X さんには窓の外から小鳥のさえずりが聞こえてくると途端に顔面蒼白になり、その時点で解いていた問題に必ず「崖」と解答してしまうという奇妙な癖がある。これは X さんが子どもの頃に山の中で美しい小鳥を追っていたところ、うっかり足をすべらせて数十メートルの崖から転落してしまい、血まみれになりながら自力で崖をよじ登って生還したという超人の子ども時代にありがちなトラウマ体験のためである。テスト受験会場のすぐ近くには大きな木があり、当然小鳥もやって来る。そして小鳥が無邪気にさえずると、それは X さんの耳にも届く。そのようなわけで X さんが 123,000 回テストを受け、仮に毎回 12 問目で小鳥がさえずったとすると、123,000 回とも X さんの 12 問目の「崖」という解答は不正解となる（言い忘れていたがこれは数学のテストであり、解答に崖が入り込む余地は微塵もない)。このような条件反射的要素に加えて、テストへの慣れも系統誤差の一例である。つまり、X さんが 123,000 回テストを繰り返し受験した場合、テストへの慣れから、回を重ねるごとに点数が上がっても不思議ではない。

このようにテストの得点には系統誤差も関係しているのだが、古典的テスト理論が関心を持っているのは偶然誤差にどううまく対処するかということである。そのため、古典的テスト理論では系統誤差にうまく対処できない。これに加えて古典的テスト理論にはほかの弱点もいくつかある（詳細は次節に譲る)。そこでこういった一連の問題に対処しようと、現代的テスト理論が立ち上がることになる。次節では、現代的テスト理論が古典的テスト理論の問題をどのように克服したかを述べる。ただし、古典的テスト理論に問題があるからといって古典的テスト理論には学ぶ価値がないということにはならない。古典的テスト理論は現代的テスト理論の礎であるため、古典的テスト理論を理解せずに現代的テスト理論を理解することはできないし、また、古典の名が示すとおり、古典的テスト理論には現代にも通じる普遍的な価値がある。

## 第２節　現代テスト理論（Modern Test Theory）の概要

古典的テスト理論に対して、その欠点を克服するために開発されたテスト理論が項目反応理論である。そのため、項目反応理論はそれ以前のテスト理論と対比的に紹介され、現代的なテスト理論として言及されることが多い。ちなみに、読者は「現代的」という言葉から相当に新しい理論という印象を持つかもしれないが、できたばかりの理論という意味ではない。第１節でも述べたように、テスト理論の重要な転換点であり、現代的なテスト理論の基礎を成す研究は 1950 年代から 60 年代にかけて発表されている。現代的と称するにはいささか古すぎるのではないかという読者の疑問をテスト理論研究者たちが甘受するかどうかは定かではないが、それまでのテスト理論とは異なるものという意味合いをもって、現代的なテスト理論（modern test theory）という名称で呼ばれることもある。本節ではこの現代的なテスト理論の一つである項目反応理論を概観する。項目反応理論の特徴と利点は、テスト項目の難易度と受験者の能力を分離できることである。これは古典的テスト理論の枠組みでは対応が難しい問題であったが、項目反応理論がまったく新しい考え方を導入したことによって取り組むことが可能となった。以下では、古典的テスト理論との違いに注意しつつ、項目反応理論の概略と利用上の注意点を述べる。

### １　古典的テスト理論の問題点

テストに求められる性質は信頼性である。ここでいう信頼性とは、受験者集団やテスト項目が異なっても、同じ能力を持つ受験者には同じ得点が割り振られることである。つまり、ある能力をもった受験者がテストを受ければ、何度テストを受けても同じ点数を与えなければならない。同時に、同じ能力の受験者が二人いれば、その二人には同じ点数を与えなければいけない。

テストの信頼性を検証するための方法論を提示するのがテスト理論である。テスト理論は伝統的に第１節１に示した（式１）に基づいていた。これはテスト得点を「受験者の真の得点」と「偶然誤差」に分割するものであった。この式に基づくテスト理論は一般的に古典的テスト理論と呼ばれている。

第１節の復習となるが、テスト得点から偶然誤差を取り除くには測定を反復しなければならない。それも生半可な回数の反復ではない。仮に超人的能力を持つ X さんが実在したとして、X さんに同じテストを際限なく何度も受けてもらうことができればよいのだが、非超人にとっては不可能である。そこで我々の非超人的世界において現実的にこの問題を解消する方法として、同一のテストを複数の受験者を対象として行うことでテストの信頼性を検証することが行われる。

古典的テスト理論では誤差をこのように扱っていたが、それでも対応できない問題が系統誤差であり、その代表例は標本依存性である。標本依存性とはどういう問題かというと、テストの信頼性を検証するために複数の受験者を集めたとしても、テストを実施して得た結果はそのとき集めた標本（受験者の集団）によって変わってしまうというものである。例えば、新しく作成したテスト問題をある集団に対して実施したときの平均点が、それ以前のテスト問題の平均点と比べて高くなったような場合を考える。このとき、古典的テスト理論では、新しいテスト問題に回答したこの受験者集団の平均的な能力が高かったと解釈する。つまり、「受験者の真の得点」が平均的に高いので、テスト得点の平均値もそれに応じて高くなったという解釈である。もし、その受験者たちの能力が

本当に高かったのであれば、平均点の高さは受験者の能力を正しく推定できたことになるため、理想的な状況といえる。しかし、これと異なる解釈も可能である。もしかすると、受験者の平均的な能力は低かったのだが、テストの難易度も低く、結果的に正答しやすかったのかもしれない。つまり、古典的テスト理論では問題の難易度と受験者の能力を切り分けることができないのである。測定結果を「本当の得点」と「偶然誤差」に分割する古典的テスト理論では標本（受験者集団）に依存した影響を取り除くことは困難である。

標本依存性の他にも、古典的テスト理論では解決が難しい問題が指摘されている。例えば、テスト得点の範囲の問題である。学力テストでは得点の最小値は0点であり、最大値は満点（通例では100点）である。これは見方を変えれば、0点から100点の範囲内でしか受験者の能力を評価できないということでもある。仮に超人Xさんが100点を取ったとしても、Xさんの能力が100点に相当するとは必ずしも言えない。Xさんの実力は10,000点くらいなのかもしれないが、テストの得点には上限があるため、実際よりも過小評価されていることになる。このように得点によって受験者の能力を区別しようとした場合、得点の範囲を超える能力は評価ができない。

以上のような問題は古典的テスト理論の枠組みでは解決が難しい。特に、標本依存性の問題はテストの難易度と受験者の能力を分離することの重要性を指摘している。項目反応理論は（式1）から離れることでこの問題に対応した。

## 2　項目反応理論

古典的テスト理論の問題点を克服するために発案されたテスト理論の一つが項目反応理論である（Load, 1952; Rasch, 1960）。ここでいう項目（item）とはテストの問題や質問紙調査の質問項目のことであり、それに対する反応（response）を説明することを目的としているため項目反応理論と呼ばれる。項目反応理論は古典的テスト理論とは異なり、受験者の「真の得点」を仮定しない。その代わりに、受験者の能力を一つのパラメーターで表し、このパラメーターの大小によってテスト問題に正答する確率が変わるという考え方をとる。このパラメーターの値はいわば受験者の「真の能力」であり、ギリシア文字の$\theta$（シータ）を使って表される。つまり、真の能力$\theta$が低ければテスト問題に正答する確率は低く、$\theta$が高ければ正答する確率が高くなるといったように表現する。

テストの総合得点や平均点はテストの結果から即座に計算できるが、パラメーター$\theta$を特定するためには複雑な数値計算を行う必要がある。そのため、項目反応理論を実施するには統計ソフトウェアを用意しなければならない。統計ソフトウェアの使用法や具体的な計算方法は本稿の範疇を超えるため解説は省略するが、近年は包括的な解説書が多数出版されており（de Ayala, 2009; 加藤・山田・川端, 2014; 豊田, 2012）、より詳細を求める場合は参考にすることができる。

### （1）項目特性曲線

項目反応理論では受験者の真の能力$\theta$とテスト問題に正答する確率の関係を項目特性曲線によって表す。項目特性曲線はロジスティック関数[3]を使った曲線として描かれることが一般的である

[3] ある結果が2値変数で、どちらに分類されるかを予想するのがロジスティック回帰分析である（なお、ロジスティック回帰分析は2値変数だけでなく、3値以上の名義尺度や順位尺度も扱うことができるが、ここでは詳細を割愛する）。回帰分析は原因と結果の関係

が、その意味を実感しやすいように、まずは直線（一次関数）として考えることから始める。

例えば、図２－１の左のグラフを見てほしい。これは、ある問題の項目特性曲線を描いたものである。この関数は右上がりの形をしており、θ の値が大きくなるほどこの問題に正解する確率が高くなることを意味する。θ は無限に小さい値（－∞）から無限に大きい値（＋∞）までを含んでいるが、ここでは－３から＋３範囲を示している。

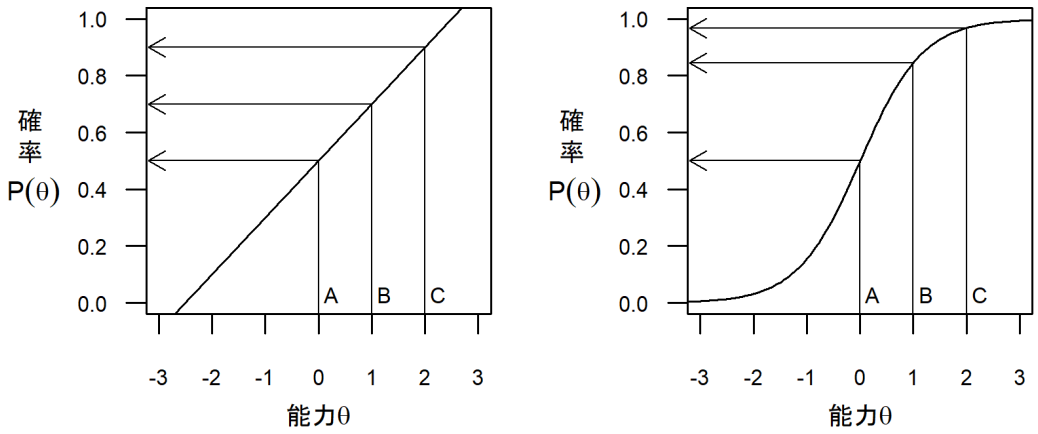



**図２－１　項目特性曲線**

ここで、ある３名の受験者たち（A さん、B さん、C さん）を例に考えてみよう（ちなみにこれら３名は X さんとは違って普通の人々である）。A さんの真の能力 θ の値はすでに判明しているものとして、その値が０だったとする。能力が０と聞いて奇異な印象を受けるかもしれないが、これは能力がまったくないという意味ではない。ここで重要なのは、θ の値が０である A さんがこの問題に正答する確率はどれほどなのかである。グラフを見ると、θ が０のとき確率は 0.5（50%）である。つまり、A さんはこの問題に 50%の確率で正答する。同様にして、受験者 B さんは θ が１であり、正解する確率は 0.7（70%）である。受験者 C さんは θ が２であり、正解する確率は 0.9（90%）である（能力が常人を超える X さんの場合は θ が極めて大きくなるのでグラフの範囲からはみ出ている）。つまり、受験者３名の中では C さんがもっとも優秀である。このように項目特性曲線によって、受験者の能力と問題に正解する確率を表現することが可能となる。

もちろん、この例のように初めから受験者の真の能力 θ がわかっているという場面はありえない。テストを実施した直後にわかることは「受験者が問題に正答したかしなかったか」であって、「受験者がこの問題に正答する確率」ではない。だから、ある問題に３名の受験者が正答したからといって、そのことだけでは彼らの能力が同じなのか違うのかはわからない。しかし、項目反応理論を用

---

が“最も当てはまりの良い直線（回帰線）”で表される。だが、回帰線を推定する際に、原因（x 軸）の値次第で、予測値＝結果（y 軸）が１を超えたり、０を下回ることもあり、推定する線形回帰が２値データ（あり・なし等）の分析に適さないといった問題が生じる。そこで、ロジスティック関数はこれらの問題を解決するための工夫が施されている。具体的には１か０かを予測するのではなく、y が１を選択する確率 p を予測し、その確率から個体が１か０のどちらかに分類されるかを判定する形となっている。確率は０～１の範囲を取るので、線形回帰の問題はこれで解決することができる。だが、予測対象を確率にしても、推定が直線関係のままだと x の値によって１以上、０未満が生じてしまう。そこで、ロジスティック関数では説明変数 x と選択確率 p の間に S 字曲線の関係を想定し、最大・最小の値が１以上、０以下にならず頭打ち・底打ちになるようにして、扱いやすい関数としているのである。

いて受験者それぞれの能力値θを特定すれば、それによって受験者を区別することができるようになる。

ここまでの説明では、真の能力θと問題に正答する確率が直線の関係にあるものと想定していた。しかし、項目反応理論が実際に利用するのは直線（一次関数）ではなくロジスティック関数による曲線である。その意図として実用的な利点を指摘できる。第一に、一次関数を用いた場合にはθの値によって確率が０を下回ったり１を上回ったりすることがある。図２－１左の直線を例に考えると、もしθが−3の受験者がいたとしたら、問題に正解する確率は−10%になってしまう。これは気象予報士からにこやかに「明日雨が降る確率は−10%です」と宣言されるようなものだ。果たして明日雨は降るのだろうか。あるいは−10%ということはだいぶ空気が乾燥しているということか。それとも我々の気づかないうちに気象予報のシステムが変わり、プラスだけでなくマイナスの確率も使われるようになったのかと混乱するに違いない。つまり、数値が確率として意味を持つためには、グラフの縦軸が０（０%）から１（100%）の範囲に収まる必要がある。第二に、一次関数は正答する確率が一定の割合で上昇していくことを表現しているが、これは現実に即さない。ある問題に正解するために一定の能力が必要であるなら、その能力に達しない受験者はほとんど正答できないが、その能力を超えている人たちは大体正答するはずだ。

こういった問題はロジスティック関数を利用することで解決する。ロジスティック関数は入力した値（この場合はθ）を変換し、０から１の範囲に割り振るための操作だと理解してほしい。そのような変換の操作を施したものが図２－１の右のグラフである。一次関数では直線だったのに対し、ロジスティック関数では曲線になっている。このグラフでもθは−∞から＋∞の範囲の値を取ることができるが、縦軸の確率は０から１の範囲に必ず収まるようになっている。

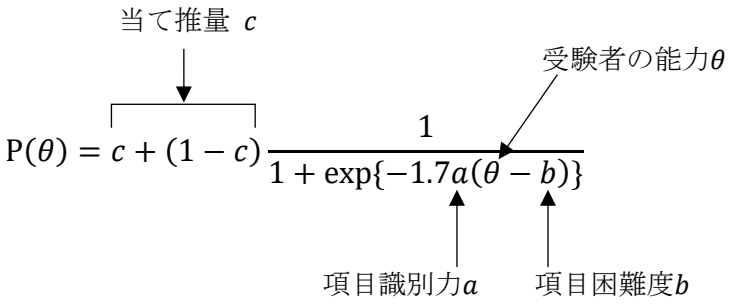
ロジスティック関数を用いればθの値に制限はなくなり、θが無限に大きくても無限に小さくても、確率を０から１の範囲に収めることができる。これは、一般的なテストの得点に付随する難点に対しても有効な解決策となっている。学力テストは得点の最小値が０点であり、最大値が満点（通例 100 点）であることが多いが、最小値０点未満の能力を評価できないのと同時に、最大値 100 点を超える能力を評価できない。受験者の能力を総合得点ではなくパラメーターθで表現すればこういった問題は解消される。

さらに、ある一定の能力を持った受験者は正答できて、それに達しない受験者は正答できないという状況も、ロジスティック関数を用いれば表現することができる。ロジスティック関数によって表現されたグラフでは傾きが急峻になっている部分があり、この前後で問題に正解できる確率が大きく変化する。つまり、このグラフの場合には真の能力θが０を超えている受験者は正答できるが、０に達しない受験者は正答できないことを表している。

これがロジスティック関数によって項目特性曲線を描く利点である。問題の正誤によって受験者の能力を測定するというテストの目的に適っている。そして、項目特性曲線はそれぞれの問題ごとに計算することができる。つまり、問題の数と同じ数の項目特性曲線を描くことができ、それぞれの問題に正答する確率を受験者の真の能力から記述することが可能となる。これが意味するのは、テストの結果からテストの難易度と受験者の能力を分離できるということである。

## ２　三つのモデル

項目反応理論で用いられる数学モデルはさまざまであり、それに合わせて項目特性曲線は形を変える。ここでは利用されることの多い三つのモデル（１PLM、２PLM、３PLM）を紹介する。これら三つのモデルはロジスティック関数によって構成されているが、その中に含まれるパラメーターの数によって異なるモデルとして扱われている。そのパラメーターとは、項目識別力パラメーターa、項目困難度パラメーターb、そして当て推量パラメーターc の三つである。モデルの全体像を数式で表現すれば以下のようになる。




左辺に置かれた「P（θ）」という記号は、受験者の真の能力θがわかれば、その受験者が問題に正解する確率 P がわかる、ということを表している。右辺に示した式の本体は非常に複雑に見えるが、グラフとして描いてしまえば曲線になることは先に示したとおりであり、ここでは式の中に三つのパラメーター（a と b と c）があることを確認してもらうだけでかまわない。そして、三つのパラメーターがどのように曲線の形を定めるかを以下で詳説する。

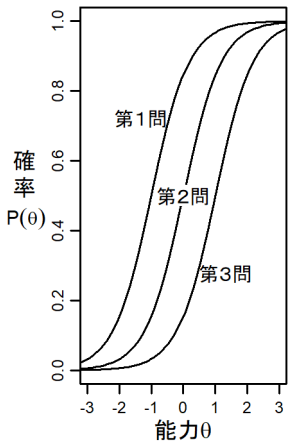
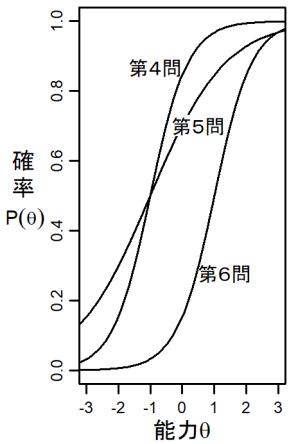
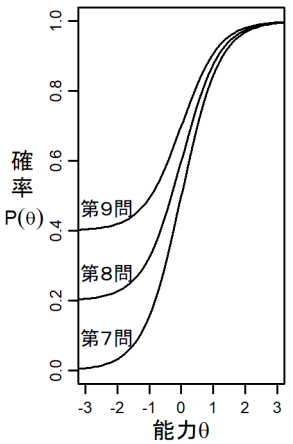
| | 1 PLM | | | 2 PLM | | | 3 PLM | | |
|---|---|---|---|---|---|---|---|---|---|
| |  | | |  | | |  | | |
| | 第１問 | 第２問 | 第３問 | 第４問 | 第５問 | 第６問 | 第７問 | 第８問 | 第９問 |
| 項目識別力 a | | | | 1.0 | 0.5 | 1.0 | 1.0 | 1.0 | 1.0 |
| 項目困難度 b | －1.0 | 0.0 | 1.0 | －1.0 | －1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 当て推量 c | | | | | | | 0.0 | 0.2 | 0.4 |

**図２－２　三つのモデル**

１パラメーター・ロジスティック・モデル（１PLM）は三つのパラメーターのうち、項目困難度 b のみを含むモデルである。項目困難度 b は問題の難しさを表しており、グラフの中心の位置を変化させるパラメーターである。図２－２には１PLM として三つの問題の項目特性曲線を例示した。

第１問、第２問、第３問のグラフは項目困難度 b をそれぞれ－1.0、0.0、＋1.0 とした場合の項目特性曲線である。項目困難度 b が大きくなると項目特性曲線は右に移動する。すると、ある能力値 $\theta$ をもつ受験者にとって正答する確率は小さくなることになるわけだから、問題の難しさが上がっていく様子を表している。実を言えば、先に紹介した項目特性曲線（図２－１の右の図）はこの１PLM について、項目困難度 b を 0.0 とした場合のグラフを描いたものであった。

２パラメーター・ロジスティック・モデル（２PLM）は１PLM に項目識別力パラメーターa を加えたモデルである。項目識別力 a はグラフの傾きを変化させるパラメーターである。第４問と第５問のグラフは項目識別力 a をそれぞれ 1.0 と 0.5 とした場合の項目特性曲線である（グラフの位置を示す項目困難度 b は－1.0 に固定した)。項目識別力 a が大きくなると、グラフの傾きは急になる（第４問の項目識別力 a は 1.0 であり第５問の 0.5 より大きいので、グラフの傾きは急になる)。すると、受験者の能力のわずかな違いが問題に正解する確率を大きく変化させることになる。つまり、項目識別力は受験者の能力の違いをどれだけ区別できるかの指標である。ところで、１PLM がそうであったように２PLM にも項目困難度 b が含まれており、グラフ全体の位置を移動することができる（第６問)。第６問は項目識別力 a が 1.0 であるため傾きは第４問と同じであるが、項目困難度 b が 1.0 になっており、グラフ全体が右に移動している。

３パラメーター・ロジスティック・モデル（３PLM）は２PLM に当て推量パラメーターc を加えたモデルである。当て推量 c は問題に偶然正解してしまう可能性を考慮したもので、グラフの底を持ち上げるパラメーターである。第７問、第８問、第９問のグラフは当て推量 c をそれぞれ 0.0、0.2、0.4 とした場合の項目特性曲線である。グラフの底が高いほど、偶然正解する確率が高い問題であることを意味している。

最後に三つのモデルの特徴についてまとめると、１PLM は左右の移動のみであり（項目困難度 b)、２PLM は左右の移動に加えてグラフの傾きの変化が加わる（項目識別力 a)。そして、３PLM ではグラフの底を持ち上げることになる（当て推量 c)。項目特性曲線をロジスティック関数によって表現し、そこに三つのパラメーターを組み込むことによって、問題ごとの特徴をこのように記述することが可能となる。

## ３　留意点

項目反応理論を実用する上で留意すべき事項に言及する。

### （１）使用するモデルの決定

項目反応理論の数学モデルとして三つのモデルを紹介したが、どれを使えばよいかという問題がある。一見すると三つのパラメーターをすべて含んだ３PLM がよさそうであるが、それが常に最適とは言えない。モデルの推薦順位を表にまとめたものが加藤・山田・川端（2014）に掲載されている。この表は三つのモデルについて「推定の正確度」や「最小標本数」などの基準によってモデ

ルの推薦順位を評価している。例えば、３PLM は「与えてくれる情報」については第一位であるが、「推定の正確度」や「最小標本数」については第三位となっている。総合的に判定すると、推薦順位の第一位は１PLM であり、続いて２PLM、最後が３PLM となっている。

**（２）必要なデータ数**

項目反応理論を用いたデータ分析を行うためには多くの参加者数が必要と言われる。参加者数はテストに含まれる問題の数やどのモデルを利用するかによって変わるため、画一的な基準は存在しない。加藤・山田・川端（2014）を参考にすると、「最小標本数」は１PLM で 100～200 程度、２PLM で 200～400 程度、３PLM で 1000～2000 程度となっている。また、当障害者職業総合センター障害者支援部門が実施した専門家ヒアリングによれば、「問題数の 10 倍」を目安とする見解もあるという。

**（３）一次元性の確認**

項目反応理論を適用するために必要な前提は一次元性である。項目反応理論が扱う「真の能力θ」はあくまで単一の能力（例えば英語の能力）であって、複数の能力が混ざったもの（例えば英語の能力と英語圏文化の知識）であってはいけない。そのため、テストが測定している能力が実際に単一の能力であるかどうかを確認する手順が必要となる。これには通常、因子分析（factor analysis）が利用されることが多い。一次元ではなく二つ以上の能力を想定した多次元の項目反応理論もモデルとしては提唱されているが、実用上の課題があり、応用段階にはないという（光永, 2017）。

**（４）局所独立の仮定**

局所独立の仮定とは、ある問題への正誤と他の問題への正誤が関連してはいけないことを指す。これに反する状況の具体例は、第１問の答えを使って第２問を解くような場合である。このとき、第１問の答えがわからなければ第２問に正解しようがない。局所独立を満たさない状況はさまざまであり、テストを実施するまではわからない場合も多く、テスト実施後の注意深い確認が必要となる。

**（５）等化**

項目反応理論の利点は問題の難易度と受験者の能力を分離できることにあった。だからといって、即座に古典的テスト理論の難点である標本依存性を回避できるわけではない。その時の受験者集団に依らず問題の難易度や受験者の能力を評価するには、テスト作成段階での「等化（equating）[4]」と呼ばれる作業が必須となる。等化の作業が古典的テスト理論よりも容易であるという点が項目反応理論の発展の一因とされる。

[4] 等化とは、各試験の各々の版における得点尺度を共通得点尺度へ調整する操作のことである（村木, 2011）。池田（1994）は、テスト得点の等化について「同一の能力や特性を異なったテストで測定するとき、それらのテスト得点を比較できるか、という問題、あるいは、二つの異なった能力や特性を比較可能（comparable）な形で表現するにはどうしたらよいかという問題は、テストの利用者にとって直接関係する重要な問題である」と述べている。特に経年比較が必要な学力テストなどでは、等化のプロセスが必須といわれる。たとえば、ある科目について月１回、同じ項目数のテストを実施して得られた結果が毎回同じ得点であっても、それらは必ずしも同じ意味であると保証されていない。そこで得られた数値に変換を施して同じ尺度上に並べる等化のプロセスによって、得点の比較が可能となるのである。

## 第３節　現代テスト理論を用いた事例

ここまで、古典的テスト理論と現代テスト理論の考え方について概観してきた。先述のとおり、古典的テスト理論の考え方は、統計的な手法として広く利用されている一般的な理論である。

一方で、我々の最大の関心は、現代テスト理論の考え方を、職業リハビリテーションのアセスメントにどのように活かせるか検討することであるが、まずは、現代テスト理論を用いた先行事例を見ていくこととしたい。

一般的に多くの試験は PBT（Paper-Based Testing)、つまり筆記試験によって実施されることが多いが、大規模試験において現代テスト理論の考え方を実装するにあたっては、CBT（Computer-based Testing : 以下「CBT」という。）の手段を用いることが重要な条件となっているようである。

国内外の状況については、大学入試センター（2021）による報告が詳しく、それによれば我が国の CBT は、主に資格検定試験での利用が代表される（表２－６)。

国内試験の多くが全都道府県単位で、随時あるいは定期的に実施されている。その年間受験者数は数万人～100 万人と大規模である。また、いずれも実施経費として受験料が設定されている。

海外においては、OECD（経済協力開発機構）が行う学力調査（PISA: Programme for International Student Assessment）において CBT が導入されている。我が国でも、文部科学省による全国学力・学習状況調査についても CBT 化に向けた検討が行われているという（大学入試センター，2021)。

**表２－６ 国内外におけるＣＢＴの実施例（「大学入試センター，2020」より引用改編）**

| 試験名称 | | 試験場 | 年間受験者数 | 受験料 | 実施回数 | 試験時間 | 回答方法 | 結果発表 |
|---|---|---|---|---|---|---|---|---|
| 国内 | 医療系大学間共用試験 | 47 都道府県／50 以上（所属大学で受験） | １万人以上 | 25,000 円 | 随時 | 360 分 | 多肢選択 | 10 日後 |
| | 英検Ｓ－ＣＢＴ | 47 都道府県／約 100 | 約 400 万人 | 4,500～12,600 円 | 原則毎週土日 | 90-135 分 | 音声吹込み・多肢選択・記述 | 3－4 週間後 |
| | ＳＰＩ | 47 都道府県／約 50 | 200 万人以上 | 5,500～6,500 円（企業負担） | 随時 | 約 70 分 | 多肢選択 | 即時採点 |
| | ＩＴパスポート | 47 都道府県／約 120 | 10 万人以上 | 5,700 円 | 随時 | 120 分 | 多肢選択 | 2-3 時間後 |
| | 損害保険代理店試験 | 47 都道府県／約 190 | 約 100 万人 | 1,900～3,900 円 | 随時 | 40－120 分 | 多肢選択 | ３営業日後の翌日 |
| 国外 | TOEFLiBT® テスト | 世界 150 か国以上<br>日本 27 都道府県／約 70 | 非公開 | 245USD（約 26,000 円） | 月３－６回<br>年 45 回以上 | 約 180 分 | 音声吹込み・多肢選択・記述 | 約６日後 |
| | TOEIC® Ｓ＆Ｗ | 12 都道府県／約 40 | ３万人以上 | 10,450 円 | 年 16 回 | 80 分 | 音声吹込み・記述 | 35 日以内 |
| | ISAT | 日本２都府／２ | - | 320USD（約 34,000 円） | 年４回試験期間 1―2 週間 | 約 180 分 | 多肢選択 | 約２週間 |
| | GRE | 世界 160 か国以上,日本２都府/２ | 約 65 万人 | 205USD（約 22,000 円） | 随時 | 約 225 分 | 多肢選択・記述 | 約 10-15 日後 |

また、大学入試センター（2021）の同報告書では、現在の大学入試である共通テストを、紙と鉛筆ではなく、パソコンやネットワークなどを利用して実施する（すなわち、CBT-IRT（Computer-based Testing -Item Response Theory）による実施）場合に生じる課題や対応について現在も検討が進んでおり、今後も調査研究を行うことが重要であるとしている。

さて、大学入試ほどに大規模でなければ、CBT-IRT 又は項目反応理論はテスト開発に活かせないのだろうか。論文検索サイト（Cinii,J-stage,Pubmed）において、「項目反応理論」のキーワードで検索したところ、ヒット数は Cinii で 626 件、J-stage で項目反応理論では 1, 071 件（Item Response Theory では、590, 509 件）、Pubmed では「Item Response Theory」として検索し、4, 892 件（2021 年 12 月 21 日時点）のヒット数があった。研究分野は心理学、教育心理学、行動計量学、工学、体育、医学、社会心理学、経営学、数学、情報学など多岐にわたった。そこで、項目反応理論の利用方法を概観するため、検索結果の内、2000 年以降の論文の中から職業リハビリテーションの分野に近接する分野で、対象者の特性評価に用いられるツールや尺度の品質改善をテーマとしたものの一部を表２－７に取り上げた。

**表２－７　尺度開発・ツール開発に項目反応理論の利用が確認された論文概要**

| **医学分野での利用例** |
|---|
| 【作業療法士の職業的アイデンティティ自己評価尺度における項目特性と構造的妥当性－若手作業療法士を対象にした検討一，鈴木渉・藪脇健司・中本久之／作業療法・38 巻 4 号，2019 年】<br>身体障害領域及び高齢者領域に勤務する臨床経験が 1 ～ 3 年目までの作業療法士を対象とし、作業療法士の職業的アイデンティティ自己評価尺度（PI 尺度）の項目反応理論を用いた項目分析を行い、構造的妥当性を検討した。PI 尺度は、中本らが回復期病棟に勤務する作業療法士用に改編したものに、加筆・修正を加えて使用した。結果として、PI 尺度は、作業療法士に独自性があると思う程度が平均的な回答者に対する測定精度が高く、4 因子 29 項目の 2 次因子モデルで適合していたことから、身体障害領域および高齢者領域の作業療法士を対象として、広く使用できる尺度であることが明らかとなった。 |
| 【日本版 WHO-QOL-26 の構成妥当性の再検討，折笠秀樹・横山奈緒美・上馬場和夫／臨床薬理・35 巻 1 号，2004 年】<br>世界保健機関（WHO）による一般向け QOL 質問票（WHO-QOL-26 日本語版）は、1997 年に出版された。それは四つの領域（身体 7 問、心理 6 問、社会 3 問、環境 8 問）からなっており、それぞれ 0~100 点が付けられる。妥当性の詳細は検討済だが、その中で思うような結果が得られなかった構成（領域間）妥当性について、さらに詳細な統計分析を行うことが研究目的である。1,013 名のうちデータの揃った 805 例を用いて因子分析、主成分分析、変数クラスター分析、項目反応理論、クロスバリデーションを実施。因子分析の結果、社会領域は三つ揃ったが、他領域は外れた項目があり、身体領域は構造が見られなかった。クロスバリデーション、変数クラスター分析においても同様の結果であった。項目反応理論を用いて判別力を見たところ、領域ごとの寄与が判明し、寄与の低い 8 問を除いて因子分析を実施し構成は明瞭となったが、それでも心理領域と環境領域は分かれなかった。質問票の構成に問題点が指摘されるとともに、日本人では領域間の境界が鮮明でないことが示唆された。 |
| **教育・心理分野での利用例** |
| 【項目反応理論を用いたストレス測定尺度短縮版構成の試み，岩田昇、菊池賢一／日本心理学会大会発表論文集 84 PR-020.2020】<br>米国国立労働安全衛生研究所職業性ストレス調査票に対する労働者 2,428 名（男 2,224，女 203，不明 1 ）のデータに多値型の項目反応理論モデル（Modified Graded Response Model; Muraki, 1992）を適用し，ストレス尺度の構成項目の項目特性パラメーターを推定した。求めた項目パラメーターに基づき，各項目の情報量を算出し，大きい順に配置し，尺度情報量の 7 割程度の情報が得られる項目を選抜した。この方法により，量的労働負荷（11 項目）の場合， 3 項目で全体の 75%の情報が得られ，短縮版作成の簡便な方法論となることが明らかとなった。 |
| 【青年のソーシャルスキルにおける汎状況的なスキルと具体的な対人場面でのスキルとの関連性の検討，吉良悠吾・尾形明子・上手由香／教育心理学研究，68，11-22.2020】<br>本研究は、ソーシャルスキルの階層性を考慮し、認知や情緒面のスキルも測定できる「成人用ソーシャルスキル自己評価尺度」が青年に適応可能であることを確認した上で、項目反応理論を用いて短縮版尺度を作成し、その短縮版尺度を用いて、具体的な対人スキルとそれらを発揮する基となる汎状況的な認知・情動・行動スキルであるコミュニ |

ケーション・スキルとの関連性を検討することが目的であった。多母集団因子分析によって、青年のソーシャルスキルを同様の因子構造で測定できることを確認した上で、項目反応理論を用いて 35 項目であった項目数が 20 項目となる短縮版尺度を作成した。また、階層的重回帰分析の結果、対人スキル発揮のためには、自分の意思を相手に伝えるための行動スキルだけでなく、認知や情動面のスキルも重要であること、その関連性は対人スキルの種類によって異なることが示された。したがって青年のソーシャルスキルを高めるためには、対人スキルの種類に合わせて、認知や情動面のスキルを含めた訓練が有効であることが示唆された。

**障害分野での利用例**

【3 歳児向け自閉症スペクトラム障害スクリーニングテストの開発，中村知靖、大神英裕、実藤和佳子、山下洋／日本心理学会大会発表論文集 84（0）, PD-066-PD-066, 2020】

自閉症スペクトラム障害スクリーニングテスト（M-CHAT）は，16～30 ヶ月児を対象としたテストだが，早期発見を継続して行うためには高い年齢を対象としたテストの開発も必要である。そこで当該研究では， 3 歳児を対象にスクリーニングテストを開発した。項目は，コミュニケーションに関する 10 項目，興味の限局と常同的・反復的行動に関する 4 項目，運動機能に関する 1 項目であった。 3 歳児 384 名を対象とし， 3 歳児健診会場，子育て支援センター，こどものこころの相談室においてテストを実施した。項目反応理論によって分析した結果，コミュニケーション能力（ 6 項目）と興味の限局と常同的・反復的行動傾向（ 4 項目）の尺度構成が可能であることが分かった。また，対象児の能力値を算出し，定型（367 名），自閉症（15 名），高機能自閉症（ 2 名）の群毎に平均を求めたところ，コミュニケーション能力では，平均の大きさが，定型＞高機能自閉症＞自閉症の順となり，興味の限局と常同的・反復的行動傾向では，高機能自閉症＞自閉症＞定型の順となった。

【自閉性スペクトラム障害の障害特性に関する知識尺度 （Literacy Scale of Characteristics of Autistic Spectrum Disorder : LS-ASD）の開発，酒井貴庸・設楽雅代・脇田貴文・金澤潤一郎・坂野雄二・園山繁樹／自閉症スペクトラム研究，12（1）、19－28，2014】

発達障害の障害特性に関する知識の程度（知識度）に着目し、自閉性スペクトラム障害（Autistic Spectrum Disorder：ASD）の障害特性知識尺度（LS-ASD）の開発を試みた。なお、本研究では、高等教育機関における発達障害関連の相談の中で相談件数が最も多く、気分障害や不安障害との合併といった二次障害のリスクが特に高い ASD に焦点をあてた。LSASD は、能力測定の分野においてさまざまな成果を上げている項目反応理論（IRT）に基づいて開発された。学生、医療福祉従事者、教師の 825 名の回答データについて分析し、最終的に 44 項目において内容的・基準関連妥当性、信頼性が確認された。IRT に基づいて開発された尺度であるため、LS-ASD は十分な信頼性を維持しつつ、回答者の知識度に合わせた項目を抜粋しての使用や Computer Adaptive Test としての使用可能性をもつ尺度となった。

【日本版 Vineland- Ⅱ適応行動尺度の開発　適応行動尺度の項目分析と年齢による推移，谷伊織・伊藤大幸・行廣隆次[他]／精神医学，54（9）889-898, 2012】

本研究は、全年齢に適用可能であり国際的に広く利用されている Vineland Adaptive Behavior Scales, Second Edition の日本版の標準化に関する研究の一環として、適応行動尺度の項目分析と年齢による推移を検討した。項目分析の結果、いずれの項目も十分な項目－合計相関を示した。さらに項目反応理論（IRT）によって項目の再配置を行い、打ち切りルールを適用し、尺度得点を求めた。打ち切り前後の得点の相関を検討したところ、いずれの下位尺度についても、99 程度の強い相関がみられた。Vineland- Ⅱ日本語版の適応尺度の年齢推移を調べた結果、いずれの尺度も全年齢の発達に対応していることが明らかとなった。これらの結果は適応行動尺度の高い信頼性・妥当性を示している。

表２－７に上げた取組例はごく一部であるものの、いずれの調査においても、項目のもつ困難度や被検者の特性値を項目反応理論によって推定し、既存の構成を再検討したり、項目選定を行っていた。このような様々な調査結果からは、従来の古典的テスト理論の問題として前節でも指摘された「調査対象の違いによって、測定の精度が異なる」という問題に対して、項目反応理論を用いることで対応可能となり、よりきめ細やかに測定精度を評価することが可能となることが示されている。

それでは、これらの取組を参考にしつつ、職業リハビリテーションで用いられるアセスメントツールにおいても項目反応理論が応用できるかどうかについて、第３章において具体的に検討することとしたい。

**【文献】**


Allen, M.J., and Yen, W.M.（1979）. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

別府正彦（野口裕之・柴山直・熊谷龍一監修）(2015),「新テスト」の学力測定方法を知る　IRT 項目反応理論入門　基礎知識からテスト開発・分析までの話，河合出版.

Crocker, L., and Algina, J. （1986）. *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

de Ayala, R. J.（2009）. The Theory and Practice of Item Response Theory. The Guilford Press.

独立行政法人大学入試センター (2021), 大規模入学者選抜における CBT 活用の可能性について (報告) 令和３年３月 24 日.

池田央（1994)，現代テスト理論，朝倉書店.

加藤健太郎・山田剛史・川端一光 （2014）． R による項目反応理論 オーム社.

Kline, J.B.T.（2005）. *Psychological Testing: A Practical Approach to Design and Evaluation*. CA: SAGE Publications.

栗原伸一（2021）．入門統計学第２版　－検定から多変量解析・実験計画法・ベイズ統計学までー，オーム社.

Load, F. M.（1952）. A theory of test scores. Psychometric Monograph, No. 7 .

光永悠彦（2017）．テストは何を測るのか: 項目反応理論の考え方 ナカニシヤ出版.

村木英治（2011）．項目反応理論，朝倉書店.

Rasch, G.（1960）. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

豊田秀樹（2012）．項目反応理論[入門編] 第 2 版 朝倉書店.

# 第３章

# テスト理論の応用可能性の検討

# 第３章　テスト理論の応用可能性の検討

本章では、職業リハビリテーションの分野で用いられるアセスメントツールにおいて、項目反応理論をどのように利用できるか検討するため、実在のアセスメントツールの中でも、障害者職業総合センター（2019）が開発した WMS の課題のうち『社内郵便物仕分』を参考として、項目反応理論を用いた分析までのプロセスを提示していく。

## 第１節　MWS『社内郵便物仕分』を対象とした探索的解析
### －シミュレーションデータ作成と項目反応理論による検討－

### １ MWS『社内郵便物仕分』の構造

『社内郵便物仕分』は、MWS の中でも難度の高い課題であり、既存のワークサンプルでは評価が難しい、作業能力の比較的高い障害者に対する支援ニーズに応じて開発された。

『社内郵便物仕分』は、架空の株式会社 JEED 宛てに届いた郵便物を、宛先の部課等に仕分ける作業である。郵便物の仕分に当たっては、「サブブック」という名称の資料（ファイル綴り）の中にある「仕分のルール」に基づき、「組織図」、「社員名簿」、「あいうえお索引」の各資料を参照しながら、個人の所属先や速達・親展の有無などを確認し、仕分棚を模した部署別の仕分ボックス・仕分フォルダーに郵便物を仕分けるという作業過程が想定されている。

『社内郵便物仕分』の課題構成としては、郵便物１通を仕分ける作業を１試行とし、１ブロック当たりの試行数は 20 試行となっている。

レベル設定は表３－１のとおりである。新規課題は、「情報処理の複雑さや認知的負荷（一時的に記憶しなければならない情報量、注意配分数、確認箇所数など）を上げることで、難易度を高めた」（障害者職業総合センター，2019）とされており、レベルが高まれば、適用ルールが拡大し、処理すべき情報量が増加すること等によって難度が高まる構成となっている。

**表3－1　レベル別にみた郵便物の宛名面に記載される情報と適用ルールの範囲**

| | 郵便物の宛名面に記載される情報 | | | | | 適用ルールの範囲（仕分のルールで指定される仕分先ボックス） | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| レベル | 部名 | 部課名 | 部課名＋個人名 | 個人名のみ | 速達・親展 | 各部代表フォルダー | 部課フォルダー | 要確認ボックス | 速達・親展ボックス | 本部外所属転送ボックス（付箋付） | 本部内他部課所属現部課フォルダー（付箋付） |
| 1 | ○ | | | | | ○ | | | | | |
| 2 | ○ | ○ | | | | ○ | ○ | ○ | | | |
| 3 | ○ | ○ | ○ | ○ | | ○ | ○ | ○ | | | |
| 4 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | |
| 5 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

『社内郵便物仕分』を含むMWSの各課題には、アセスメント・体験版としての機能を持たせた「簡易版」(20試行）と、トレーニングとしての利用等を想定した「訓練版」（5レベル各30ブロック）が用意されている。今回の分析に当たっては、先行研究（障害者職業総合センター，2019）において比較的多くのデータが得られている「簡易版」を取り上げることとした。「簡易版」は、20試行の中に、表３－１の、レベル１～レベル５の全レベルの内容が含まれる構成となっている。

## ２　分析に当たっての問題

『社内郵便物仕分（簡易版)』に関して、公開済みで利用可能なデータは、障害者職業総合センター(2019）による記述統計量（正答数、作業時間、年齢（各男女別）に関する平均値及び標準偏差）と一般参考値(具体的には、年代ごとの平均正答率、作業時間のパーセンタイル順位）のみである。これらのデータからは、当時の被検者全体の傾向をおおよそ把握することは可能だが、項目反応理論を実施する上で必要な、各個人の正答状況のデータを得ることができない。具体的には「簡易版」における各被験者の20試行毎の正答状況や作業時間が現時点では公開されていない。

項目反応理論は、被検者ごとのスコアを取り入れて解析を行う必要があることから、現状のままでは分析にかかれないことが明らかとなった。

このため、代替的な方法として、個人の得点をシミュレーションによって生成し、模擬データを用いて項目反応理論を用いた分析までのプロセスを提示することとした。したがって、本章において得られた結果は、あくまで模擬データに基づいており、将来的には実測データそのものを用いて、同様の検証を行うことが必須である。しかしながら、第１章で触れたように、ワークサンプルの基準値を作成する上で課題となっている「一般労働者を対象としたデータ収集は極めてコストが高く、実現可能性が乏しい現状」に対して、シミュレーションによる模擬的データの作成は、大規模データを収集した結果と同等の情報量が得られる可能性を持つ。

そこで、本章では、単に項目反応理論の実施可能性のみならず、シミュレーションデータを利用することの有効性についても同時に検討することとしたい。

## ３　シミュレーションの方法

『社内郵便物仕分（簡易版)』の模擬データを乱数シミュレーションによって生成する方法を解説する。この模擬データは、障害者職業総合センター（2019）が一般参考値を算出するために収集した162名（年代別には「20代26名、30代34名、40代56名、50代46名」であったことを、当時の研究担当者において確認）の生データに可能な限り類似させなければならない。そのため、模擬データを構成するのは架空の162名の実施結果であり、各々について20試行の正誤の情報（○あるいは×）を含むものとなる。完成した模擬データは表３－２のような構造となる。例えば、参加者１は年代が20代であり、第１試行から第18試行は正答（○）であったが、第19試行と第20試行が誤答（×）となり、結果的に正答数は18となる。このようなデータを162名分生成する。

**表３－２　模擬データの構造**

| 参加者 | 年代 | 20試行の正誤 | 正答数 |
|---|---|---|---|
| 参加者 1 | 20代 | ○○○○○○○○○○○○○○○○○○○○×× | 18 |
| 参加者 2 | 30代 | ○○○○○○○○○○○○○×××○×○××× | 14 |
| … | … | …… | … |
| 参加者 161 | 40代 | ○○○○○○○○○○○×××○○×○×○○ | 16 |
| 参加者 162 | 50代 | ○○○○○○○○○○○○○○○○○○○○ | 20 |

一般参考値の生データに特徴が一致する模擬データをまったくの偶然に頼って生成するのは効率的でない。そこで、一般参考値として与えられている基礎統計値に基づいて母集団の分布を推定し、仮想的に構築した母集団からのランダムサンプリングを行うことによって模擬データを生成する方法を採用する。もし、推定した母集団が真の母集団を正しく反映しており、実際に行ったサンプリングの手順を正しく模倣することができれば、得られた乱数が実際の生データに類似する可能性は高くなるはずである。このような発想に基づいて模擬データを得る一連の手続をここではシミュレーションと呼ぶ。

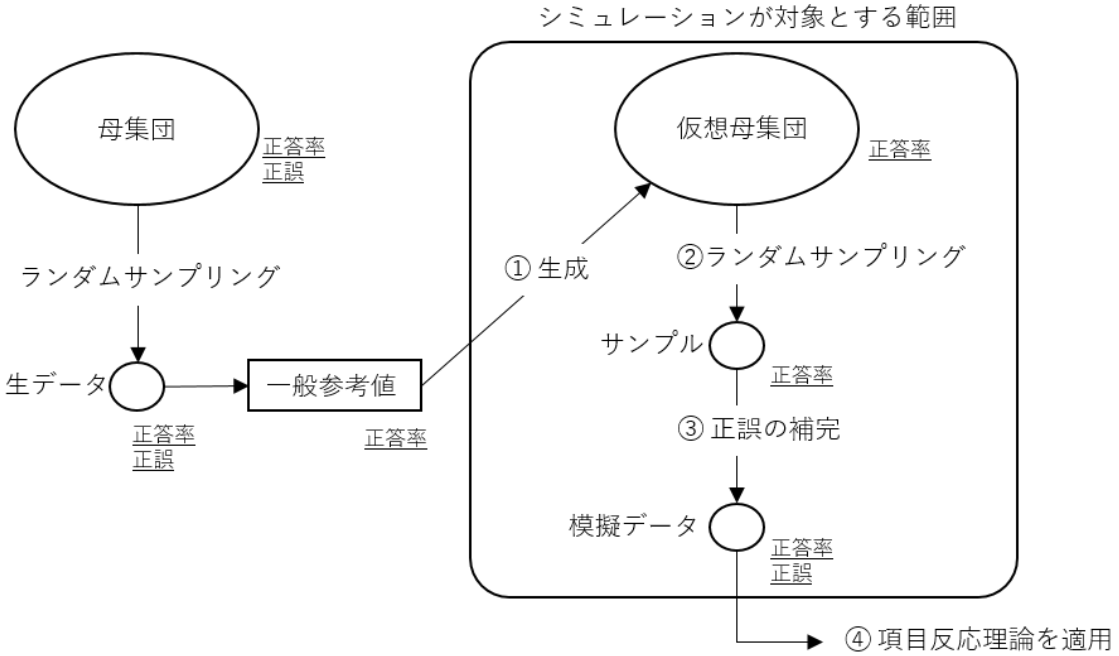



**図３－１　シミュレーションの概要**

図３－１には母集団から得られたサンプルとしての生データと、そこから算出された一般参考値に基づいて実施されたシミュレーションの結果である模擬データの関係が図示されている。図の左半分は既に完了した研究の範囲（一般参考値の作成）を示し、大枠で囲まれた右半分は今回の研究におけるシミュレーションが対象とする範囲である。この図が示しているとおり、生データと模擬データを橋渡ししているのは一般参考値だけである。

生データは 162 名の参加者の『社内郵便物仕分（簡易版)』の実施結果であり、一般参考値の作

成のために実施されたデータ収集において母集団からのランダムサンプリングによって得られたものである。ここで得られた生データの要約統計量が一般参考値として算出されている。ただし、一般参考値が示す指標は正答率に関する統計量のみであり、試行ごとの正誤の情報は与えられていない。そのため、母集団の推定に利用できる情報は正答率のみである。

この一般参考値に基づいてシミュレーションを実施する。シミュレーションは全体で四つの工程から成り、概略は以下のとおりである。

第１の工程（図３－１の矢印①）では、正答率に関する母集団の分布を仮定し、これに従う乱数データセットを生成する。この乱数データセットを仮想母集団と呼ぶ。ただし、母集団分布として想定しうる形状は複数存在するため、それぞれの想定の下で母集団の推定を行った。

第２の工程（矢印②）では、構築した仮想母集団からランダムサンプリングを行う。ここでは参加者の年代別にサンプリングを行う方法と、参加者の年代を区別せず全体からサンプリングを行う方法の二つの方法を実施した。

第３の工程（矢印③）では正誤の情報の補完を行う。これによって模擬データが完成し、表３－２のようになる。

第４の工程（矢印④）では試作した模擬データの一つに対して項目反応理論を適用する。

以下では、各々の工程の詳細を個別に解説する。

**（１）工程１：仮想母集団の生成**

シミュレーションの第１の工程は、母集団分布と同じとみなせる仮想的な母集団を生成することである。これはシミュレーションの概要として図３－１に示した矢印①に相当する。しかし、一般参考値に示された統計値から母集団分布を完全に再現することは不可能である。そこで、母集団分布として妥当と思われる形状のそれぞれについてシミュレーションを実施した。用いた３種類のシナリオの概念図を図３－２に示し、以下で詳述する。

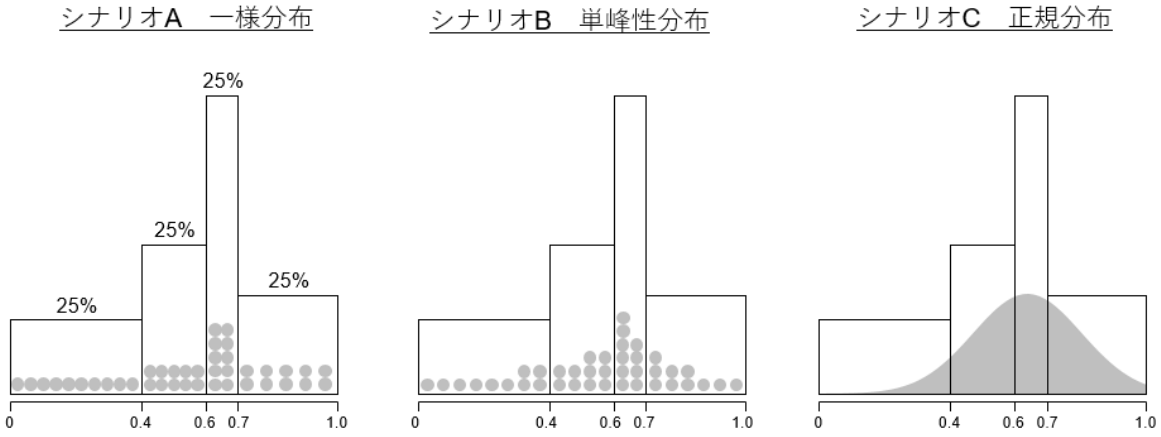



**図３－２　３種類の母集団分布の概念図**

**ア　パーセンタイルの意味**

母集団分布の形状を推定する根拠となるのは一般参考値に示されているパーセンタイルである。そこで、３種類の分布の違いを説明する前に、パーセンタイルの意味を明らかにしておく。

いま、40 個のデータがあり、それは 0.0 から 1.0 の範囲を取る値であるとする。例えば、0.1 や 0.2 や 0.999 などである。この 40 個のデータを座標軸上に並べて置いていくことを考える。すると、いくつかのデータは値が重複するため、すでに置かれているデータの上に積み上げることにする。この作業を続けていき、40 個のデータを並べ終えると、最終的に図３－２（シナリオ A）のような分布が得られたとする。

この図を見ると、データはちょうど 10 個ずつでまとまりを作っているようである。そこで、値が小さい順にデータを数えていき、データを 10 個ずつに区切っていく。すると、最初の 10 個は 0.4 より小さい値であることがわかった。次の 10 個を数えると、0.6 より小さい値であることがわかった。さらに次の 10 個を数えると、0.7 より小さい値であった。残りの 10 個は必然的に 0.7 より大きなデータだということになる。結果的に、座標軸上に三つの境界を引くことができた。すなわち、0.4 と 0.6 と 0.7 である。

このようにデータを数えたときに得られる数値がパーセンタイルである。例えば 0.4 は 25%点と呼ばれ、この値より小さいデータは全体の 25%（10 個）に相当することを意味する。同様にして、0.6 は 50%点であり、この値より小さいデータは全体の 50%（20 個）に相当する。0.7 は 75%点であり、この値より小さいデータは全体の 75%（30 個）に相当する。すると、残りの 25%である 10 個のデータがカウントされずに残るが、これは 0.7 より大きい数値であることがわかる。つまり、パーセンタイルとは、データを小さい順に数えたときのデータ全体に占める割合のことである。

それでは、パーセンタイルの値から類推した分布であるシナリオ A は母集団の分布とみなせるだろうか。実は、パーセンタイルの値（0.4、0.6、0.7）がまったく同じであっても、形状が異なる分布はいくつでも想定することができる。パーセンタイルは一定区間の中にデータがいくつあるかを示す指標でしかなく、その区間内にデータがどのように分布しているかは未知のままである。例えば、シナリオ A では、25%点は 0.4 であった。これは、0.4 より小さいデータは全体の 25%（つまり 10 個）であることを意味しているが、10 個のデータが 0 から 0.4 の範囲に等間隔に存在しているという意味ではない。あくまで、最初の 10 個のデータが 0 から 0.4 の範囲にあった、という事実を述べているだけである。そのため、値が 0 のデータが 10 個あったのかもしれないし、値が 0.1 のデータが５個と値が 0.2 のデータが５個だったのかもしれない。最初の 10 個のデータが実際にはどのような値であり、どのような偏りを持っていたかは定かではなく、その可能性は無限に存在する。図３－２に示されているデータの並びは無限に存在する可能性のうちの一つにすぎない。

このように、一般参考値に示されたパーセンタイルの情報をそのまま解釈しただけでは母集団分布の形状を定めることはできない。そこで、新たな仮定を加えることによってシミュレーションで使用できる仮想母集団を作ることにした。ここでは、それぞれの仮定によるシミュレーションをシナリオ A、シナリオ B、シナリオ C のように区別して説明する。

なお、以下で導入した仮定は理論的に導き出されたものではなく、恣意的に設定されたものである。よって、この仮定は検証されなければならない。すなわち、加えた仮定が母集団分布の特徴として妥当なものであったかという問題について、生成した模擬データを評価することによって検証する必要がある。

**イ　シナリオ A：一様な分布**

パーセンタイルから推定される素朴な分布に仮定を加える。ただし、加える仮定はできるだけ現実に即しており、無理がないものが望ましい。

パーセンタイルのみが与えられている状態で母集団分布を推定するなら、まずは一様分布を仮定すべきであろう。ある区間内での分布が不明であり、どのような偏りを持っているのか未知であるなら、偏りはないと想定するしかない。つまり、その区間内での分布は一様であると考える。

一様分布の想定では、パーセンタイル順位が示す値で区切られた範囲内では、どのような正答率も発生頻度は等しいと仮定する。例えば、図３－２（シナリオ A）では 25%点が 0.4 であるため、値が０から 0.4 のデータが発生する頻度は一様であって偏りはないとする。どこかの正答率の割合が他より高いということも、その逆に、どれかの正答率の割合が他より低いということもないと仮定する。

これは一般参考値から読み取れる母集団分布の形状として最も単純な想定であり、逆に言えば、何も想定しないことに近い。とはいえ、厳密に言うと何も想定がないというわけではない。一般参考値は正答率の最小値と最大値を示していないため、この分布の最小値も最大値もわからない。ただし、これが正答率という確率で表された数値である以上、理論上の最小値は０であり最大値は１である。よって、シナリオ A に限らず、このシミュレーション研究においては値の範囲を０から１であるものとして進める。

**ウ　シナリオ B：単峰性の分布**

一様分布の想定は恣意的な仮定が少ないという意味では望ましいが、データを機械的に割り振るものであり、母集団の近似として優れているとはいえない。近似の精度を向上させるためには仮定を加える必要がある。

ここで、厳密とは言えないが、データの分布として妥当な形状を考えたい。それは、分布が山のような形状をしているとしたら、山頂は一つしかないという単峰性の仮定である。そして、山の高さは山頂から離れるに従い直線的に低くなっていくと仮定する。これを表現したものが図３－２のシナリオ B である。このシナリオでは、山頂は一つだけであり、その両端は徐々に落ち込んでいき山の裾になっている。単峰性の仮定に反して山頂がいくつもあれば、実在の連峰がそうであるように複雑な稜線が生じていたであろう。山頂が一つであると仮定すれば、山頂に近いほど頻度が高く、山頂から離れるほど頻度が低くなるような分布を作ることができる。

単峰性の仮定を利用するためには山頂の位置を定める必要がある。山頂の位置を表すなら最頻値を使用するべきだが、後述するシナリオ C では平均値を分布の中心位置と定めている関係上、ここでも平均値を用いた場合の結果を示す。

**エ　シナリオ C：正規分布**

数学的に記述可能な確率分布によって母集団を近似することができれば、より精密な分布の検討が可能となる。そこで、データの分布として典型的であると考えられている正規分布を仮定する。正規分布はデータが最も集中しやすいある一点を中心として、そこから離れるほどデータが急激に少なくなっていく様子を示している（図３－２のシナリオ C)。通常の正規分布は裾が左右に無限

に伸びていく。しかし、課題成績には必然的に最小値と最大値が存在するため、この分布は実際には途中で切断されたようになる。そのため、厳密に言えばこれは打ち切り正規分布である。

正規分布を利用するため中心の位置（平均値）と裾野の広さ（標準偏差）を特定する必要がある。ここでは、年代別のパーセンタイルの値に累積正規分布を当てはめ、平均値と標準偏差を得た。この値を正規分布のパラメーターとして用いて母集団分布を仮想的に構築した。

**（２）工程２：仮想母集団からのランダムサンプリング**

シミュレーションの第２の工程は仮想母集団からランダムサンプリングを行うことである。これは、シミュレーションの概要として図３－１に示した矢印②に相当する。ここでは、年代を区別するサンプリングと、年代を区別しない全体からのサンプリングという２通りのサンプリング方法を試みる。

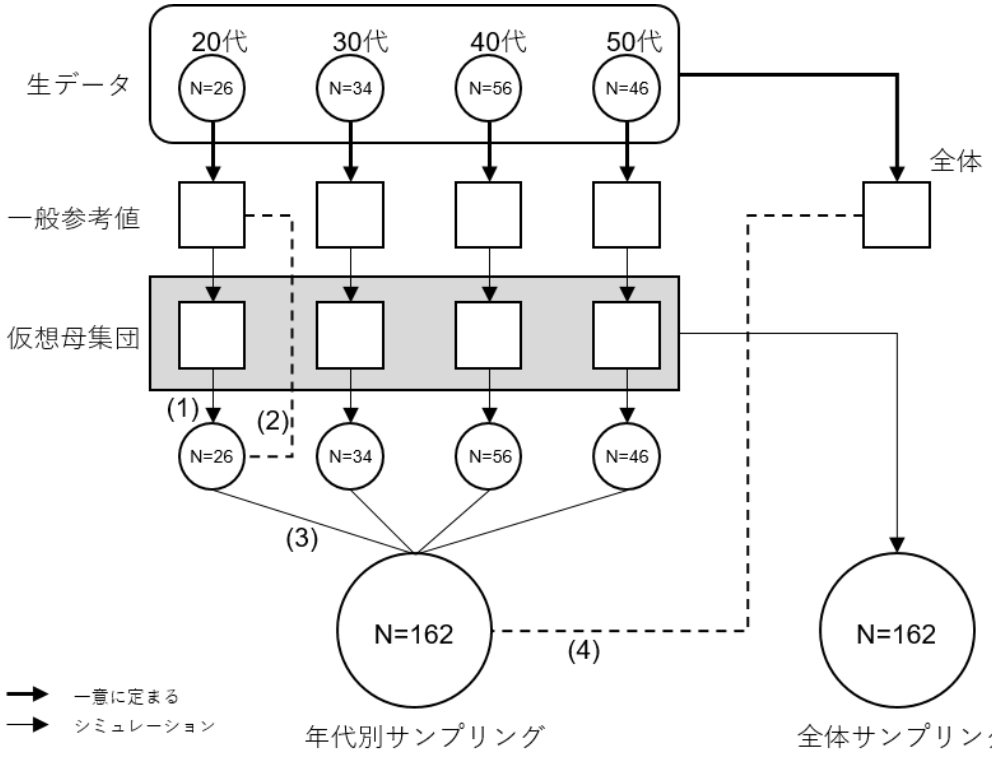



**図３－３　年代別サンプリングと全体サンプリング**

サンプリングの手順は図３－３の上から下へという順番で実施される。図中の太い矢印は既存の研究ですでに確定している不動の情報（生データと一般参考値）を表しており、これに基づいたシミュレーションの手順が細い矢印に相当する。生データの取得（全体で 162 名）とそれを集計した一般参考値は障害者職業総合センター（2019）による研究の範疇である。一般参考値は年代ごとに集計されており、これに基づいて各年代の母集団をランダム生成した（ランダム生成の過程で三つのシナリオがあることは工程１で説明した）。そして、それを合体させることで仮想母集団を生成した。各年代の仮想母集団は１万件の疑似乱数から構成されているため有限母集団であるが、重複を認めるランダムサンプリングを行うことで事実上の無限母集団とみなせる。

また、この図に示されているとおり、一般参考値は年代別だけでなく、各年代を合算した全体としての数値も示されているが、これを仮想母集団の生成に使用しているわけではない。ただし、後述するように、最終的に得られたサンプルの適否を確認するための比較照合として利用しており、シミュレーションの中でも重要な情報として扱っていることを付記しておきたい。

これ以後の手順は年代を区別するサンプリング方法と年代を区別しないサンプリング方法によって異なる。以下でそれぞれの方法を説明する。

**ア　年代を区別するサンプリング方法**

一般参考値には年代別のパーセンタイルと、全年代を合算した全体のパーセンタイルが示されている。年代を区別するサンプリング方法では、模擬データをこれらすべてに合致するものとして作成する。そのため、サンプリングは図３－３の（１）から（４）に示した順序で行う。（１）仮想母集団から年代別にランダムサンプリングを行い、一般参考値とサンプルサイズが一致するサンプルを年代ごとに都合四つ得る。（２）得られた四つのサンプルをそれぞれの年代の一般参考値と比較し、相関が十分に高いことを確認する。（３）四つの年代別サンプルを結合する。最後に、（４）全体の一般参考値と比較し、相関が十分に高いことを確認する、という手順である。このような手順によって、一般参考値に極めて類似した模擬データを生成できる可能性が高くなる。

なお、年代を区別するサンプリングを行うということは、年代ごとのサンプルサイズを事前に決定していたという想定に基づいたシミュレーションを実施したことに等しい。つまり、一般参考値の参加者 162 名は、年代ごとにサンプルする人数を事前に定めており、その定められた人数を各年代の母集団からランダムサンプリングしたという想定となる。具体的に言えば、生データに 20 代が 26 名含まれているのは、20 代から 26 名をランダムサンプリングすることを事前に定めていたからである、という想定である。

この手法を用いれば、既存の研究で取得した 162 名の年代構成と一致する模擬データを得ることができる。そのため、実測した 162 名のデータを対象とした項目反応理論の適用を意図するのであれば、年代を区別するサンプリングによる模擬データを扱うべきかもしれない。ただし、実際のデータ収集（障害者職業総合センター，2019）では年代別のサンプルサイズが事前に決定されていたわけではないため、このシミュレーションは実際のサンプリングの手順と異なるという欠点がある。

**イ　年代を区別しないサンプリング方法**

年代を区別しないサンプリング方法では、各年代をまとめた全体の仮想母集団から 162 名全員分の値を一度に取得する。そして、パーセンタイルが一般参考値と一致しているかどうかを、年代別及び全体について同時に確認する。

この場合、一般参考値の参加者 162 名は年代を区別せずにサンプリングされたという想定に基づく。つまり、生データの 162 名の年代配分はランダムサンプリングの結果であり、人数を事前に決めていたわけではないと考える。例えば、一般参考値には 20 代が 26 名含まれているが、20 代から 26 名を集めるという意図があったわけではなく、母集団からランダムにサンプルを取った結果、偶然にも 20 代が 26 名集まったということを意味する。同時に、年代ごとの偏りもランダムサンプリングの結果であるとみなす。具体的に言えば、一般参考値の生データにおける年代の構成は 20 代

が最も少ない 26 名であり、40 代が最も多い 56 名となっている。これは、年代ごとの人数配分に最大で２倍の偏りがあることを意味するが、このような配分もランダムサンプリングによって偶然発生した傾向とみなす立場である。

さらに言えば、この想定は一般参考値の年代配分をそのまま母集団の年代配分の推定とみなす立場を取ることに等しい。この立場を取れば、真の母集団の内訳として 40 代の人数（56 名）は 20 代の人数（26 名）の約２倍であると推定する。実在の 162 名のデータを取得した際のサンプリングの手順はこちらに近いはずである。

**（３）工程３：正誤のシミュレーション**

第３の工程はそれぞれの参加者がどの試行に正答し、どの試行に誤答したのかという正誤の情報を補完することである（表３－２の○×に相当する）。正誤を補完するために必要となるのは各試行の難易度に関する情報であるが、一般参考値は個別の試行の正答率を示していない。つまり、全 20 問のうち易しい試行（つまり正答率の高い試行）はどれであり、難しい試行（正答率が低い試行）はどれであるかについて、推定のために依拠できる根拠が存在しない。これを補うために、３名の評定者（『社内郵便物仕分（簡易版）』を採点した経験あり）による難易度の得点化を行った。

３名の評定者が回答したのは「試行の難しさ」という単一の基準についてであり、２名の評定者は１点から３点で割り振った３件尺度によって、残りの１名の評定者は１点から５点で割り振った５件尺度によって評定した。評定の基準が一つに限定されているのは、この評定が正答率の暫定的な仮定を目的としているからであり、正答率の高低を導く原因の特定を目的としていないからである。また、評定者により得点化の尺度が異なっているが、できるだけ評定者の知見を反映することができるようにあえて尺度化の方法を指定しなかったためである。

図３－４左に評定の結果を示す。ここに示された数値は評定者による得点そのものではなく、正答する確率として読み替えてプロットしたものである。そのため、値が高い試行ほど正答しやすく、値が低い試行ほど正答しにくいことを示している。図中の灰色の各シンボルは３名の評定者を示し、黒いシンボルは３名の平均値である（誤差棒は±１標準偏差である）。この図から、正答率が高い試行が多いことや、極端に正答率が低い試行がほとんどないことが読み取れる。

３名の評定はおおよそ一致しているが、多少なりともばらつきがあることから、３名の平均値を推定の根拠として用いた。このようにして与えられた試行ごとの正答率を合計が１となるように基準化することによって、参加者がどの試行に正答したのかを表す確率分布が得られた。これを図３－４右に示す。

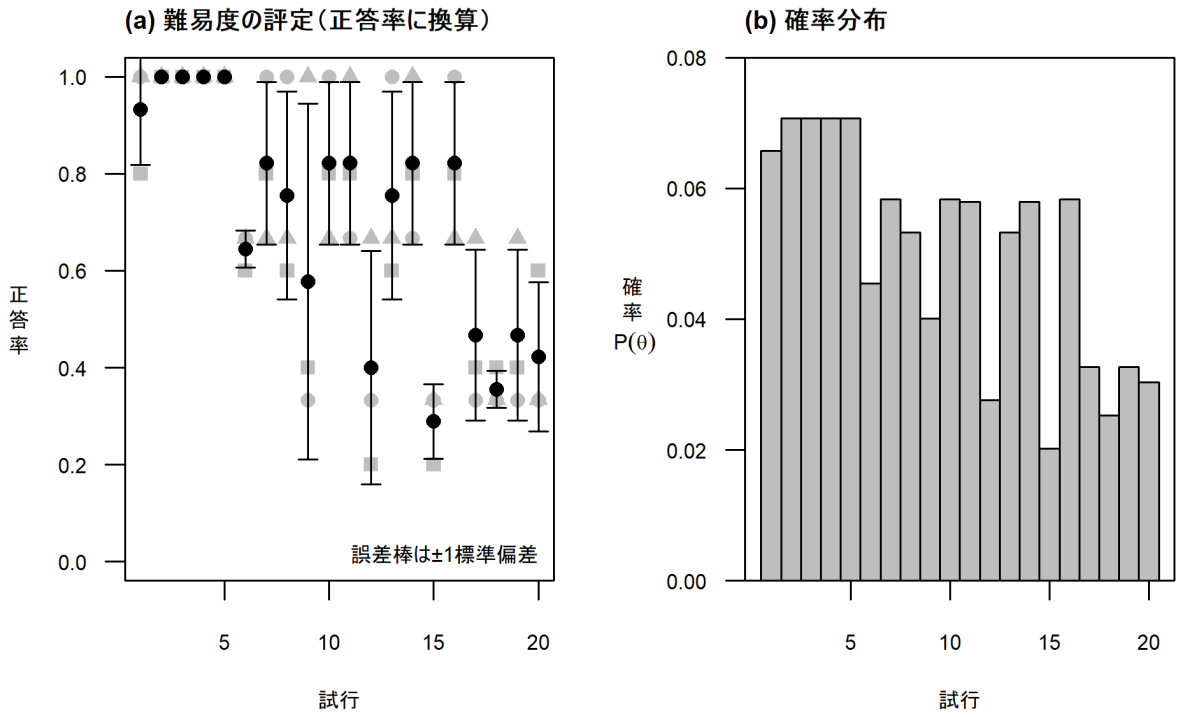



**図３－４　(a)難易度の評定及び(b)確率分布**

正誤の補完は 20 問から決められた数の試行を選び出す非復元抽出として実行した。つまり、20 問から正答数の数だけ試行を選び出したときに、選び出されたその試行をその参加者の正答した試行とみなし、選ばれずに残った試行を誤答した試行とみなすという方法である。このとき、どの試行が選ばれるかは一律ではなく、図３－４右に示した確率分布に従って抽出される。

この手順は壺から玉を取り出す作業によく似ている。いま、一人の参加者の目の前に中の見えない壺が置いてあり、壺の中には１から 20 までの番号が書かれた 20 個の玉が入っている。その参加者は壺から玉を一つずつ取り出し、玉の番号を記録していく。このとき、取り出した玉は壺の外に置かれるため、同じ玉を二回以上取り出すことはない。特徴的なのは、壺から取り出す玉の数が参加者ごとに異なることである。ある参加者は 10 個だけ取り出すことが許されているが、別の参加者は 20 個すべてを取り出すことが許されているなど様々である。これは、参加者ごとに正答数が異なるからである。もう一つ特徴的なのは、どの玉が選ばれるかは一定ではなく、取り出されやすい玉もあれば、取り出されにくい玉もあるということである。玉が取り出される確率の違いを表しているのが図３－４右である。このようにして取り出された玉こそが正答した試行であり、壺の中に残った玉が誤答した試行ということになる。

**（４）工程４：項目反応理論の適用**

三つのシナリオと二つのサンプリング方法の組合せによって都合６種類のシミュレーションを実施し、そこから得られたサンプルのうち最も精度のよいサンプルを模擬データとして利用することとし、これに対して項目反応理論を適用する。ここでは項目反応理論の利用方法の紹介を目的として、２パラメーター・ロジスティック・モデル（２PLM）を適用した。これによって、それぞれ

の試行の困難度と識別力を評価することができる。

## ４　シミュレーションの結果

### （１）シナリオとサンプリング方法の比較

図３－５はそれぞれのシナリオ（シナリオA、B、C）及びサンプリング方法（年代別と全体）のもとで実施したシミュレーションの結果である。都合六つのシミュレーションの結果が示されている。ここには、シミュレーションによって生成した 10 万個のサンプルが灰色の点としてプロットされている。横軸はサンプルの平均値、縦軸はサンプルの標準偏差を示す。白い実線は二次元カーネル密度推定によって算出した境界線であり、半分のサンプル（すなわち５万件）が存在する範囲を示している。図中の十字マーク（＋）は 10 万件のサンプルの平均値である。

また、黒い実線で描かれた四角形は一般参考値の値を示しており、全体の平均値（平均は 75.6%、標準偏差は 16.6%）を中心として、その±0.5 の範囲を描いたものである。つまり、この四角形の内部にプロットされたサンプルは一般参考値に極めて類似しているとみなせる。この四角形で囲まれた範囲にあるサンプルの中でも特に精度のよいサンプルを黒い点で示した。これは、各年代及び全体のパーセンタイルが一般参考値と 0.99 以上の高い相関を持つサンプルであり、全体の平均値と標準偏差が一般参考値の±0.1 の範囲にあるサンプルである。都合六つのシミュレーションの結果、８件の精度のよいサンプルが得られた。

図３－５を概観すると、シナリオＡとシナリオＢが類似した結果を示しており、シナリオＣはそれと異なることがわかる。また、全体サンプリングと年代別サンプリングには目立った違いがないこともわかる。ここでは主にシナリオの違いに着目してシミュレーションの結果を検証する。

シナリオ A とシナリオ B はよく似た結果となった。平均や標準偏差の範囲もほぼ等しく、分布の形状も類似していることがわかる。そして、両者とも分布の中心位置は一般参考値よりやや離れていた。いずれのシナリオでも平均値は一般参考値より２%ほど低く、標準偏差は一般参考値より３ほど高くなった。言い換えると、シナリオ A とシナリオ B のシミュレーションでは、一般参考値より正答数が少なく、かつ、正答数のばらつきが大きくなるようなデータが多く生じたことを意味している。ただし、正答率のパーセンタイルは一般参考値とよく一致していた。パーセンタイルの相関を示したものが図３－５の各図の右上に挿入されたヒストグラムである。これを見ると、ほとんどすべてのサンプルが相関係数 0.99 を超えており、一般参考値と相関が高いサンプルが多く得られたことがわかる。

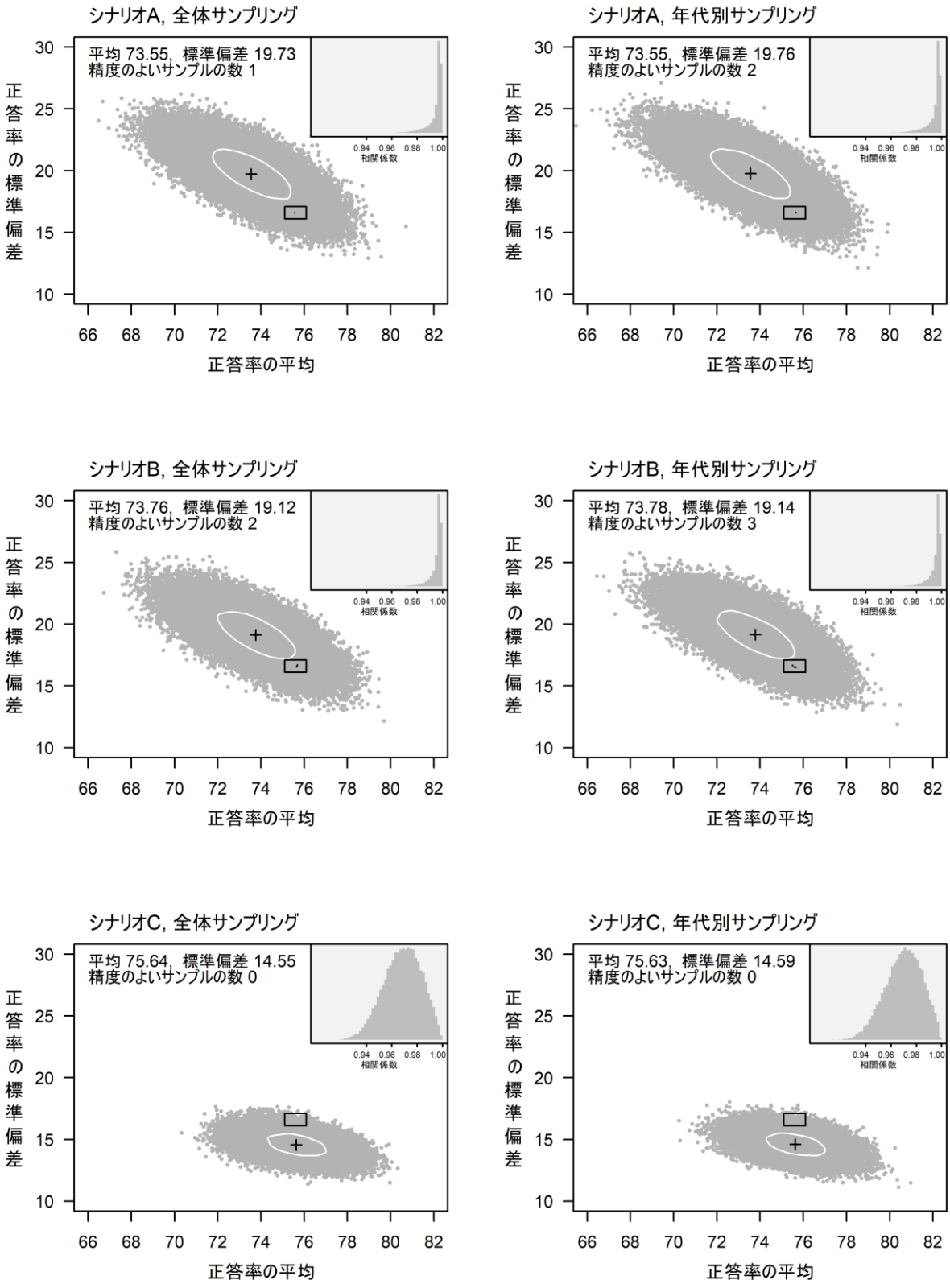



図３－５　正答率のシミュレーション

精度のよいサンプルは、全体サンプリングと年代別サンプリングを合わせると、シナリオＡで三つ、シナリオＢで五つ得られた。一つのシミュレーションが 10 万件のサンプルを生成していることを勘案するとわずかな違いでしかないが、模擬データの生成はシナリオＢの方が若干優れるようである。ちなみに、シナリオＡの年代別サンプリングでは精度のよいサンプルが図中に一つしか描かれていないように見えるが、これは二つのサンプルが近接しているためである。

このように二つのシナリオは非常に類似した結果となったが、サンプルの 50%が集中的に分布する範囲（図の白い線で囲った範囲）を見るとシナリオＢの方がわずかに一般参考値に近い。よって、シナリオＢはシナリオＡより母集団の推定として優れたシナリオであるといえる。

一方で、シナリオ Ｃ の結果はこれらと異なっている。サンプルが分布する範囲は狭くなっており、特に標準偏差のばらつき具合の減少が顕著である。分布の中心位置は一般参考値に近づいており、平均値は一般参考値とほぼ同じ値が得られた。ただし、標準偏差は一般参考値を下回り、２ほど小さい値が得られた。これはいわば、一般参考値の平均値に近づけようとする試みがサンプルのばらつきを減少させてしまい、結果的に標準偏差が小さくなりすぎてしまったことを意味している。さらに、シナリオ Ａ とシナリオ Ｂ との違いは一般参考値との相関を見ると明瞭である。シナリオ Ｃ のパーセンタイルは一般参考値との相関があまり優れていない。シナリオ Ａ とシナリオ Ｂ ではほとんどのサンプルが 0.99 を超える相関を示したのに対して、シナリオ Ｃ の場合は 0.99 を超えるものは少なく、ほとんどが 0.96 から 0.98 の範囲にあった。こういった事実からも、シナリオＣは模擬データの生成のシミュレーションとして採用できる精度を満たさないといえる。

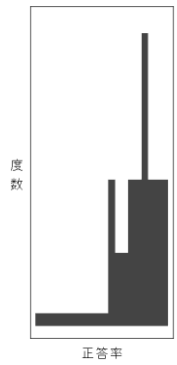
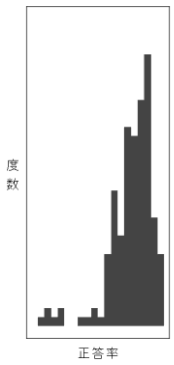
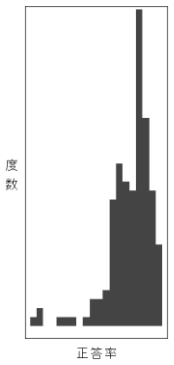
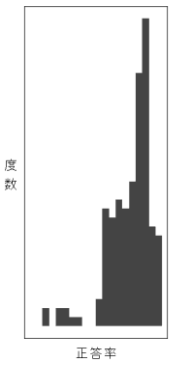
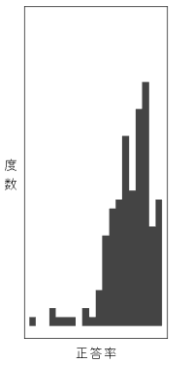
**（２）サンプルの検証**

六つのシミュレーションによって精度のよいサンプルが８件得られた。表３－３には各シミュレーションの中でもっとも精度のよかったサンプルを一つだけ示す。シナリオ Ｃ では精度のよいサンプルが得られなかったので、シナリオ Ａ とシナリオ Ｂ のサンプルだけが示されている。シミュレーションの前提である一般参考値の情報を最左列に示した。

平均値と標準偏差は一般参考値の±0.1 を基準としてサンプルを選定しているため、当然ながら一般参考値とほぼ同じ値が得られている。また、分布の形状の記述として利用されることの多い歪度[1]と尖度[2]を表示した（一般参考値には歪度と尖度は未掲載のため表中では N/A となっている）。正規分布であれば歪度は必ず０となることが知られているが、サンプルの歪度はいずれもマイナスの値であり、サンプルを正規分布と見なすことはできないことがわかる。シナリオＣの精度が他と比較して低いのはこういった理由によるかもしれない。

[1] 分布の偏りや歪みを表す指標。
[2] 分布のとがり具合を示す指標。

**表３－３　シミュレーションによって得られた精度のよいサンプルの基礎統計値**

| | | 一般参考値 | サンプル１ | サンプル２ | サンプル３ | サンプル４ |
|---|---|---|---|---|---|---|
| シナリオ | | ― | A | A | B | B |
| サンプリング | | ― | 全体 | 年代別 | 全体 | 年代別 |
| | |  |  |  |  |  |
| 平均値 | | 75.6 | 75.57 | 75.65 | 75.65 | 75.51 |
| 標準偏差 | | 16.6 | 16.60 | 16.58 | 16.56 | 16.66 |
| 歪度 | | N/A | -1.70 | -1.70 | -1.46 | -1.41 |
| 尖度 | | N/A | 3.94 | 4.28 | 2.77 | 3.06 |
| パーセンタイル | 100 | N/A | 99.60 | 99.95 | 100.00 | 99.90 |
| | 90 | 95 | 91.27 | 91.86 | 92.65 | 93.92 |
| | 80 | 90 | 88.23 | 88.23 | 87.87 | 88.31 |
| | 70 | 85 | 85.61 | 84.18 | 85.54 | 85.53 |
| | 60 | 85 | 82.09 | 82.51 | 83.44 | 83.01 |
| | 50 | 80 | 79.07 | 80.20 | 80.35 | 78.18 |
| | 40 | 75 | 74.87 | 75.39 | 75.36 | 74.13 |
| | 30 | 70 | 71.36 | 69.68 | 69.49 | 69.73 |
| | 20 | 60 | 64.74 | 65.24 | 63.03 | 64.85 |
| | 10 | 55 | 59.61 | 59.95 | 58.78 | 57.58 |
| | 0 | N/A | 7.22 | 4.55 | 10.46 | 3.33 |
| 一般参考値との相関 | | ― | 0.9977 | 0.9971 | 0.9989 | 0.9980 |

パーセンタイルについて０％点から 100%点までを 10%区切りで算出した。さらに、サンプルのパーセンタイルと一般参考値のパーセンタイルの相関を最下段に示した（一般参考値には０％と100%は未掲載のため表中では N/A となっている)。ここに示した四つのサンプルの中ではサンプル３において相関の値が 0.9989 となっており、生データの再現としてもっとも精度のよいサンプルである。

## ５　項目反応理論を用いた模擬データの分析

表３－３に示した四つのサンプルのうち、一般参考値のパーセンタイルと相関がもっとも高かったサンプル３を模擬データとして用いることに決定した。さらに工程３（第３章第１節３（３）工程３：正誤のシミュレーション）の手順に基づいて正誤の情報を補完することで模擬データを完成させ、これに対して項目反応理論による分析を行った。ここでは項目反応理論の例示を目的として、困難度と識別力を想定する２PLM を用いた。

なお、第２章において項目反応理論を紹介した際には「問題」という用語を用いて個別の試験項目を記述したが、『社内郵便物仕分（簡易版）』では「試行」という用語を用いるため、以下の分析では「試行」に統一した。

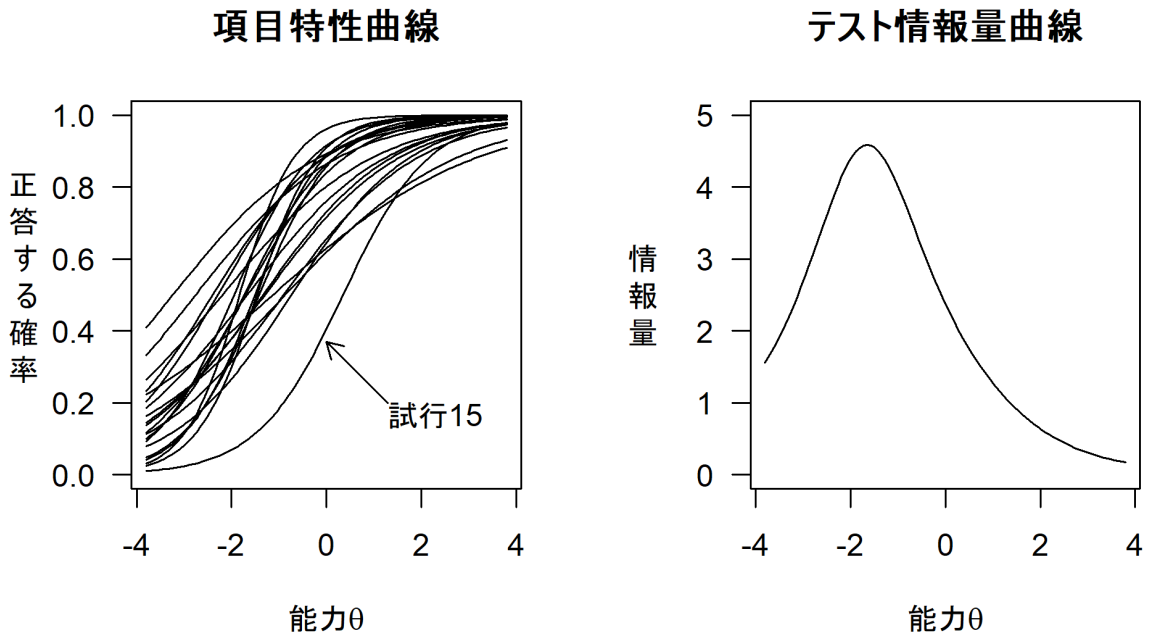



図３－６　２PLM によるサンプル３の分析結果

図３－６の左の図には『社内郵便物仕分（簡易版）』の 20 試行それぞれについて項目特性曲線が描かれている。ここではグラフの見やすさを重視したため、どの曲線がどの試行であるかを示さなかった（ただし、一例として試行 15 だけ矢印で示した）。項目特性曲線の形は試行の特性を表しており、その試行に正答する確率が受験者によってどのように変わるのかを読み取ることができる。今回分析に用いたモデルは２PLM であり、グラフの左右の位置の違いは試行の難しさの違いであり、グラフの傾き具合の違いは識別力の違いである。

試行の難しさの指標である困難度については、グラフの左側に位置する曲線ほど困難度は低く、易しい試行であることを意味している。逆に、グラフの右に位置する曲線ほど困難度は高くなり、難しい試行であることを意味している。このグラフを見ると、全体的に左に寄った曲線が多く、易しい試行が多い様子が読み取れる。正答する確率が 0.5（すなわち 50%）となる能力 $\theta$ を探すと、

ほぼすべての試行がマイナスの値であり、能力θがプラスの曲線は矢印で示した試行 15 の１問だけであった。これは、平均的な能力の受験者であれば（つまり、能力θが０の受験者)、試行 15 以外であればどの試行であっても 50%以上の確率で正答できることを意味している。逆に言えば、平均以上の能力を持つ受験者は 20 問のほぼすべての試行に正解してしまう。すると、受験者同士の能力の違いを区別することが難しくなる。つまり、この 20 問が判別できるのは能力の高い受験者達ではなく、能力の低い受験者達である。

これと同様の解釈はテスト情報量曲線[3]（図３－６の右の図）によっても裏付けられる。テスト情報量曲線は 20 問全体としての推定の精度を表している。この曲線の値がもっとも高くなっているのは能力θが－２に近いあたりである。つまり、『社内郵便物仕分（簡易版)』の 20 問は受験者の能力が－２であるかどうかを判別することに優れた課題であるということがわかる。

困難度についてさらに検討するため、20 試行を表３－１の分類に基づいてレベルごとに分けて表示したのが図３－７である。これを見ると、曲線の左右の位置のずれはレベルごとに違いがあることがわかる。レベル２に示した曲線はそれぞれ左右に大きくずれているのに対して、レベル４の曲線は比較的まとまっているように見える。つまり、レベル２の試行は難易度のばらつきが大きく、標準的な易しさの試行もあれば極端に易しすぎる試行もある。これに対してレベル４の試行は難易度が一定に保たれており、同程度の難しさの試行で構成されていることがわかる。

２PLM のもう一つの指標は識別力である。識別力は受験者の能力の違いをどれほどの精度で区別できるかの指標であり、グラフの傾きの違いとして現れる。具体例としてレベル２に相当する試行を見てみよう。図３－７に示したレベル２のグラフには試行４と試行 14 の曲線を黒い実線で描いて強調させてある。この二つの試行について、グラフの位置の違いは無視して、傾き具合に注目してほしい。試行４の傾きは急峻であるのに対して、試行 14 はなだらかになっていることがわかる。なぜこのような傾き具合の違いが能力の区別の精度となるのかを以下で説明する。

試行４の困難度は－1.83 であり、グラフの中心は横軸の－1.83 の位置にある。改めて解説すると、これは能力θが－1.83 の受験者であれば 50%の確率で試行４に正答することを意味している。ここで、50%というピンポイントの確率ではなく、「大体 50%」の確率で正答できる能力について考えたい。試しに、45%から 55%の範囲を「大体 50%」としよう。試行４の識別力は 1.76 であり、計算してみると、能力θが－1.90 から－1.76 までの 0.14 の範囲にあれば試行４に正答する確率が大体 50%であることがわかった。同様に試行 14 についても計算してみると、識別力は 0.65 であり、「大体 50%」の能力θの範囲は－3.42 から－3.06 までの 0.36 の幅に相当することがわかった。

ここで試行４と試行 14 を比較してみると、試行４（識別力は 1.76）では能力θの 0.14 の幅が「大体 50%」に相当したのに対して、試行 14（識別力は 0.65）では能力θの 0.36 の幅が「大体 50%」である。つまり、試行 14 では「大体 50%」に相当する能力θの幅が広がってしまっている。もし幅の広い基準で受験者を区別しようとすれば、精度は必然的に粗くなってしまう。別の言い方

[3] テスト情報量とは能力θの推定がどれほど精確であるかを表す指標である。テスト情報量の逆数を計算したものがθの推定に伴う誤差の大きさであり、テスト情報量が高いことは推定誤差が小さいことを意味する。テスト情報量は試行ごとに算出することができ、これをすべての試行について足し合わせたものがテスト全体のテスト情報量曲線である。

をすれば、試行４は能力θの 0.14 の違いを区別できるが、試行 14 では 0.36 違ってようやく区別できるということでもある。これが識別力の高低の意味である。

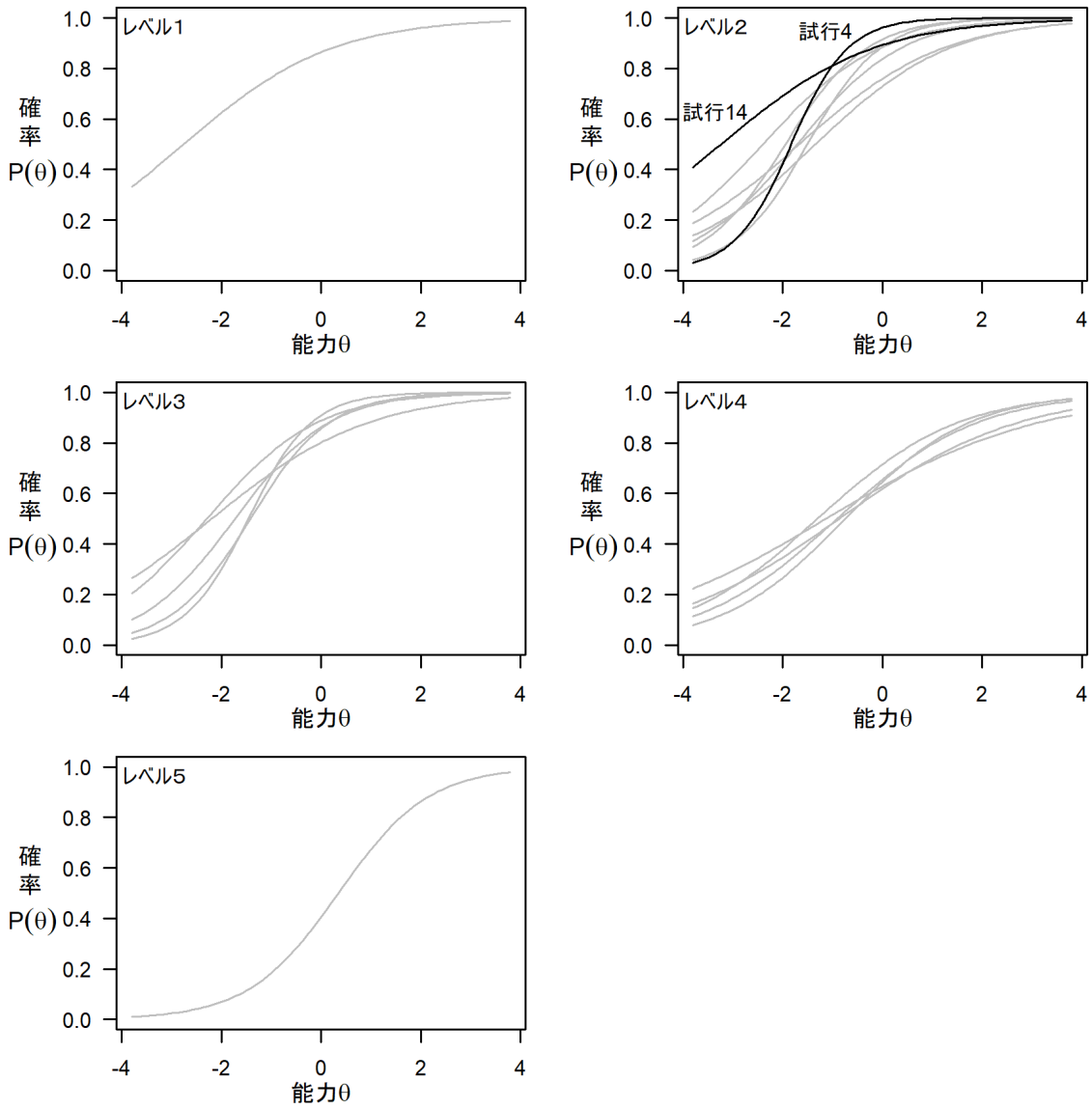



**図３－７　レベルごとの項目特性曲線（図３－６左を再掲）**

このように、試行ごとに識別力は異なっている。そこで、識別力の低い（精度の悪い）試行を取り除き、試行数を減らすことを考える。例えば 20 問から識別力の高い上位 10 問だけを抜粋して課題を実施したとすると、テスト情報量曲線は図３－８の灰色の曲線のようになる。黒い実線は 20 問すべてを用いた場合のテスト情報量曲線であるが、ピークの位置がほとんど変わっていない。このことから、識別力の高い試行だけを抜き出した場合でも、高い精度で測定できる能力θの値は変わらないようだ。ただし、曲線は全体的に下に移動しており、情報量は減少している。それでも曲線の全体的な形状は変わらず、20 問すべてを用いた場合と同じ性質の課題が維持されている。

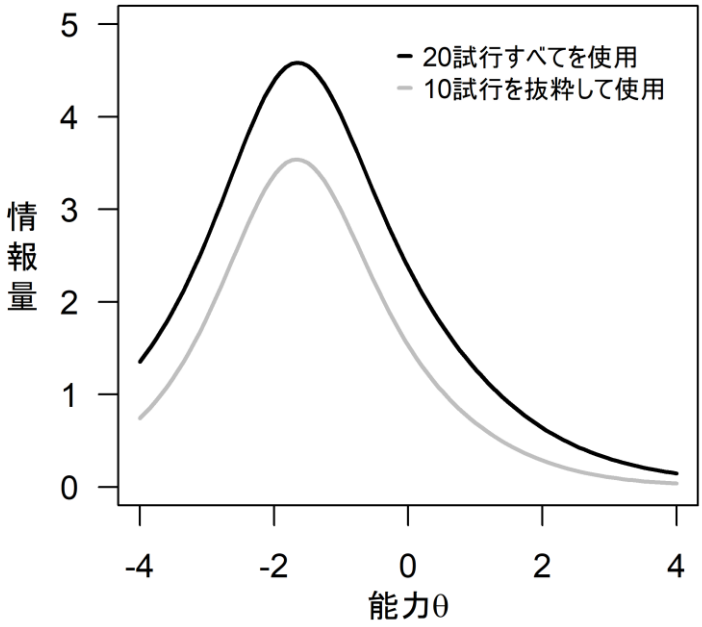



**図３－８　識別力による試行の選抜**

## ６　シミュレーション実行環境

本章のシミュレーションは統計ソフト R 4.0.1（R Core Team, 2021）を用いて行った。巻末資料に示した二つのスクリプト（scriptA_sampling.R と scriptB_irt.R）によって図３－５及び図３－６に示した結果を再現することができる。スクリプトを実行するために必要なパッケージは ks 及び ltm であり、事前にインストールする必要がある。

二つのスクリプトを scriptA_sampling.R と scriptB_irt.R の順番で実行すれば本章に掲載した通りの結果を得ることができる。結果のグラフは拡張子.png 形式のファイルとして出力され、生成したデータは拡張子.RData 形式のファイルに保存される。

一つ目のスクリプト scriptA_ sampling.R は工程１及び工程２に相当する処理を行う。このスクリプトは処理が完了するまでに長い時間を必要とする。PC によって処理時間は変わるが、クロック周波数が 3.6 GHz のプロセッサーを持つ PC で実行したところ処理の完了までにおよそ 15 分を要した。このように時間がかかるのは取得するサンプルの数が 10 万件と膨大だからである。試験的な実行を目的として処理時間を短縮する場合はスクリプトの冒頭にある nSample を小さい値に設定することでサンプル数を減らせばよい。nSample を 100 とした場合の処理時間はおよそ２秒であった。ただし、スクリプトに変更を加えてしまうと図３－５と同一の結果は得られないこと、また、二つ目のスクリプト scriptB_irt.R を実行できないことを注記しておく。

二つ目のスクリプト scriptB_irt.R は工程３及び工程４に相当する処理を行う。これによって図３－６を再現することができる。

## 第2節　応用面での課題（開発・運用面での条件整理）と展望

第１節では、MWS『社内郵便物仕分』を参考に、一般参考値の模擬データの作成プロセスとともに、項目反応理論を用いた分析プロセスを提示した。これらの分析によって、先行研究において提示された課題にどのような対応がなされたかについて検討したい。

### １　一般参考値に対する統計的検討

まず、第１章第３節（３）で指摘したイ.「一般的な標準値に相当するデータをワークサンプルにおいて整備することは極めて困難である」という点についてである。ここでいう困難とは、先行研究（障害者職業総合センター，2004）ではA.「標準化に必要な相当数の偏りないデータを集めること」、B.「障害者データについては母集団が規定しにくい」、C.「正規分布が想定できない」として挙げられていた。

A の問題については、コストと実現可能性の点で難しい問題となっている。どの程度の人数からデータを収集すれば標準値として妥当な規模となるのかは、作成しようとする評価指標の性質によって異なるといえる。一般的には心理検査等は、数百～千単位のデータ数で標準値が整備されているが、ワークサンプル幕張版は課題ごとに数十～百数十人程度のデータを元に一般参考値が作成されていた。

また、B の問題は、障害の個別性や多様性から、その障害のある母集団の分布状況を正確に想定することが難しい状況を意味している。したがって、仮に障害のある人の標準値を作成する場合には、診断名や障害種別といった大まかな情報だけではなく、障害程度やその他の属性に関する情報についての条件を厳密に定めた上で、条件に合致する障害者からサンプリングを行う必要があるということである。このような条件に合致する対象者から協力を得ることが現実的に難しいために、A の問題が喚起されているとみられた。

C の問題は、障害者のデータを標準値とする場合に、障害特性によって単純な正規分布を仮定することが必ずしも妥当でないことを示唆している。これらの問題に対して、第１節で行ったシミュレーションの取組は、次のような改善策を提案したといえる。

#### ①　シミュレーションによる大規模な模擬データの取得

A の問題に対して、既に得られた被検者データの統計量から、想定される大規模データの統計的モデルが提示された。なお、シミュレーションによる模擬データが、どれだけ現実のデータに近いのかは、既に得ている被験者データが母集団を適切に代表するようサンプルリングされていることが、必須の条件である。だが、このシミュレーションの利点は、実際には現実的に得ることのできない規模のデータを、実際のデータに近似させ作成することによって、仮想的なデータではあるものの、課題構成についての統計的な検討（例えば本研究では項目反応理論の利用）を可能とさせるものであることが明らかとなった。

#### ②　母集団の分布の想定

第１節では、シミュレーションによってサンプルの分布を仮定するプロセスが提示された。仮に、標準値とするサンプルの母集団が想定しにくいとしても、サンプリングを行う上では対象者の属性を研究者が定めた上で収集が行われるわけであり、事前に母集団の分布について仮説を立てること

は不可能ではない。したがって、想定される分布のモデルを三つ用意し、そのあてはまりの妥当性を検討したわけである。分析の結果、シナリオ C は最も妥当でないことが判明しており、「正規分布は想定されにくい」という先行研究の指摘は、まずは健常者データ（派遣労働者）において支持されたといえる。障害者に関する分布の検討は別途検討が必要である。

シミュレーションによる大規模データを用いることで、統計的に検討可能となる範囲が拡大することが判明した。

### 2 一般参考値の信頼性･妥当性の検討

次に、第１章第３節（３）で示したロ.「一般的な標準値に準ずるものとして扱っている“パーセンタイル順位による参考値”であっても、信頼性・妥当性の向上を検討することは必要性である」という点についての本調査研究における対応に触れたい。

#### ① 信頼性･妥当性向上のための項目分析の実施

本調査研究では、MWS『社内郵便物仕分』の品質向上につながる信頼性・妥当性の検討のため、それまでとは異なったデータの解析方法に取り組んだ。既存の一般参考値は、20 試行のうち何試行が正確に仕分けられたかの平均正答率及び、作業時間についてのパーセンタイル順位を提供していた。だが、正答率に関して言えば、20 試行全体（１ブロック）の平均正答率１値のみが提供されていたにすぎない（エラー内容に関する情報は先行研究で示されているが、この度の分析では取り扱っていない)。本調査研究では、シミュレーションによって被検者の 20 試行すべてについての正答の有無を変数として扱うことを可能にし、郵便物１試行（１枚）単位の難易度も取り入れた上で模擬データが作成され、これらを用いて項目反応理論による分析を行うことが可能であることを示した。これによって、これまで区別なく提供されていた同レベルの試行において、“困難度”や“識別力”という点での違いを見出せることが明らかとなった。試行ごとの特性をより詳細に把握することができれば、今以上に精度の高い課題設定が可能となり、アセスメントツールとしての信頼性・妥当性の向上や効率性の向上を図ることが可能になるといえる。

#### ② 項目分析による課題の改善方法

20 試行について各々に統計的な特徴（困難度、識別力を含む。）が得られた結果、どのようなメリットが生まれるだろうか。現在の「簡易版」の一つの機能である“MWS の課題の体験的な実施”が利用の目的であれば、時間をかけて 20 試行全てを行うのではなく、課題の品質を保ったまま 10 試行を実施するだけで済むかもしれない。MWS 以外のアセスメントツールであっても、例えば、大量の項目で構成されたチェックリストや、各作業課題についての統計的な特徴（困難度、識別力等）に基づいて項目の選定を行うことで、障害者の作業特性やその背景にある認知面・情報処理の特徴に対して、精度よく効率的に測定できる項目を支援者が選択することができる可能性がある。また、コンピューターの利用によって、作業遂行の状況に合わせてリアルタイムで最適な問題を自動で出題することも実現するかもしれない。

このように、ユーザーの目的やニーズに応じた実施方法の可能性が広がることからも、改めて、項目反応理論による各項目の性質を精査することの意義が示されたといえる。

## 3 今後の課題

本章では、MWS の『社内郵便物仕分』を参考にシミュレーションデータを作成した上で項目反応理論を実施した。前述したとおり、あくまで模擬データによる検討である。したがって、実際のデータを用いた再検討が別途必要である。また、本検討は、MWS の課題のみならず、その他のアセスメントツールにおいても、そのツールで何らかの評価指標が提供されているのであれば、適用可能性の高い方法である。今後は、アセスメントツールの作成・改修を行うに当たり、これらのシミュレーションや項目反応理論等を含む新たなデータ解析手法を積極的に取り入れていくことが必要であると考える。

## 【文献】

R Core Team（2021）. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

障害者職業総合センター　調査研究報告書№57（2004 年）「精神障害者等を中心とする職業リハビリテーション技法に関する総合的研究（最終報告書）」

障害者職業総合センター　調査研究報告書№145（2019 年）「障害の多様化に対応した職業リハビリテーション支援ツールの開発（その２）―ワークサンプル幕張版（MWS）新規課題の開発」

# おわりに

# おわりに

本調査研究では、職業リハビリテーションサービスの対象者における状態像の多様化に伴って、職業評価やアセスメントにおいて客観性や効率性が求められる現状に対して、利用されるアセスメントツールの品質向上が必要であることを示した。また、品質向上のためには、ツール開発のプロセスにおいて適切な手続を踏まえること、加えて、開発に伴う体制整備も必要となる可能性があること等を示したうえで、近年、様々な分野で利用されているテスト理論の考え方を、アセスメントツールの品質向上の手段として用いる可能性について指摘した。

テスト理論の利用可能性を具体的に検討するため、MWS の『社内郵便物仕分』を題材に、現代テスト理論のうち項目反応理論による統計手法を試みた。その結果、課題を構成する各項目について、さらに詳細な基礎統計量を算出することが可能であることが明らかとなった。また、これらの統計的な手法は、他のアセスメントツールで広く応用できる可能性があることから、今後、ツール開発・改修のプロセスにこれらの手法を積極的に加えていくことが期待される。

# 巻末資料

【統計ソフトR 4.0.1を用いたスクリプト】

# 統計ソフトR 4.0.1を用いたスクリプト

scriptA_sampling.R

```
# source('scriptA_sampling.R')

### シナリオとサンプリング方法の指定
nSample <- 100000 # 生成するサンプルの数

#DistName <- "unif"; fAgeFilter <- 0 # シナリオA, 全体サンプリング
#DistName <- "unif"; fAgeFilter <- 1 # シナリオA, 年代別サンプリング
DistName <- "unim"; fAgeFilter <- 0 # シナリオB, 全体サンプリング
#DistName <- "unim"; fAgeFilter <- 1 # シナリオB, 年代別サンプリング
#DistName <- "norm"; fAgeFilter <- 0 # シナリオC, 全体サンプリング
#DistName <- "norm"; fAgeFilter <- 1 # シナリオC, 年代別サンプリング

### 以下、シミュレーションの本体
cat(" ### ", format(Sys.time(), "%H:%M:%S"), " START¥n", sep="")
flush.console()
set.seed(SetSeed <- 1)
t1 <- Sys.time()
para <- list(SetSeed=SetSeed, nRep=1000, nSample=nSample, DistName=DistName, fAgeFilter=fAgeFilter)
savefileName <- paste0("sampling-", DistName, "-", fAgeFilter)

# 一般参考値
ref <- list(
  age20=list(
    title="20s",
    bks=c(0, 57.5, 65, 65, 75, 77.5, 85, 87.5, 90, 90, 100),
    N=c(10, 20, 10, 10, 10, 10, 20, 10),
    mean=76.3, sd=14.5, sampleSize=26, peak=76.3,
    estMean=0.76992, estSD=0.1410640),
  age30=list(
    title="30s",
    bks=c(0, 55, 60, 64.5, 70, 80, 85, 90, 90, 95, 100),
    N=c(10, 10, 10, 10, 10, 10, 20, 10, 10),
    mean=75.7, sd=16.7, sampleSize=34, peak=75.7,
    estMean=0.7694092, estSD=0.1713303),
  age40=list(
    title="40s",
    bks=c(0, 50, 65, 70, 75, 80, 85, 85, 90, 95, 100),
    N=c(10, 10, 10, 10, 10, 20, 10, 10, 10),
    mean=76.6, sd=18, sampleSize=56, peak=76.6,
    estMean=0.7733987, estSD=0.1579981),
  age50=list(
    title="50s",
    bks=c(0, 55, 65, 70, 75, 75, 80, 85, 85, 90, 100),
    N=c(10, 10, 10, 20, 10, 20, 10, 10),
    mean=73.9, sd=16.2, sampleSize=46, peak=73.9,
    estMean=0.7550417, estSD=0.1341714)
  )

res <- matrix(NA, ncol=7, nrow=nSample)
colnames(res) <- c("mean", "sd", "age20", "age30", "age40", "age50", "all")

samplePool <- list()

# サンプリング
for (nsam in 1:nSample) {

  d <- data.frame()

  # 乱数生成
  for (page in 1:length(ref)) {

    x_m <- ref[[page]]$mean
    x_sd <- ref[[page]]$sd

    q <- ref[[page]]$bks
    names(q) <- paste(seq(0, 100, 10), "%", sep="")
    q.tick <- q
    q.tick <- unique(q.tick)

    N <- ref[[page]]$N / 10 * para$nRep

    ss <- c()
    r.seq.res <- c()
    probs.res <- c()
    for (i in 1:(length(q.tick)-1)) {

      r <- c(q.tick[i] + 1e-3, q.tick[i+1] - 1e-3)
      r.seq <- seq(r[1], r[2], length.out=100)

      if (DistName=="unif") { # シナリオA
        s <- sample(x=r.seq, size=N[i], replace=T)
```

```
      probs <- rep(1, 100)
      }

    if (DistName=="unim") { # シナリオB
      a <- matrix(c(ref[[page]]$peak, 1, 0, 1), nrow=2, byrow=T)
      b <- matrix(c(2, 1))
      UpLine <- solve(a, b)
      a <- matrix(c(ref[[page]]$peak, 1, 100, 1), nrow=2, byrow=T)
      b <- matrix(c(2, 1))
      DownLine <- solve(a, b)
      probs <- ifelse(r.seq < ref[[page]]$peak,
      UpLine[1] * r.seq + UpLine[2],
      DownLine[1] * r.seq + DownLine[2])
      }

    if (DistName=="norm") { # シナリオC
      probs <- dnorm(x=r.seq, mean=ref[[page]]$estMean*100, sd=ref[[page]]$estSD*100)
      }

    r.seq.res <- c(r.seq.res, r.seq)
    probs.res <- c(probs.res, probs)

    s <- sample(x=r.seq, size=N[i], replace=T, prob=probs)
    ss <- c(ss, s)
    }

  new <- data.frame(g=ref[[page]]$title, val=ss)
  d <- rbind(d, new)
  }

# 取得したサンプルの基礎統計（平均値、標準偏差、一般参考値との相関）
q <- c(0, 55, 60, 70, 75, 80, 85, 85, 90, 95, 100) # 郵便物仕分（簡易版）　全体
cors <- rep(NA, 4)
if (fAgeFilter==1) { # 年代別サンプリング
  sfd <- rep(NA, 162)
  sfd[  1: 26] <- sample(x=d$val[d$g=="20s"], size=26, replace=TRUE)
  sfd[ 27: 60] <- sample(x=d$val[d$g=="30s"], size=34, replace=TRUE)
  sfd[ 61:116] <- sample(x=d$val[d$g=="40s"], size=56, replace=TRUE)
  sfd[117:162] <- sample(x=d$val[d$g=="50s"], size=46, replace=TRUE)
  sampled <- data.frame(g=c(rep("20s", 26), rep("30s", 34), rep("40s", 56), rep("50s", 46)), val=sfd)

  } else { # 全体サンプリング
  select_sampling_order <- sample(x=1:dim(d)[1], size=162, replace=TRUE)
  sampled <- d[select_sampling_order, ]
  sfd <- sampled$val

  }

cors[1] <- cor(ref$age20$bks, quantile(sampled$val[sampled$g=="20s"], probs=seq(0, 1, .1)))
cors[2] <- cor(ref$age30$bks, quantile(sampled$val[sampled$g=="30s"], probs=seq(0, 1, .1)))
cors[3] <- cor(ref$age40$bks, quantile(sampled$val[sampled$g=="40s"], probs=seq(0, 1, .1)))
cors[4] <- cor(ref$age50$bks, quantile(sampled$val[sampled$g=="50s"], probs=seq(0, 1, .1)))
x_m <- mean(sfd)
x_sd <- sd(sfd)
q.est <- quantile(sfd, prob=seq(0, 1, .1))

# サンプルの要約統計量を保存
res[nsam, 1:7] <- c(x_m, x_sd, cors[1], cors[2], cors[3], cors[4], cor(q, q.est))

# 精度のよいサンプルを保存
if (mean(res[nsam, 3:7]>0.99)==1) {
  if (x_m > 75.6-0.1 & x_m < 75.6+0.1 & x_sd > 16.6-0.1 & x_sd < 16.6+0.1) {
    samplePool <- c(samplePool, list(list(group=sampled$g, value=sampled$val, mean=x_m, sd=x_sd)))
    }
  }

# コンソールに時刻を表示
if (nsam %% (nSample/10) == 0) {
  cat(" ### ", format(Sys.time(), "%H:%M:%S"), " ", round(nsam/nSample*100), "%¥n", sep="")
  flush.console()
  }
}

# データの保存
t2 <- Sys.time()
env <- list(sessionInfo=sessionInfo(), start=t1, finish=t2, timeElapsed=t2-t1)
save(res, para, env, samplePool, file=paste0(savefileName, ".RData"))

# 作図
png(file=paste0(savefileName, ".png"), width=200, height=200, units="mm", res=300)
xR <- c(66, 82); yR <- c(10, 30)
res <- data.frame(res)
plot(res[, 1], res[, 2], xlim=xR, ylim=yR, pch=16, col="gray70", ann=F, axes=F, cex=0.5)
par(new=T)
plot(mean(res[, 1]), mean(res[, 2]), xlim=xR, ylim=yR, pch=3, ann=F, axes=F)
```

```
res2 <- res[rowMeans(res[, 3:7]>0.99)==1, ]
res2 <- res2[res2[,1] > 75.6-0.1 & res2[,1] < 75.6+0.1 & res2[,2] > 16.6-0.1 & res2[,2] < 16.6+0.1, ]
par(new=T)
plot(res2[,1], res2[,2], xlim=xR, ylim=yR, pch=16, col="black", cex=0.25, ann=F, axes=F)
box()
axis(side=1, at=seq(66,82,2), labels=seq(66,82,2))
axis(side=2)
title(main="", xlab="正答率の平均", ylab="正答率の標準偏差")
text(x=min(xR), y=max(yR), label=paste0("平均 ", round(mean(res[,1]), 2), ",  標準偏差 ", round(mean(res[,2]), 2)),
adj=c(0, 1))
text(x=min(xR), y=max(yR)-1.3, label=paste0("精度のよいサンプルの数 ", dim(res2)[1]), adj=c(0, 1))

# サンプルが主に分布している位置を示す
kd <- ks::kde(data.frame(x=res[,1], y=res[,2]), compute.cont=TRUE)
contour_50 <- with(kd, contourLines(x=eval.points[[1]], y=eval.points[[2]], z=estimate, levels=cont["50%"])[[1]])
contour_50 <- data.frame(contour_50)
polygon(contour_50[,2], contour_50[,3], border="white", lty=1)

# 一般参考値の範囲
polygon(x=c(75.6-0.5, 75.6+0.5, 75.6+0.5, 75.6-0.5), y=c(16.6-0.5,16.6-0.5,16.6+0.5,16.6+0.5), border="black")

# 図のタイトル
scenario <- switch(DistName, "unif" = "シナリオA", "unim" = "シナリオB", "norm" = "シナリオC")
sampling <- switch(as.character(fAgeFilter), "0" = "全体サンプリング", "1" = "年代別サンプリング")
par(xpd=T)
text(x=min(xR), y=max(yR)+2, labels=paste0(scenario, ", ", sampling), adj=c(0,0), cex=1.2)
par(xpd=F)

# 相関係数のヒストグラム
u <- par("usr")
v <- c(grconvertX(u[1:2], "user", "ndc"), grconvertY(u[3:4], "user", "ndc"))
v <- c( (v[1]+v[2])/2, v[2], (v[3]+v[4])/2, v[4] ) + c(0.1, 0, 0.1, 0) # Upper-right
par(fig=v, new=TRUE, mar=c(0,0,0,0))

res3 <- res[rowMeans(res[, 3:7]>0.90)==1, ]
plot(NA, xlim=xR, ylim=yR, ann=F, axes=F)
rect(par("usr")[1],par("usr")[3],par("usr")[2],par("usr")[4],col = "grey95")
par(new=T)
hist(res3[, 7], xlim=c(0.9, 1), breaks=seq(0.9, 1, 0.002), col="grey", border="grey", main="", ann=F, axes=F)
box()
axis(side=1, at=c(0.94, 0.96, 0.98, 1), tck=-0.03, cex.axis=0.5, padj=-3.5, hadj=0.8)

# 終了
cat(" ### ", format(Sys.time(), "%H:%M:%S"), " FINISH¥n", sep="")
dev.off()
```

## scriptB_irt.R

```
# source("scriptB_irt.R")

set.seed(1)
library(ltm)

# データの読み込み
load("sampling-unim-0.RData")
d <- data.frame(g=samplePool[[2]]$group, pCor=samplePool[[2]]$value)
d$nCor <- round(d$pCor/100*20)

# 難易度の設定
model1 <- (-1*(-5 + c(2,1,1,1,1,3,2,3,4,2,2,5,3,2,5,2,4,4,4,3))+1)/5 # 評定者A
model2 <- c(3,3,3,3,3,2,3,3,1,3,2,1,3,2,1,3,1,1,1,1)/3 # 評定者B
model3 <- c(3,3,3,3,3,2,2,2,3,2,3,2,2,3,1,2,2,1,2,1)/3 # 評定者C
item.p <- colMeans(rbind(model1, model2, model3)) # 3名の評定者の平均

# 正誤の補完
u <- matrix(data=NA, nrow=dim(d)[1], ncol=20)
for (i in 1:dim(d)[1]) {
  s <- sample(x=1:20, size=d$nCor[i], replace=F, prob=item.p)
  soten <- rep(0, 20)
  soten[s] <- 1
  u[i, 1:20] <- soten
  }

# 項目反応理論を適用
d <- data.frame(d, u)
irt <- ltm(u~z1, IRT.param=T)

# 作図
png(file="irt.png", width=210-40, height=100, units="mm", res=300)
par(mfrow=c(1,2))
plot(irt, xlab="能力", ylab="正答する確率", main="項目特性曲線", label=rep(NA,20), col=rep(1,20))
arrows(x0=1.3, y0=0.2, x1=0, y1=0.37, length=0.1)
```

```
text(x=1.3, y=0.2, label="問題 15", adj=c(0,1))
plot(irt, xlab="能力", ylab="情報量", main="テスト情報量曲線", ylim=c(0,5), type="IIC", item=0)
dev.off()
```

ホームページについて

本資料のほか、障害者職業総合センターの研究成果物については、一部を除いて、下記のホームページから PDF ファイルによりダウンロードできます。

【障害者職業総合センター研究部門ホームページ】

https://www.nivr.jeed.go.jp/



【連絡先】

障害者職業総合センター研究企画部企画調整室

電話　０４３－２９７－９０６７

FAX　０４３－２９７－９０５７

資料シリーズ　No.104

職業リハビリテーションのアセスメントにおける現代テスト理論の応用可能性に関する基礎的研究

編集　　独立行政法人高齢・障害・求職者雇用支援機構
障害者職業総合センター
〒261-0014
千葉市美浜区若葉３－１－３
電話 043-297-9067
FAX 043-297-9057

発　行　日　　2022 年３月