資料論文

項目応答理論の分析モデル概容と日本の数学関連テストにおける利用動向

Review on the Application and Implementation of Item Response Theory in Mathematics-Related Test Making in Japan

大床　太郎[*1]，堀江　郁美[*2]
Taro Ohdoko, Ikumi Horie

e-mail:ohdoko@dokkyo.ac.jp

keywords: Item Response Theory, Mathematics-Related Test Making, Computer Adaptive Test

日本の数学教育における高大連携・大学リメディアル教育の必要性に鑑みるに，数学テスト作成における項目応答理論ないし項目反応理論（item response theory: IRT）の利用拡大，システム実装と実践が喫緊の課題である．IRT を利用することで，適切に数学的素養が能力値として測定され，コンピュータ適応型テスト（computerized adaptive test: CAT）との統合で，テスト受験者に無理のないテスト項目が提示でき，テスト作成も効率的に行うことが可能となる．

本稿では，独自の数学テスト作成・実施システム構築を見据えて，まず重要となる IRT の分析モデルの概容を整理し，数学のテスト受験者・テスト作成者・システム構築者を含めた「ユーザー」を念頭に日本の数学関連テスト分野における利用動向を整理した．結果として，受験者サンプルの確保，数学テスト項目の項目銀行の作成，CBT(computer based test)版数学テストの実現可能性模索，CAT との統合の順に推進すべきことを確認した．

It has been a pressing issue in Japanese universities to apply and implement the item response theory (IRT) to mathematical test-making because of the necessity of having to offer high school and university collaborative education and/or remedial education at university. Application and implementation of IRT enables us to estimate the mathematical ability of students properly, to provide examinees with reasonable test items along with a computer adaptive test (CAT), and to make mathematics tests efficiently.

In this article, we summarize previous Japanese research on IRT along with related foreign research studies before we create a unique mathematical test-making system. As a result, when we are to apply and implement IRT in mathematical test-making, it is recommended that, first, we secure a sufficiently large sample of examinees; second, we create mathematical test item banking; third, we examine feasibility of a computer based mathematical test with IRT; and, fourth, we integrate IRT with CATs.

[*1] 獨協大学経済学部国際環境経済学科，獨協大学情報学研究所
[*2] 獨協大学経済学部経営学科，獨協大学情報学研究所

1. はじめに

日本における近年の高大連携教育の必要性あるいは少子化への教育的対応に鑑みるに，数学関連分野を必要とする大学の学部教育において，大学のリメディアル教育は必要不可欠なものになりつつある．理学・医学・薬学・工学などに代表される自然科学分野のみならず，経済学・心理学などの社会科学分野においても数学的素養が必須となった昨今，今まで以上に有効かつ継ぎ目ない数学教育の展開が必要とされている．数学的素養あるいは能力について適切に測定・評価することで，より社会科学的に教育指導するための有効な方針と目標が立てられる．したがって，今まで以上に適切な数学能力の測定・評価が希求されている．

普遍的能力測定に関しては，特に言語テストの文脈で発展が著しい．TOEFL・TOEIC に代表される英語能力試験の分野では，項目応答理論ないし項目反応理論（item response theory: IRT）の利用がなされ，よりよい能力推定に向けた研究と実践が盛んになされている（大友 2009）．TOEFL・TOEIC などは，日本の就職活動やキャリアデザインに用いられてもいる．しかしながら，たとえば各種業務で実際に扱うソフトウェアの理論を構成し，あるいは論理的思考と切り離すことの難しい数学において，IRT の研究と実践は緒についたばかりであると言ってよい．大学の学部生が社会人候補生であるならば，言語的・コミュニケーション的な能力のみならず，数学的な素養も適切に評価されることで，その人が社会におけるどのような未来を担うのかを把握するシグナルとして有効に機能することになろう．

実際に，数学的素養に対する社会の注目は著しい[i]．統計学に関するテキストや解説書が多数刊行され，インターネット上のビッグデータの利用が謳われている．数学と不可分である情報という分野が高校教育課程に明示的に導入されてもいる．1995 年に第 1 回として開催された TIMSS（Trends in International Mathematics and Science Study）や，2000 年から実施されている PISA（Programme for International Student Assessment）の数学的リテラシーに関する調査項目に鑑みるに，国際的な学力比較において数学の占めている位置は極めて高い（Cf. 川口 2014）．浦坂ほか（2002，2010）にいたっては，日本の主要 3 私立大学の社会科学系・自然科学系の入試試験において数学を受験するか否かが，将来の所得や昇進といったキャリア形成に影響を及ぼしていることを指摘している．以上に鑑みるに，高大連携や大学教育における数学に対して，社会から高い期待が寄せられているといえよう．

本稿は，特に数学的素養あるいは能力の把握に，IRT がどのように寄与するのか，邦文の文献を中心として整理することを目的とする．今後，独自に数学テスト作成・実施システムを構築する際に，どのような点に留意すべきかを整理する．特に，高大連携と大学のリメディアル教育，および「ユーザー」を念頭におく．ここで「ユーザー」とは，数学のテスト受験者のみを指すものではない．テスト作成者，あるいはシステム構築者と想定される，教員を中心とした高校・大学関係者も含めることとする．

本稿の構成は以下のとおりである．まず 2 節にて，IRT が何を測ろうとしているのか，それはなぜか，どのように測るのか，分析モデルを概観し，「ユーザー」を念頭に整理する．3 節にて，数学関連テスト分野でどのように IRT の利用が取り組まれているのか，邦文文献を整理し，とりわけ本稿の筆者が今後に独自に数学テスト作成・実施システム構築を実施する際，どのような留意点があるか吟味する．4 節にて，総括と今後の課題を提示し，結語とする．

2. IRT 分析モデルの概容

2.1 IRT の分析モデル

本節では，豊田（2002）・大友（1996）を中心とした IRT の分析モデルの整理を通じて，IRT が何を測定しようとしているのかを明らかにし，「ユーザー」を念頭に整理する．なお，応答変数データは 0，1 で表現されていること，すなわち，テスト項目に誤答したか正答したかを応答変数とすること，そして，受験者の能力などを独立変数として定式化することを仮定する．

IRT は，受験者の能力値を連続変数の潜在特性（latent trait）として捉える潜在変数モデルである．潜在特性すなわち能力値θをもつ受験者がテスト項目jに正答する確率$p_j(\theta)$は項目特性曲線（item characteristic curve）という．この項目特性曲線に標準正規分布の累積分布をあてはめたものを正規累積モデル（normal ogive model），ロジスティック分布の累積分布をあてはめたものをロジスティックモデル（logistic model）という．前者が Lord (1952) において提示されたが，後者は以下に示すように累積分布の積分がない陽表的なモデルであり，取り扱いが容易であるため，現在では頻繁に用いられている．

ロジスティックモデルとして，3 つのモデルが提案されている．なお，以下におけるDは，正規

累積モデルにロジスティックモデルを近似させる定数として Birnbaum（1968）が提案した尺度係数（scaling constant）であり，通常$D = 1.7$と設定される[ii].

・1 パラメータロジスティックモデル
受験者の正答確率が以下のように表現されるモデルである.

$$p_j(\theta) = 1/\left(1 + \exp\left(-Da\left(\theta - b_j\right)\right)\right) \quad \text{(Eq.1)}$$

ここで，b_j はテスト項目 j の困難度（item difficulty），つまりテスト項目jがどれほど難しいかを示すパラメータである．項目困難度b_jに加えて能力値θがパラメータとして加わり，推定されることとなる．豊田（2002）によれば，米国流 1 パラメータ正規累積モデルの近似がこのモデルであり，また，欧州で独自に発展したラッシュモデル（Rasch model）とも対応するモデルである.

・2 パラメータロジスティックモデル
受験者の正答確率が以下のように表現されるモデルである.

$$p_j(\theta) = 1/\left(1 + \exp\left(-Da_j\left(\theta - b_j\right)\right)\right) \quad \text{(Eq.2)}$$

ここで，a_j はテスト項目 j の識別力（item discrimination），つまりテスト項目jが正答者と誤答者をどれほど明確に区別するかを表すパラメータである. 項目困難度b_jと能力値θに加えて項目識別力a_jがパラメータとして加わり，推定されることとなる.

・3 パラメータロジスティックモデル
受験者の正答確率が以下のように表現されるモデルである.

$$p_j(\theta) = c_j + \left(1 - c_j\right)/\left(1 + \exp\left(-Da_j\left(\theta - b_j\right)\right)\right) \quad \text{(Eq.3)}$$

ここで, c_jは擬似偶然水準（pseudo-chance level）ないし当て推量パラメータ（guessing parameter），つまり，「実力では全く正解できない」受験者が「偶然に正答してしまう」確率を表すパラメータである（豊田 2002）.

以上のモデルにおいて，3 パラメータロジスティックモデルは 2 パラメータ・1 パラメータのモデルの上位モデルと考えることができる．すなわち，2 パラメータモデルは擬似偶然基準をゼロ，1 パラメータモデルはそれに加えて項目識別力を一定とみなして分析するモデルである.

以上より IRT は，受験者の能力値（θ），テスト項目の困難度（b_j），テスト項目が正答者と誤答者を明確に区別する度合い（a_j），偶然に正答してしまう可能性（c_j）を完全に分離して測定する．また，受験者能力値θは$(-\infty, \infty)$として定義されることから，0 点～100 点などの素点の持つ天井効果・床面効果と呼ばれる定義域の課題も克服しようとしている.

「ユーザー」を念頭におけば，能力値の提示は受験者にとって必要とされるテストのフィードバックである．困難度・識別力・疑似偶然水準の各パラメータは，たとえば予備的なテストによって推定された結果をテスト項目と合わせてデータベース化することで，テスト作成者・テストシステム構築者にとって有用な指標となる．ただし，植野ほか（1994）の言葉を借りるならば，IRT の分析モデルは「学習者選抜のための効用関数」を仮定しているため，識別力や疑似偶然水準のパラメータは「ユーザー」の一端である受験者へのフィードバックを主眼にする際には優先順位が低い．数学テスト作成・実施システム構築の際には，1 パラメータの，あるいは既往研究で適用例の多い 2 パラメータのモデルまでが現実的と思われる.

2.2　IRT におけるパラメータの推定

まず，IRT の推定においては，能力値θの標準化が仮定されることが多い．IRT では，最も未知パラメータの多い 3 パラメータモデルでは，能力値θ・項目困難度b_j・項目識別力a_j・擬似偶然水準c_jを推定する必要がある. 受験者$n\ (n = 1, \cdots, N)$, テスト項目$j\ (j = 1, \cdots J)$とすると，未知のパラメータは受験者の能力パラメータN個，テスト項目のパラメータ 3 種類で計3J個を推定しなければならない．すべてのパラメータが未知の状態で尺度値の平均や分散に意味を見出し難いθについては, 通常平均 0，標準偏差 1 として標準化されて推定が行われ, 未知パラメータの数も$(N + 3J - 2$へ）減少することとなる．したがって，0 点以上の尺度などの受験者に意味のある「IRT 得点」として提示したい場合には，たとえば項目困難度パラメータb_jを予備的テスト実施で得られた推定値を用いて所与とするなど，別途工夫が必要となる[iii]. この場合，予備的テスト段階において豊かなサンプル数を確保する必要がある.

次に，パラメータの推定において，局所独立の仮定（local independence assumption）がおかれる（大友 1996）. すなわち, 受験者のテスト項目に対する反応は互いに独立である，ないし「用いられるテスト項目はただひとつの能力分野を測定する（大友 1996）」ものであるという仮定である．たとえば，項目反応の因子分析によって,「第 1 因子の説明できる分散が全体の 50%前後以上であり，第 2 以下

の因子に格段の差を生じていること（大友 1996）」が指標となる．一次元性を検証する方法は，塗師（1989）によれば，因子分析・回答パターン・IRT を用いる方法がある．しかしながら，塗師（1989）は「一次元性という概念が本来的に明確でない」とも指摘しており，長崎（1994）によれば，数学的素養は，1) 行動類型：数学の知識・理解・思考・技能・態度，2) 数学内容：数式的・図形的・関係的，3) 数学過程：数学化・数学的処理・数学的検証として 3 次元の構成概念に整理されるなど，多次元性を有している可能性がある．複数の能力の構成概念が項目反応に影響を及ぼしていると想定できる場合は多次元性（multidimensionality）へと分析モデルを拡張する必要がある．たとえば孫（1997）の一般項目反応モデル（generalized item response model），Adams et al.（1997）の多次元ランダムパラメータロジットモデル（multidimensional random coefficients multinomial logit model），Lindsay et al.（1991）の潜在クラス分析（latent class analysis）などへ展開する必要が生じる．モデルが高度になるにしたがって，より豊かなサンプル数が必要となることは容易に予想される．また，「たとえば同じような質問を何回も繰り返せば一次元性はたしかに高くなる（塗師 1989）」ことから，テスト項目をたとえばただ単に数値を変えただけで提示するようなことは避け，同一単元に対して多様なテスト項目を用意する必要もある．

局所独立が仮定されれば，受験者 n $(n = 1, \cdots, N)$ のテスト項目 j $(j = 1, \cdots J)$ に対する反応を u_{nj} とすると，正答 1・誤答 0 の 2 値反応を仮定すれば，受験者 n の反応ベクトル $u_n = [u_{n1}, u_{n2}, \cdots, u_{nJ}]$ が観察される確率は

$$f(u_n|\theta_n) = \prod_{j=1}^{J} p_j(\theta_n)^{u_{nj}} \left(1 - p_j(\theta_n)\right)^{1-u_{nj}} \quad \text{(Eq.4)}$$

という確率の積の表現としての同時確率として表される．あるテスト項目がより上位（大問）のテスト項目の一部を構成するような入れ子型構造をとっていたり，あるテスト項目の回答能力が次のテスト項目の回答能力に必須であるような逐次型構造をとっていたりする場合には，入れ子型のテスト構造に対応した段階反応モデル（graded response model，渡部・平井 1993，豊田 2002 など）や，項目間の相関を許す安田・田中（2012）のようなアプローチなどによって，この仮定の緩和を検討する必要がある[iv]．仮定を緩和すれば，より豊かなサンプル数が必要となろう．

項目パラメータや能力値を推定する際には，最尤推定やベイズ推定が用いられる（豊田 2002, de Ayala 2009, Fox 2010 など）[v]．最尤推定は全問正答（100 点）ないし全問誤答（0 点）の場合の能力尺度の最尤推定値が無限大に発散するために，事実上，解を求められていないに等しくなる（豊田 2002）．全受験者が正答・誤答した項目についても，同様の課題が生じる．したがって最尤推定では，全問正答・全問誤答者，全受験者が正答・誤答した項目に対する反応データを除く処理が必要となる（大友 1996）．ベイズ推定は，全問正答・誤答の場合でも能力値を推定することができる一方で，適切な事前分布設定が必須となる．すなわち，全問正答・誤答者や全受験者が正答・誤答したテスト項目の扱いをどのようにすべきか，適切な事前分布が設定できるかによって，推定方法を選択しなければならない．また，同時最尤推定（joint maximum likelihood estimator）はとりわけサンプル数が必要で，大友（1996）ではロジスティックモデルを想定したときに，1 パラメータのモデルでは 100～200, 2 パラメータのモデルでは 200～400，3 パラメータのモデルでは 1000～2000 の受験者サンプル数が必要であることに触れている．また de Ayala（2009）は，たとえばテスト項目 25 以上で 2 パラメータロジスティックモデルを想定した場合，1000 サンプルは必要であろうと述べている．「必要とする最小標本数（minimal sample size required）」（大友 1996）と確保できるサンプル数も念頭に，適切な推定方法を選択する必要がある．

加えてパラメータ推定について，田中ほか（2004）は，受験者能力値 θ が既知のときの同時最尤推定，および未知の θ について周辺化したうえで推定する周辺最尤推定（marginal maximum likelihood estimation）を古典的な最尤推定として「直接法」，EM アルゴリズムを用いる方法を「EM 法」と整理して，手法の特性を数値解析しており，サンプル数やテスト項目数に応じて適切な推定方法も選択する必要があることを示唆している．

「ユーザー」を念頭におけば，全問正答・全問誤答が発生しやすいであろう，テスト作成・実施システム構築初期は当該回答データも扱うようなベイズ推定が魅力的であり，項目困難度パラメータ b_j などを推定する予備的テスト段階もそれに準ずる（Cf. 村木 2001，Bock and Mislevy 1988）．システム構築が進むにつれて，段階反応モデル，項目間の相関を許すモデル，さらには多次元性を考慮したモデルも推定できるように，そして「直接法」や「EM 法」による最尤推定とベイズ推定すべての実施と比較もできるように，サンプル数を豊かに確保することが重要となる．また，テス

ト項目としては，数学テスト作成・実施システム構築においては，数値を変えただけの問題を提示するようなことは避け，同一単元について豊かなテスト項目を蓄積する必要があろう．

2.3　IRT におけるテスト/項目情報関数

IRT におけるテスト情報関数とは，一般的に以下のような式で表される（de Ayala 2009）．

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta) \qquad \text{(Eq.5)}$$

ここで$I_j(\theta)$は項目情報関数であり，受験者の正答確率$p_j = p_j(\theta)$で表される．

$$I_j(\theta) = {p_j}^2 / p_j(1 - p_j) \qquad \text{(Eq.6)}$$

テスト情報関数は項目情報関数の和で表現されており，3 パラメータロジスティックモデルで整理すると，以下のような形をとる（豊田 2002，田中ほか 2004 など）[vi]．

$$I(\theta) = D^2 \sum_{j=1}^{J} [{a_j}^2 \left(p_j(\theta_j) - c_j\right)^2 \left(1 - p_j(\theta_j)\right) / p_j(\theta_j)\left(1 - c_j\right)^2]$$
$$= E\left[\left(\partial \log L(\theta|u_n)/\partial\theta\right)^2|_{\theta=\theta_n}\right]$$
$$= 1/V(\hat{\theta}|\theta) \qquad \text{(Eq.7)}$$

ここで$\log L(\theta|u_n)$は受験者能力値θの対数尤度関数であるため，テスト情報関数は Fisher 情報量に他ならず，θが与えられたときの$\hat{\theta}$の分散の逆数，すなわち推定の精度を測定している．θを横軸にとってテスト情報関数をプロットすればテスト全体の，項目情報関数をプロットすれば項目ごとに，どのような能力値θの推定精度が最も高かったかを示すことができる．

テスト情報関数や項目情報関数の利用によって，テスト・項目のターゲットにすべき受験者の能力値が分かる[vii]．したがって，能力に応じて無理のないテストを実施する適応型テスト（adaptive test）に効力を発揮する．たとえば阿久津・石亀（2012）は，岩手大学で実施された心理学の講義の中間・期末試験の分析から得られたテスト情報関数を用いて，テスト情報関数のすそ野が広い（fat tail）試験の方が，能力がとりわけ高い，あるいは低い受験者にも適切な試験であったことを示唆している．

「ユーザー」を念頭におけば，テスト情報関数・項目情報関数は，事前の予備的テスト実施によってテスト項目の組合せ・テスト項目それぞれについてデータベース化されることで，テスト作成者・実施システム構築者がよりよいテストを作成・実施するための指標が提示される．また，とりわけコンピュータ適応型テスト（computerized adaptive test: CAT）で実施された場合は，「ユーザー」の一端である受験者にとって無理のない問題が提示されることにつながる．

2.4　IRT 関連ソフトウェア

Rasch モデル専門のソフトウェアとして，有償である Widely used, versatile Rasch Analysis and Rasch Measurement Software (WINSTEPS) & Facets[viii]が，また，IRT 全般では，有償の BILOG-MG/MULTILOG（Rupp 2003）や，無償の統計解析環境 R の専用パッケージ ltm[ix]，日本発のソフトウェアとして無償の EasyEstiamtion [x]や Exametrika [xi]などがある．また，ベイズ推定できるソフトウェアとしては，Java で書かれた無償の Waikato Environment for Knowledge Analysis (Weka): Class BayesianLogisticRegression（Genkin et al. 2007）[xii]，R のパッケージである Bayesian inference Using Gibbs Sampling (BUGS)/WinBUGS [xiii]などがある．もしランダムパラメータロジットモデルに拡張する必要があれば mlogit[xiv]，潜在クラス分析に拡張する場合は flexmix (Grun and Leisch 2008)という無償の R パッケージも存在する．

「ユーザー」を念頭におけば，テスト作成者・実施システム構築者にとっては，1) 無償であることがまずは重要であり，2) CAT などとの親和性の高いソフトウェアが魅力的といえる．ただし，「直接法」「EM 法」による最尤推定やベイズ推定さらには多次元性を考慮した推定方法に対応した豊かなサンプル数の確保，同一単元について豊かなテスト項目の蓄積の必要性に鑑みるに，まずはシステム構築初期ないし予備的テスト段階からサンプル数を確保し，項目銀行（item banking）を作成してテスト作成者・実施システム構築者支援を充実させる必要がある．しかるのちに受験者のニーズも踏まえて CAT との統合を考え，利用するサーバー等システム要件を実験的に精査し，CAT 内に能力値等のパラメータ推定機能を含めるかどうかを判断すべきである．CAT 内に機能を含めるかどうかを検討する段階でようやく，Weka などの Java プログラムを参考に，独自に CAT システム内に実装すること[xv]，計算量があまり大きな場合はいったん CAT システム外に分析を受け渡すことが比較されよう．一方で受験者からは，もちろん自身に無理のない問題提示が望ましいものの，自身の能力値が TOEFL ・TOEIC のようにフィードバックされることがまず重要である．それに比べて，そもそも紙筆版（paper and pencil test）ではなくコンピュータを用いたテスト（computer based test: CBT）で実施されるか

どうかは，受験機会がどれだけ提供されるかに依存することから，サンプル数・項目銀行整備・CBT の実現可能性の順に検討され，しかるのちに CAT への展開を考えるべきと思われる[xvixvii].

2.5 数学テスト作成・実施システム構築への示唆

本節を概観して，「ユーザー」を念頭におけば，分析モデルとしては，1 パラメータロジスティックモデル，あるいは既往研究で適用例の多い 2 パラメータロジスティックモデルまでで検討を進めるべきことが示唆された．また，システム構築の手順として，サンプル数・項目銀行整備・CBT の実現可能性，しかるのちに CAT への展開を考えるべきことも確認した．

3. 数学関連テスト分野における IRT 利用

前節において，IRT の分析モデル概容から，「ユーザー」を念頭においた際に重要となる視点，すなわち豊かなサンプル数・項目銀行整備・CBT の実現可能性，しかるのちに CAT への展開を考えるべきことが示唆された．

そこで本節では，とりわけ以上の点に着目して，日本の数学関連テスト分野における IRT の利用状況について邦文文献を整理する．テスト分野として，数学のみならず情報関連テスト分野も収集・整理した．整理の軸として，マークシート形式を含む紙筆版テスト，CBT/CAT 版テストの順に整理する．なお，以下の既往文献はすべてテスト項目に誤答したか正答したかを応答変数（0, 1）とすることを仮定している．

3.1 紙筆版テストの事例

前出の田中ほか（2004）は，IRT のパラメータ推定に関する傾向を観察するために，福井県立大学の 2003 年度「情報科学」期末試験結果を利用している．期末試験受験者 400 名に，30 項目の試験問題について，4 肢択一形式のデータを IRT で分析した．結果として，1）上記の「直接法」の最尤推定は計算精度が高い一方でテスト項目数が多い場合に計算が困難になること，2）「EM 法」の最尤推定はパラメータ数やテスト項目数が多くとも推定できる一方で収束計算が遅く近似的計算であることに起因する誤差に留意が必要となること，3）2 パラメータロジスティックモデルから 3 パラメータロジスティックモデルに変更することで必要サンプル数が増大し計算速度が格段に遅くなる一方でテスト受験者特性を豊かに得られること，4）IRT に最尤推定を用いると能力値などパラメータの大きいあるいは小さいテスト受験者について真の値からのずれが大きくなる一方でベイズ推定を用いると過小推定になる傾向があることを確認した．ただし，本研究のテストは紙筆版と思われるものの，CBT で実施することも可能であるが，明記されていなかった．

飯島（2006）は，2004 年度に実施された埼玉県立大宮工業高校の数学科受験データを用いて，2 パラメータロジスティックモデルでの分析を試みているが，『日本数学教育学会誌』臨時増刊総会特集号 88 の記載では，本稿筆者らはサンプル数の情報を得ることができなかった．こちらも紙筆版と思われるが，明記されていなかった．

尾崎・松坂（2007）は八戸工業大学で 1996 年～2007 年の入学直後リメディアル科目のうち，数学のデータを 2 パラメータロジスティックモデルで分析した．成績処理はマークシートで行っている．サンプル数は明示されていないが，テスト項目数は 20 問で実施している．1 問 1 点として素点 0 点～20 点のグループに分け，同時最尤推定で分析するために全問正答（20 点）・全問誤答（0 点）グループを除いた 19 グループ（1 点～19 点）を用いて分析結果を精査している．全サンプルの能力値の中心を 0 として，IRT で推定されたテスト受験者能力値を素点で分けた 19 グループと合わせてクロス集計したところ，1996 年から 2007 年に向けて，テスト受験者能力値が 0 以下，すなわち全サンプルの平均能力値以下のサンプルが増えていくことが確認され，当該リメディアル数学科目受講者の数学能力が低下していることが示唆されている．ただし，テスト項目は「比例式の計算」「絶対値の計算」と，内容分野のみ表示されており，問題自体が変更されてきたのかどうか、明記されていなかった．

矢野・藤井（2007）は，1989 年～2006 年までの宮崎県数学一斉テストのデータを 2 パラメータロジスティックモデルによって分析している．まず，分析対象年の間に実施された数学一斉テストの問題の中から，「基本的な計算問題」（103 問）・「基本的な知識を問う小問集合」（84 問）・「方程式」（18 問）・「数学的な見方を見る問題」（論文では項目数不明）によって項目銀行を作成した．次に，テスト項目 20 問（「基本的な計算問題」（12 問）・「基本的な知識を問う小問集合」（6 問）・「方程式」（1 問）・「数学的な見方を見る問題」（1 問））を項目銀行から抽出し，「基本的な計算問題」を基準として 153 パターンの調査票を作成した．そして，宮崎県の中学校 3 校を選定，中学生 653 サンプルを抽出し，ランダムに調査票を配布して数学

授業中の 30 分で解答してもらい，まずテスト項目困難度を調べた．その後，各年度 1 万人以上の数学一斉テスト回答結果を分析している．各年度の能力値の平均を，年度を説明変数として単回帰分析したところ，「文字式」のテスト項目（論文では詳細不明）については若干の増加傾向，「方程式」のテスト項目（論文では詳細不明）についてはほぼ横ばいであることが観察され，宮崎県下の中学生の数学能力について，明確な学力低下傾向が観察できなかったとしている．

古谷・愛甲（2007）は，情報系大学生が受験した C 言語のデバッグおよびプログラムの出力予想テストについて 2 パラメータロジスティックモデルで能力値の推定を試み，プログラミング教育における IRT 利用の可能性を示している．推定には「EM 法」による周辺最尤推定を用いた．C 言語のデバッグテストは受験者 63 サンプルで 20 問，C 言語のプログラムの出力予想テストは受験者 62 サンプルで 8 問であったが，本論文の著者自身，後者のテストの項目数（8 問）が少なかったことを認めている．

松宮ほか（2012）では 2003，2006，2009 年度にマークシート方式で実施された京都府中学校学力診断テストの数学項目について能力値を推定し，経年的に変化のないことを確認している．推定には 3 パラメータロジスティックモデルを採用している．分析対象とした項目は 2003 年度が 24 項目，2006 年度が 25 項目，2009 年度が 25 項目であり，まずこちらで項目銀行を作成した．次に，各年度からそれぞれ 17 項目を，4 肢択一問題に絞って抽出し，年度ごとの問題セットを作成した（セット③・セット⑥・セット⑨）．この問題セットを 2 種類組み合わせることで 34 問構成の調査票を 6 パターン（③⑥・⑥③・③⑨・⑨③・⑥⑨・⑨⑥）作成，各パターンに調査協力サンプルを 70 前後割り当てている．その結果，各年度（③・⑥・⑨）でそれぞれサンプル数が 280 前後，テスト項目は 17 項目で困難度・識別力・疑似偶然水準パラメータを計測している．

その他，大橋（2010）は，高校数学テストデザインに IRT を利用する授業実践を行っている旨を報告しているが，内容，測定結果などは日本数学教育学会数学教育論文発表会での報告時点では得られていない．

3.2 CBT/CAT 版テストの事例

IRT と CBT/CAT を統合したシステムは，予備的テスト実施などで暫定的に得られた困難度や能力値等パラメータを漸進的に修正していくことから，紙筆版よりも受験者それぞれに提示するテスト項目数が少なくなると予想されている（村木 2001）．近年，その特性を活かし，数学テスト作成・実施システムを運用した事例が蓄積されつつある[xviii]．

月原ほか（2008）は，高校数学（数 I, II, III, A, B, C）について，Moodle[xix]による e-ラーニングシステムを九州工業大学情報工学部の新入生を対象として導入を試み，2 パラメータロジスティックモデルによって分析を行っている．学生の習熟度を正確に把握し，学生への能力評価レスポンスを早めることで，意欲の向上につながるようなシステム構築を模索している．サンプル数は 432，テスト項目は 35 問，困難度と能力値のパラメータ推定には「EM 法」による周辺最尤推定を用いている．分析の結果，困難度の高い問題を解いた受験者ほど IRT での能力値評価が高くなる傾向が示唆され，素点評価より IRT の能力値の方が，困難度の高い課題にチャレンジした受験者を適切に評価している可能性が示された．

九島・小玉（2011）は八戸工業大学工学部システム情報工学科の 2 年生に実施した「情報ネットワーク入門」の Moodle による e-ラーニングシステムに IRT を組み込むことで，習熟度に合わせた自習用問題を提示する実験を試みた．分析は計算の単純化のために 1 パラメータロジスティックモデル（あるいは Rasch モデル）を用いており，課題として 3 パラメータロジスティックモデルへの拡張を挙げている．項目銀行は 5 肢択一問題で 100 問，実験として抽出したサンプル数は 67，テスト項目はランダムに 50 問提示した．実験設定として，まず予備的テスト（習熟度テスト）を実施し，計算の簡便な PROX 法によって困難度と能力値を推定する[xx]．次に，約半数のサンプルには能力値に適した教材を用意，自習を促したうえで，もう一度習熟度テストを実施した．なお，システムのモジュール構築には PHP 言語を用いている．結果として，自習を促したサンプルの能力値の方が高いことが示唆された[xxi]．ただし，習熟度テストを，自習を促さなかったサンプルにもう一度行っていたかどうか，論文には明記されていなかった．

野口ほか（2014）では，九州工業大学情報工学部の推薦入試合格者を対象に，入学前教育システムとして IRT を応用した適応型のオンラインシステム「愛あるって」を構築し，試験運用している．大学生サンプル 50 人に予備的テスト（モニターテスト）を実施，項目銀行内のテスト項目数，予備的テストにて提示したテスト項目数とも明記されていない．システムはブラウザ上で動作する Web ページとして PHP 言語で作成され，MySQL

データベースシステム・能力値推定用の Java プログラムを併用した.

実際に，情報処理に関して IRT と CBT/CAT を統合した運営を行っている検定試験も存在する．独立行政法人情報処理推進機構（Information-technology Promotion Agency, Japan: IPA）[xxii]は，2012 年に日本の国家試験としてはじめて IRT に基づいた CBT テスト方式を IT パスポート試験[xxiii]に採用した．IT パスポート試験は情報処理技術者試験のひとつとして 2009 年から実施されており，2012 年に IRT/CBT 統合型のテスト方式に移行した．4 肢択一形式で，小問 84 問，1 問につき 4 つの小問が出題される中問 4 問の合計 100 問から構成される．この 100 問の出題のうち，評価は 92 問で行われ，8 問は今後の出題への問題評価のために利用される[xxiv]．実施に伴い，最初はテスト方式変更への対応もあってのことか，受験者数が減少したが，年々応募者数が上昇し，2013 年度で約 75000 人が応募した．CBT 方式への移行に伴い，これまで春と秋の 2 回しか実施されなかった試験が随時開催できるようになった．毎日試験が実施される試験場もあり，受験生の受験機会が大幅に増加したことがわかる.

3.3 数学テスト作成･実施システム構築への示唆

以上を概観すると，分析モデルとしては 2 パラメータロジスティックモデル，サンプル数としては 400 以上，一度のテスト実施にテスト項目 20 問以上が目安となろう．また，推定法は「EM 法」による周辺最尤推定を基礎とし，全問正答・全問誤答者や，全サンプル正答・全サンプル誤答のテスト項目を含める必要がある際にはベイズ推定を用いることが目安となる.

IRT と CBT/CAT を統合して独自に数学テスト作成・実施システムを構築する際は，Moodle を利用したブラウザ上で動作するシステム構築が進んでいる．また，野口ほか（2014）のように，能力値推定に Java プログラムを併用するなどの工夫が必要となる．これは，Waikato Environment for Knowledge Analysis (Weka): Class BayesianLogisticRegression （Genkin et al. 2007）が無償で用意されていることからも推奨されよう．また，IT パスポート試験の事例に鑑みるに，CBT/CAT の採否において「どれだけ受験機会を提供するか」も検討しなければならない.

4. おわりに

川口（2011）の言葉を借りるならば，日本の数学に関する学力研究においても,「良質な学力データの蓄積と，その公開」が必要である．IRT は，能力値などパラメータ推定に関する課題を克服すべく理論を発展させ，多様な分野で事例が蓄積されており，テスト項目の項目銀行を活用できる，CBT/CAT と統合が進んでいるなど，数学テスト作成・実施システム構築においても，ひいては数学教育研究においても,「良質な学力データの蓄積と，その公開」に資する有望株である.

本稿の筆者は，既往邦文文献を中心とした以上のレビューをもとに，独自の数学テスト作成・実施システム構築を模索する．その際に，以下の点に留意して検討を進める.

・受験者サンプル数の確保

まず考えるべきは，相当数の受験者サンプル確保である. 比較的高度な分析にも耐えるように, 400 以上，できれば 1000 サンプル程度集められそうか，検討が必要となる．サンプル数の確保如何によって，比較的少ないサンプルでも対応できる分析モデルを用いるなど，別途対応が必要なことも理由として挙げられよう.

・項目銀行の作成

受験者サンプル数に見通しを立てたのちに，項目銀行として，どの程度の問題数を確保していくかが課題となろう．以下に述べる CBT/CAT との統合を，高大連携・大学リメディアル教育として実践するならば，公開されている入学試験問題を主軸として, 25～50 程度は項目数を確保したい. その際，単に数値を変えただけの問題は避けるべきだが，入学試験問題であればこれは容易に解決できるものと思われる.

・CBT/CAT との統合

とりわけ，高大連携と大学リメディアル教育における CBT/CAT と IRT の統合の研究が緒についたばかりである．小・中・高については，県下一斉テストなど地理的に広範な試験が整備されており，IRT による検証の土壌が既に存在する．それは「国家として，自治体として教育をどのようにしていくか」が視野にあるからであろう．一方で高大連携については，地理的な広範さよりは各大学で受け入れたい学生像が異なることから，個別対応が中心となるため，CBT/CAT との統合によって，テストの作成・実施者の負担軽減を図り，受験生へのフィードバックもできる限り遅滞なく行いたい．もちろん，受験機会の豊かな提供ひいては受験率の向上にも資するものである.

以上を念頭に，問題作成が効率的に実施可能であり，学習者の様々な事情に対応可能な独自のIRT・CBT/CAT 統合型数学テスト作成・実施システムを実装・実践に結び付けたい.

最後に，既往文献レビューに関する課題を整理する.

まず，IRT と同様に受験者の潜在特性を分析する方法論として，ニューラルテスト理論（neural test theory: NTT）が Shojima (2007) によって提案されて以来，研究と実践が進んでいる（山川・荘島 2007，木村 2009，小泉・飯村 2010）．IRT が連続変数として受験者の能力値を推定するのに対して，NTT は能力をランクで表現するもので，大学教育においてAA，A，B，C などと段階的に評価することを前提とするならば自然な想定の分析手法である．能力推定値を連続変数として示すべき場合，離散変数として示した方が使いよい場合それぞれに対して，IRT と NTT の使い分けも可能であるため，数学テスト作成・実施システムへの導入可能性を探りたい.

さらに，ベイジアンネットワーク（Bayesian network）によって，たとえば「正負の数の四則演算の知識が前提となって方程式が解けるようになる」など，教員「ユーザー」の事前知識や，学習の過程で習熟度が増していくという，学生「ユーザー」の成長を明示的に組み込んだテスト理論も提案されてきた（植野ほか 1994，加藤・赤堀 1999 など）．植野ほか（1994）では，ベイジアンネットワークを組み込んだ場合と IRT を利用した場合とでどのような出題方略の異同があるかを考察し，IRT の提供するテスト情報関数が学習者選抜のための情報であること，学習者自身にフィードバックすることを主目的とするのであれば，成長を織り込んだベイジアンネットワークのモデルの方が有効であることを指摘している．IRT との両立可能性を中心に整理したい.

以上, NTT とベイジアンネットワークと IRT の比較検討に加え，論文に明記されていなかった部分の精査について，本稿筆者の今後の課題として挙げ，結語とする.

謝辞

本研究の一部は、情報学研究所研究助成によるものであり，『情報学研究』の編集者，2 名の匿名査読者とともに，ここに記して謝意を表する.

参考文献

(1)　阿久津洋巳，石亀雅哉（2012）「項目反応理論を用いた試験問題の検討：共通教育心理学の例」，『岩手大学教育学部附属教育実践総合センター研究紀要』11: 167-175.

(2)　飯島研一（2006）「項目反応理論による埼玉県数学科標準テスト分析」『日本数学教育学会誌』臨時増刊総会特集号 88: 478.

(3)　今井新悟，伊東祐郎，中村洋一，菊地賢一，赤木彌生，中園博美，本田明子，平村健勝（2009）「項目応答理論に基づくテストの得点―J-CAT の得点換算・解釈・利用法について―」，『大学教育』6: 93-106.

(4)　植野真臣，大西仁，繁桝算男（1994）「確率ネットワークを組み込んだテスト理論の提案」，『電子情報通信学会論文誌. A, 基礎・境界』J77-A(10): 1398-1408.

(5)　浦坂純子，西村和雄，平田純一，八木匡（2002）「数学教育と大学教育・所得・昇進―「経済学部出身者の大学教育とキャリア形成に関する実態調査」に基づく実証分析」，『日本経済研究』46: 22-43.

(6)　浦坂純子，西村和雄，平田純一，八木匡（2010）「数学教育と人的資本蓄積--日本における実証分析」『クオリティ・エデュケーション』3: 1-14.

(7)　大友賢二（2009）「項目応答理論―TOEFL・TOEIC の仕組み―」，『電子情報通信学会誌』92(12): 1008-1012.

(8)　大友賢二（1996）『項目応答理論入門』，大修館書店.

(9)　大橋真也（2010）「項目反応理論を用いた高等学校数学におけるテストデザイン」，『日本数学教育学会数学教育論文発表会論文集』43(1): 425-426.

(10)　尾崎康弘，松坂知行（2007）「項目反応理論による数学の基礎能力の推移分析」，『八戸工業大学紀要』27: 61-67.

(11)　加藤浩，赤堀侃司（1999）「ベイズ推定による適応的問題演習システムのための問題選択方式」，『電子情報通信学会論文誌. D-II, 情報・システム, II-情報処理』J82-D-II(1): 147-157.

(12)　川合治男，福山裕宣，岩瀬弘和，半田勝久（2010）「項目反応理論による新入生のコンピュータ・リテラシーの測定」，『東京成徳大学研究紀要』17: 33-47.

(13)　川口俊明（2014）「国際学力調査からみる日本の学力の変化」，『福岡教育大学紀要』63(4): 1-11.

(14)　川口俊明(2011)「日本の学力研究の現状と課題」，『日本労働研究雑誌』 53(9): 6-15.

(15)　木村哲夫（2009）「ニューラルテスト理論による英語プレイスメントテストの作成と評価」，『関東甲信越英語教育学会研究紀要』23: 23-34.

(16)　九島新，小玉成人（2011）「項目反応理論を用いた習熟度の測定と WBT への応用」，計測自動制御学会東北支部第 265 回研究集会，資料番号 265-6.

(17)　熊谷龍一（2002）「語彙理解尺度における CBT 版と紙筆版の同等性の検証―項目反応理論によるテスト作成・分析を通した検討―」，名古屋大学大学院教育発達科学研究科紀要. 心理発達科学 49, 47-54.

(18)　小泉利恵，飯村英樹（2010）「ニューラルテスト理論の特徴：古典的テスト理論・ラッシュモデリングとの比較から」，『日本言語テスト学会研究紀要』13: 91-109.

(19)　孫媛（1997）「多次元データに対する項目反応モデル」，『国立情報学研究所　学術情報センター紀要』9: 103-111.

(20)　田中武之，山川修，菊沢正裕（2004）「項目応答理論に基づく母数推定法とテストの分析」，『福井県立大学論集』24: 105-124.

(21)　月原由紀，鈴木敬一，廣瀬英雄（2008）「項目反応理論による評価を加味した数学テストと elearning シス

テムへの実装の試み」, 九州工業大学学術機関リポジトリ.
(22) 登藤直弥（2012）「大問形式の問題の項目群への項目反応に対する確率モデルの比較」,『日本テスト学会誌』8(1): 85-100.
(23) 豊田秀樹（2002）『項目反応理論 [入門編] ―テストと測定の科学―』, 朝倉書店.
(24) 長崎栄三, 國宗進, 太田伸也, 長尾篤志, 吉川, 成夫, 五十嵐一博, 牛場正則, 小俣弘子, 久保良宏, 熊倉啓之, 島崎晃, 島田功, 榛葉伸吾, 滝井章, 西村圭一, 藤森章弘, 牧野宏, 松元新一郎, 森照明（2006）「現在の学問や職業で使われている算数・数学 :「数学教育に関する研究者調査」の結果の分析」, 日本数学教育学会誌 88(3): 29-43.
(25) 長崎栄三（1994）「児童・生徒の基礎学力の形成と指導方法との関連に関する総合的研究 : 算数・数学」,『国立教育研究所紀要』123: 53-104.
(26) 塗師斌（1989）「二値データに基づく尺度の一次元性の評価の方法」,『横浜国立大学教育紀要』29: 137-148.
(27) 野口和久, 大山修平, 桑幡隆行, 作村建紀, 田上真, 廣瀬英雄（2014）「アダプティブオンライン IRT による大学数学教育への準備」,『電気学会研究会資料』FIE 2014(1-14・16・17): 45-51.
(28) 廣瀬浩二（1998）「英語歯科用語に関するテスト項目の研究(1) : 項目困難度,項目弁別力指数,モデルとデータの適合度の検討」,『明倫歯科保健技工学雑誌』1(1): 39-43.
(29) 廣瀬浩二（1999）「英語歯科用語に関するテスト項目の研究(2) : オプションの数と種類の違い」,『明倫歯科保健技工学雑誌』2(1): 51-56.
(30) 廣瀬浩二（2000）「英語歯科用語に関するテスト項目の研究(3) : 項目難易度表示の付与」,『明倫歯科保健技工学雑誌』3(1): 19-24.
(31) 廣瀬浩二（2001）「英語歯科用語に関するテスト項目の研究(4) : 項目銀行の拡張」,『明倫歯科保健技工学雑誌』4(1): 54-62.
(32) 廣瀬英雄（2009）「デジタルネイティブ度から見える九州工業大学の学生の傾向」, http://hdl.handle.net/10228/4618（retrieved on Sep. 26th 2014）.
(33) 藤森進（1998）「同時尺度調整法による垂直的等化の検討」,『人間科学研究』20: 34-47.
(34) 古谷博史, 愛甲弥生（2007）「項目反応理論を用いたプログラミングテストの分析」,『宮崎大學工學部紀要』36: 333-338.
(35) 松宮功, 永砂正弘, 荘島宏二郎（2012）「項目反応理論による学力テストの経年比較 : 京都府中学校学力診断テストの等化」『日本数学教育学会誌』94(9): 12-21.
(36) 村木英治 (2001)「コンピュータ版テスト (CBT) の実施と理論的研究」,『計測と制御』40(8): 549-554.
(37) 安田宗樹, 田中和之（2012）「項目間の相関を考慮した項目応答理論」,『電子情報通信学会技術研究報告 NC ニューロコンピューティング』111(483): 387-391.
(38) 矢野愛子, 藤井良宜（2007）「宮崎県数学一斉テストにおける中学生の学力の変化についての分析」,『日本数学教育学会数学教育論文発表会論文集』40: 61-66.
(39) 山川修, 荘島宏二郎（2007）「項目応答理論とニューラルテスト理論の比較研究」,『日本教育工学会研究報告集』2007(5): 223-226.
(40) 渡部洋, 平井洋子（1993）「段階反応モデルによる小論文データの解析」,『東京大学教育学部紀要』33: 143-150.
(41) Adams RJ, M Wilson, W Wang (1997) The Multidimensional Random Coefficients Multinomial Logit Model, Applied Psychological Measurement 21(1): 1-23.
(42) Birnbaum A (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability, In Lord FM and MR Novick (eds.), Statistical Theories of Mental Test Scores 397-479, Reading MA: Addison-Wesley Publishing.
(43) Bock RD, RJ Mislevy (1988) Adaptive EAP Estimation of Ability in a Microcomputer Environment, Applied Psychological Measurement 6: 431-444.
(44) Camilli G (1994) Origin of the Scaling Constant d = 1.7 in Item Response Theory, Journal of Educational and Behavioral Statistics 19(3): 293-295.
(45) de Ayala RJ (2009) The Theory and Practice of Item Response Theory, The Guilford Press, New York.
(46) Fox JP (2010) Bayesian Item Response Modeling: Theory and Applications, Springer New York Dordrecht Heidelberg London.
(47) Genkin A, DD Lewis, D Madigan (2007) Large-Scale Bayesian Logistic Regression for Text Categorization, Technometrics 49(3): 291-304.
(48) Grun B, Leisch F (2008) FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters, Journal of Statistical Software 28, Issue 4.
(49) Lindsay B, CC Clifford, J Grego (1991) Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis, Journal of the American Statistical Association 86(413): 96-107.
(50) Lord FM (1952) A Theory of Test Scores, Psychometric Monograph No.7, retrieved from http://www.psychometrika.org/journal/online/MN07.pdf.
(51) Rupp AA (2003) Item Response Modeling With BILOG-MG and MULTILOG for Windows, International Journal of Testing 3(4): 365-384.
(52) Shojima K (2007) Neural Test Theory, DNC Research Note 07-02, http://www.rd.dnc.ac.jp/~shojima/ntt/Shojima2007RN07-02.pdf (retrieved on Sep. 26th 2014)
(53) Wang T and MJ Kolen (2001) Equating Comparability in Computerized Testing: Issues, Criteria and Example, Journal of Educational Measurement 38: 19-49.

(2014 年 9 月 30 日受付)
(2014 年 12 月 3 日採録)

[i] 日本国内の文系・理系双方の研究者を対象に「算数・数学が使われている場面や実例」に関して調査した結果をまとめた長崎ほか（2006）は, 1）数学的知識・処理のみならず論理的思考力という算数・数学の能力・技能の価値, 2）統計学的な見方や考え方, 3）現象と数学をつなげて考える数学的モデル化が教育指導場面で必要とされていることを明らかにしている.

[ii] Camilli（1994）において検討が加えられている.

[iii] 今井ほか(2009)は IRT を利用した日本語能力の CAT を提案し, IRT での能力測定から能力値の 100 点満点換算までの実践を試みている.

[iv] 登藤（2012）は段階反応モデル・Bayesian testlet model・constant combination model を 2 パラメータロジスティックモデルとシミュレーションによって比較検討し, パラメータの推定精度に違いがないことを示唆している.

[v] 豊田（2002）では構造方程式モデリング（structural equation modeling）ないし共分散構造分析の下位概念として，カテゴリカル因子分析で能力尺度を推定する方法も提示している．

[vi] 2パラメータモデルや1パラメータモデルのテスト情報量は，Eq.7にそれぞれ$c_j = 0, a_j = a$を適宜導入すればよい．

[vii] 項目情報関数は，どの項目が他の項目よりもよくパラメータを推定しているか，という項目選択基準にも用いられる（村木2001）．

[viii] http://www.winsteps.com/index.htm

[ix] http://cran.r-project.org/web/packages/ltm/ltm.pdf

[x] http://irtanalysis.main.jp/

[xi] http://www.rd.dnc.ac.jp/~shojima/exmk/

[xii] http://weka.sourceforge.net/doc.stable/weka/classifiers/bayes/BayesianLogisticRegression.html

[xiii] http://www.mrc-bsu.cam.ac.uk/software/bugs/

[xiv] http://cran.r-project.org/web/packages/mlogit/mlogit.pdf

[xv] 尾崎・松坂（2007）はMATLAB，藤森（1998）は自作のFORTRANプログラムを用いている．IRTは計算負荷が高いと予想されるために，田中ほか(2004)をさらに拡張し，数学テスト作成・実施システム要件の十分な精査が必要と考えられる．

[xvi] 廣瀬（1998, 1999, 2000, 2001）でも，英語歯科用語に関して，テスト項目の洗練から項目銀行の整備拡張という手順に重きをおいて研究されている．

[xvii] 村木（2001）は，Wang and Kolen (2001) を引いて，いままで紙筆版で実施していたものをCBTにすることは大変であることを述べている．また一方で，熊谷（2002）は，日本語の語彙理解尺度を用いて，紙筆版（paper and pencil）とCBT版のテスト結果をIRTで検証し，同等であるという示唆を導いている．ただし，熊谷（2002）で用いているのは，コンピュータに触れる機会の多い大学生・大学院生でCBT版受験者は30サンプルであったため，より詳細な同等性の検証が必要とされる．

[xviii] 収集した研究事例はブラウザ上で動作することから，iBT (internet based test)，WBT (web based training)ともいわれる．

[xix] https://moodle.org/

[xx] テスト項目それぞれについて「全受験者のうち何人が誤答したか（誤答率）」をロジットスコアにして困難度を，受験者ごとに「全問中いくつ正解したか（正答率）」をロジットスコアにして能力値を得る手法であり，Raschモデルとして分析する際にはよく用いられる（大友1996）．

[xxi] 実験実施前に，システムへの同時ログイン数最大値を100と仮定してサーバーの負荷テストを実施している．

[xxii] http://www.ipa.go.jp/index.html

[xxiii] https://www3.jitec.ipa.go.jp/JitesCbt/index.html

[xxiv] テストに使用した項目はテストの等化（equating）を行うことで項目銀行としてプールされる．藤森(1998)は，6〜8個の共通項目をテストに含めておくことで，能力値やテスト項目の違いによる等化が可能であると示唆しており，ITパスポート試験が8問を今後の問題評価に使用することとも親和性が高い．