様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １７ 日現在

機関番号：１３９０１
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１７３１２

研究課題名（和文）項目バンク構築における等化済み項目特性の推定方法に関する研究

研究課題名（英文）A comparison of method for estimating equated item parametrs for building an item bank.

研究代表者

光永　悠彦（Mitsunaga, Haruhiko）

名古屋大学・教育発達科学研究科・准教授

研究者番号：７０７４２２９５

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：項目反応理論を応用し，テストを行うたびに項目バンクの中に困難度等の値が蓄積されていく方式のテストを行う場面について，複数存在する共通尺度構成（等化）のための手法を，実データやシミュレーションデータを用いて比較検討した。結果として，個別推定法を用いるか，もしくは同時推定法に個別推定法を組み合わせた方法をとることが理論通りの結果となることが示唆された。ただし，測定すべき構成概念について，尺度の一次元性が十分に仮定できない場合は，最適な方法を慎重に検討する必要性が指摘された。

研究成果の学術的意義や社会的意義
学校教育の現場では，たとえば学力の経年変化を題材とした研究が多く試みられているが，それらの研究において標準化された学力指標を用いて議論が行われることは少ない。そのためにはある程度の厳密性をもって学力の一次元尺度化が行われることが前提となる。本研究の成果は，毎年項目バンクの中身が増えていくという限定的なテスト計画の場合において，どのような等化方法が有効かを示すことであり，標準化された学力指標を構築し，さまざまな問題項目を用いて学力を測定していくための手段を，テスト理論の考え方に沿った形で提供することに寄与することにつながる。また等化の方法論を研究するうえでも有益な示唆を与えることが期待される。

研究成果の概要（英文）：Constructing an item bank plays an important role in administrating standardized test iteratively based on item response theory (IRT). Standardized item characteristics indices are estimated on the field test and stored to the item bank. However, there were few research about the method of equating, or obtaining common scale of item parameter and examinee ability, especially test administration with breeding contents of an item bank. In this study, several types of test equating method were compared on the simulation using artificial and real test data. Results show that concurrent calibration method with separate calibration marked much accurate, as well as separate calibration method does, but further consideration is needed when the scale of concept does not meet the assumption of unidimensionality, the fundamental assumption on equating.

研究分野：心理統計学，教育測定学

キーワード：項目反応理論　等化　多母集団IRTモデル　重複テスト分冊法　同時推定　個別推定

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

学力検査，学力調査や入試等，社会においてテストが果たす役割は大きい。しかしながら，日本で行われてきたこれらのテストは，異なる実施回をまたいで行われたテストにおいて，スコアの意味が異なるという問題点が指摘されてきた。すなわち，年度単位で行われるテストにおいて，ある年度のテストで 50 点をとった受験者 A と，別の年度で 50 点をとった受験者 B のスコアは，互いに互換可能ではない可能性が高い。なぜなら，異なる年度において出題される問題（以下「項目」と表記）は異なるものにならざるを得ず，スコアの意味を同一とするためには項目の困難度を統制しなければならないのに，出題者の経験的予測によってのみ今年度を検討しなければならないためである。

この点を解決するために，項目困難度を標準化した形で表す試みが行われてきた。標準化された困難度とは，受験者の能力分布によらない形で表現された困難度のことを指す。素朴に考えると,項目に正答する確率が低いことが,すなわち困難度が高いことを表すようにみえるが，それには受験者の能力がある程度の水準であり，極端に低いわけではないという前提がある。仮に極端に低い能力を持つ受験者集団が，中程度の困難度の項目を提示された場合でも，正答確率は低くなるであろうし,能力が中程度の分布である集団に困難度が高い項目を提示しても，正答率は同様に低くなるだろう。標準化された困難度を表現するためには，受験者の能力分布によらない形で困難度を表現することが必要であり，複数の年度をまたいで共通の意味をもつスコアを返すためには，年度ごとの受験者の違いによらない困難度をあらかじめ推定するというアプローチが必要不可欠である。

標準化された困難度を推定して，問題文とともに「項目バンク」に記録しておき，毎年行われる本試験の際には項目バンクから出題し，受験者のスコアは標準化された困難度を根拠として推定する，というテストの実施方法は，多くのテストで広く行われてきた。このような，項目バンクを用いたテストの実施においては，それに先立ち，標準化された困難度を推定する必要がある。この過程は「フィールドテスト」「予備試験」などとよばれ，本試験実施前に行う必要がある。テストの実施という観点からすると，この実施方法は部分的に重複した項目を含む複数のテスト版（問題冊子）を用いることから,「重複テスト分冊法」とよばれている。しかしながら，本試験とは別に，それぞれの本試験に出題するためのフィールドテストを毎年行い続けるためには，フィールドテストの受験者を確保し続ける体制を維持する必要があり，効率的ではなかった。

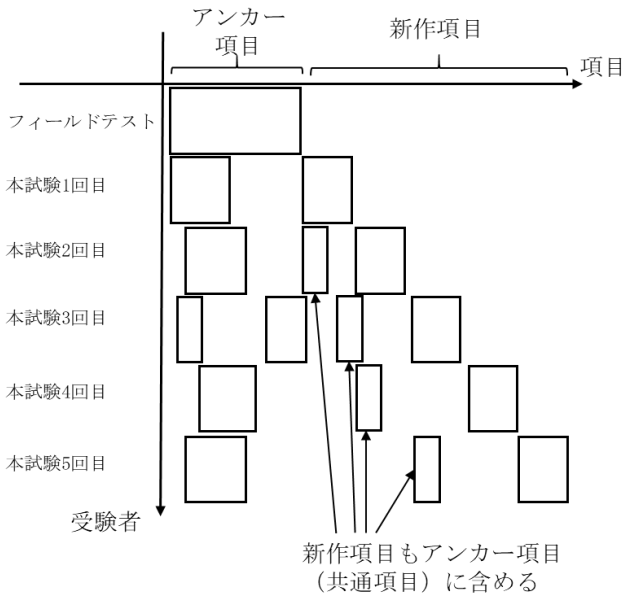

図 1：項目バンク増殖法によるテストのデザイン。規準集団を定めるためのフィールドテストののち，本試験を 5 回行う場合について示した。

また，フィールドテストの実施結果を用いて，毎年行われる本試験をまたいで共通の困難度（実際には項目識別力などの他の特性値を含む場合があるため，以下，困難度や識別力などを含めて「項目特性値」と表記する）を推定するためには，フィールドテストの間で共通の項目または受験者を用いて，規準となる集団を一つ設定し，その集団の尺度上に毎回行われるフィールドテストの尺度を変換する作業（等化 equating）が必要となる。また，等化の操作に先立ち，テストの標準化のために行われる項目反応理論（IRT）を用いた項目特性値の推定が必要となる。これらの手続きにはいくつかの前提が必要であり，等化の前提としては等化対象となる各テストで尺度が共通の一次元であること（テストで測定されている構成概念が単一のものである）などが，IRT の項目特性の推定には(尺度の一次元性を含む)IRT モデルにデータが適合していることが必要とされている。しかし，現実のテストではこれらの前提から外れたデータであることが多い。特に一次元性については，IRT 分析以外の方法で検証する必要がある。

２．研究の目的

本研究では，以上の背景を踏まえ，IRT を用いた等化を行うことを前提とした重複テスト分冊法による項目バンクの構築を題材に，より効率的な項目バンク構築方法である「項目バンク増殖法」をとりあげる（図 1 参照）。

項目バンク増殖法は，毎回のフィールドテストを，本試験の実施に組み込む方法であり，すでにいくつかのテストで行われているものである。受験者は本試験の項目と共に，フィールド

テストに相当する「新作項目」を提示され，解答することが求められる。このデザインでは，本試験受験者がフィールドテストの受験者を兼ねることとなる。毎回の本試験終了後，本試験に存在する共通項目を用いて，等化を行い，結果を項目バンクに入れていくことで，本試験が終わるたびに項目バンクのサイズが大きくなっていく。

IRT を前提とした場合の等化の手法には，いくつかの方法が提案されてきていた。その方法を大きく分けると，(a)等化の対象となるテスト版 A と B（両者には共通項目を含む）のそれぞれのデータを個別に IRT により分析して別々の受験者集団からのデータから項目特性値を求め，テスト版 A の受験者を規準集団としてテスト版 B の項目特性値をテスト版 A の尺度上で表現する，という「個別推定法（separate calibration）」，(b)テスト版 A と B の両方のデータを,同じ項目が同一列になるようにまとめ,テスト版Aとテスト版Bの受験者集団について別々の能力値分布を仮定した IRT モデルを用いて分析する「同時推定法(concurrent calibration)」，の二種が存在する。またそれぞれの種類の中でも方法が複数あり,最も改良が進んだ例として，(a)では一つの規準集団上の尺度に複数の集団上の尺度を一度の操作で等化することができる「Mayekawa の方法」が提案されている。また(b)では規準集団をテスト版 A 受験者集団と考えた場合，テスト版 B 受験者集団の項目特性値を推定する必要があるが，一旦テスト版 A だけで IRT による項目特性値を推定しておき，あらためてテスト版 A と B で同時推定を行い，テスト版 A だけで得られた項目特性値にテスト版 A と B の同時推定で得られた項目特性値を(a)で示した個別推定の手法で等化するという，個別推定を含んだ同時推定の手法が提案されている。これらの推定手法については先行研究がいくつか存在する（Arai & Mayekawa (2011)や Hanson & Béguin (2002)など）が，項目バンク増殖法によって等化を繰り返す場合に，繰り返しの効果を検証することがなされていなかった。また実データを用いたシミュレーション研究のアプローチをとっていなかった。

本研究においては，これらの等化方法を比較し，実際にテストで用いられたデータやシミュレーションによる人工データの分析結果を用いて，どの等化手法を用いれば理論通りの方法となるかを探ることを目的とした。

３．研究の方法

項目バンク増殖法に関して、(1)人工データを用いたシミュレーション，(2)実データを用いた分析、の 2 点を行った.

(1) については,項目バンク増殖法を模した人工データを用いたシミュレーションを行った。項目バンク内に存在する項目パラメタの真値を操作することで，テスト実践場面においてどのような等化手法を用いるべきかを，真値と推定値の近さを指標として検討した。また(2)については，すでに平成 29 年度において、データ提供元 2 社（株式会社エヌ・ティ・エス及び学研メディカル秀潤社（実際の契約は学研ホールディングス））と秘密保持契約を結んだ上で，平成 27・28・29 年度実施分テストについて，それぞれデータの提供を受けた。(2)については，得られたデータは完全データ（すべての項目に対して全受験者が解答している）であるが，組織的な欠測を生じさせることにより複数のテスト版からの同時推定データであるかのように模し，このデータを用いて項目バンク増殖法を行ったかのように仮定した状況において IRT による分析及び等化を行った。

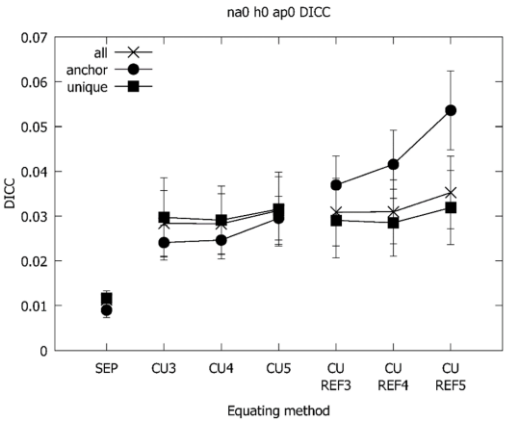

図 2：等化精度の比較。等化手法及び本試験実施回数（3～5 回）ごとに示した。

４．研究成果

人工データを用いたシミュレーションの結果,次のことがわかった：(1)個別推定を含んだ同時推定を行ったほうが,多母集団 IRT モデルのみを用いた同時推定を行った場合に比べて，真の値に近い結果となった。また個別推定の場合は，Mayekawa の方法など，項目特性曲線を近づけるアプローチをとったほうが，真値に近い結果となった（図 2 参照)。この傾向は，新作項目の数や項目特性値の真値を変化させた場合であっても大きく変わらなかったが，新作項目についてはテスト実践場面においてその数が多く設定される場合があり，この場合は真の値に近づけるために個別推定を行ったほうが同時推定よりも真値に近い結果となる傾向がみられた。

実データを用いた検討の結果，おおむね人工データを用いた結果と同様の傾向がみられたものの，測定尺度の一次元性が十分でない場合において，同時推定と個別推定のいずれの方がより良い結果となるかの一致した傾向がみられなかった。テストの場面においては，新作項目が既存の項目バンクと同一の構成概念を測定していることを，妥当性の検証の一環として行う必

要がある（この点は，等化の前提の一つである）ものの，一次元性が十分確認されていない場合においても，毎年テストを行う関係上，等化の前提を満たさないことを承知で等化を行う場合が存在する。シミュレーション研究では，一次元の尺度上で能力値が表現できるという IRT モデルが「正しい」と考え，そのモデルに従って人工データを発生させ，結果を得ていた。よって，シミュレーションの結果とは異なる傾向が得られることは研究当初から予測されたことであったが，それが一次元性の程度の違いではない他の要素，たとえば新作項目数や真値の違いといった要素によって，また欠測とするデータの数によっても左右されることがわかった以上，これらの要素を統制したうえで，あらためてどの要素が等化の精度に影響しているのかを検討する必要性が指摘された。

以上の結果より，実際のテストの実施場面において，以下のような示唆が与えられるものと考える。(1)等化の前提を満たす形の，一次元性が確保されたテストである場合は，個別推定と「個別推定を併用した同時推定」の双方が同じ程度の等化の精度を有する。この傾向は，特に新作項目を毎回の本試験で多く出題した場合に顕著となるため，一度に多くの項目を項目バンクに追加したい場合は同時推定のみで等化することは避けたほうが良い,(2)等化の前提を満たさない場合であっても,基本的には(1)で見出された結論と同様の結果であると解釈できる部分もあるため，個別推定と「個別推定を併用した同時推定」のいずれかを用いたほうが良いとされた。しかし，一次元性以外の要素によっては個別推定と「個別推定を併用した同時推定」のいずれで精度が高まるかについては，一貫した傾向がみられなかったことから，テストデータの一次元性をあらかじめフィールドテストの段階で測定しておき，その結果を用いて疑似データを用いたシミュレーションを本研究で行ったシミュレーションの手法により行うことで，当該テストの場合においていずれの手法で精度が高まるかを検討できるであろう。テストの実施主体からすれば，一貫してよい結果を返すような等化手法が一意に定まらないという結論は有益ではないかもしれないが，もともと等化が行われている前提として「一次元性」が挙げられ，それを無視する形で等化の手続きをせざるを得ない場合は,より頑健な方法をとったとしても,モデルに沿った解釈ができない場合がありうることをテスト実施主体は自覚すべきではないだろうか。

しかしながら，本研究ではテストの一次元性が，項目バンク増殖法による等化に影響する傾向が指摘された。一次元性の程度の大小と,新作項目数等の他の要因との交互作用については，今後の研究課題として残された。また，シミュレーションの手法だけではなく，実践研究の一環として等化方法を検討する事例があることが理想ではあるが，本研究ではそのような事例研修をとることはできなかった。これらの点については，研究手法のあり方を含め今後さらなる検討が必要であろう。

（参考文献）

Arai, S. & Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika,* **38**, 1-16.

Hanson, B.A. & Béguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement,* **26(1)**, 3-24.

５．主な発表論文等

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　3　件）

① Mitsunaga, H. (2018). Investigating Practicality for Building an Item Bank Using CALR Method. International Meeting of the Psychometric Society (IMPS 2018), Columbia Univ., New York, NY.

② 光永悠彦（2017）．等化手続きの違いが項目バンクのパラメタ推定値に及ぼす影響―等化係数を用いた等化法を同時推定法に併用する効果の実践的検討―．　日本教育心理学会第 59 回総会，名古屋国際会議場

③ 光永悠彦(2017).項目反応理論に基づいた項目バンク構築における同時推定手続きの実践的効果―規準集団上と比較可能な項目特性を継続的に項目バンクに追加する場合の検討―. 日本テスト学会第 15 回大会，東北大学

〔図書〕（計　1　件）

① 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