2015 年度 日本テスト学会誌　Vol.12, No.1 37 - 53

事例研究論文

わが国の TIMSS2011 数学データにおける 多次元 IRT を使った妥当性の検証について

The verification of validity in TIMSS 2011 mathematics data in Japan using multidimensional item response theory

坂本　佑太朗[1]

Yutaro Sakamoto[1]

[1]株式会社リクルートマネジメントソリューションズ

[1] Recruit Management Solutions Co., Ltd.

Correspondence concerning this article should be sent to: Yutaro Sakamoto, Recruit Management Solutions Co., Ltd., Gate City Ohsaki, West Tower 1-11-1 Ohsaki, Shinagawa-ku Tokyo 141-0032
Email：yutaro_sakamoto@recruit-ms.co.jp

わが国の TIMSS2011 数学データにおける多次元 IRT を使った妥当性の検証について

坂本　佑太朗[1]

[1]株式会社リクルートマネジメントソリューションズ

科学的な根拠に基づいた教育論議が求められている中で，テストの品質保証という観点もまた重要である．現状としても，構成概念を精緻に測定することの重要性が再認識される必要があり，先行研究自体も不足していることが指摘されている．そこで本研究では，多次元 IRT を使ってわが国の TIMSS2011 中学校 2 年生数学データにおける構成概念妥当性の「構造的な側面」について検討した．その際，TIMSS2011 の「認知的領域」である「知識」「推論」「応用」が持つ下位領域特有の影響について双因子モデルを用いることで項目情報量の観点から検証することを試みた．その結果，下位領域特有の影響が相対的に大きい項目は 214 項目中 23 項目存在し，通常の IRT 分析では拾えなかった下位領域に関する特徴を定量的・定性的に確認することができた．つまり，この結果は多次元 IRT により構成概念についての測定論的な特徴を詳細に表現出来たことを意味し，今後のさらなる応用可能性が示唆された．

キーワード：構成概念妥当性，多次元 IRT，双因子モデル

The verification of validity in TIMSS 2011 mathematics data in Japan using multidimensional item response theory

Yutaro Sakamoto[1]

[1] Recruit Management Solutions Co., Ltd.

While it is said that the evidence based discussion about education is needed, the quality assurance of test is also important. They say that we need to reaffirm the significance of measuring constructs correctly and the previous studies are not enough. The present study examined the verification of validity in TIMSS 2011 mathematics data in Japan using multidimensional IRT. In addition, the present study tried to investigate what the subscales “knowing” “reasoning” “applying” measure using bifactor model in terms of item information . As a result, there are 23 items which group factors have more impact on than general factor, so the present study proved characteristics which unidimensional IRT can not express. In other words, the present study can express characteristics about constructs which this test try to measure using multidimensional IRT and the application possibility was suggested.

Keywords：construct validity, multidimensional IRT, bifactor model

はじめに

科学的な根拠に基づいた（evidence based）教育論議の必要性が主に教育経済学の領域で言われている（中室，2015）中で，「学力」を測定するテストそれ自体の品質保証という観点もまた重要である．教育経済学的な領域では，テスト得点を従属変数，それに影響を与える要因として考えられる変数を独立変数として設定することによって「学力」の規定要因を探ることに関心がある．その際，従属変数としてのテスト得点は「学力」を示していることを前提に用いられるが，そもそもそのテスト得点が本当に測りたい構成概念（construct）の特徴を反映しているのか，またそのテストが十分な測定精度を担保しているのかという測定論的な観点がどうしても軽視されてしまう傾向がある．しかしながら，受検者の人生を左右してしまうハイステイクスなテストや，教育政策に影響するようなテストであればあるほど，その後の経済学的な分析を支えるテストそれ自体の検証が重要であると言える．

最近では，大学入学者選抜における項目反応理論（IRT, Lord, 1952）の応用可能性が議論され（中央教育審議会，2014a, 2014b），わが国でも教育測定評価への関心は少なからず高まってきていることは伺える．その中で，教科・科目の枠を超えた合教科科目型あるいは総合型の問題を従来の教科型に加えて出題を検討する（文部科学省，2015）などというように，より一層複雑な心理学的特性を測定することが求められてきている．しかし，測定論的なエビデンスを担保しながらそれらを実現するには，「日本のテスト文化」（池田，1970；Arai & Mayekawa, 2005；柴山，2008）の制約下においては簡単ではなく，これまで以上に IRT を中心としたテストの測定技術研究の蓄積が急務であるというのが現状であろう．

テストによる心理学的な測定について議論する上で，テストの信頼性（reliability）と妥当性（validity）は必須要素である．信頼性は妥当性の必要条件として位置づけられるため，言い換えれば，信頼性をどれだけ高めたテストであっても，測りたい特性を反映していない妥当性の低いテストであれば本末転倒である．テストの妥当性概念については，現在では構成概念妥当性（construct validity）を 6 つの側面から捉え直したアプローチが主流となっており（AERA，2014；Messick，1989，1995），その中でもテストの次元性に関わる「構造的な側面」については，「テストがどのような能力を測定しているのか」という問いを検証する上で重要な側面の 1 つであると判断できる．

IRT をテスト開発の基盤とする際には，1 次元性の仮定と，局所独立の仮定という 2 つの仮定を置いた IRT モデルが使用されることが多い．1 次元性を確認するためには，たとえば正答/誤答の 2 値データの場合にはテトラコリック相関係数行列を推定し，固有値の減衰状況によって判断する方法が一般的であろう．たしかに，この方法による推定値は実用上十分な精度を持つことは知られている（Parry & McArdle, 1991）．しかしながら，厳密にはテトラコリック相関係数の推定値が正定値とならず固有値が計算できない場合や，その相関係数行列を因子分析することで項目の困難度を反映する因子が抽出されてしまう可能性があること，さらに受検者がどの項目に正答/誤答したのかという情報が含まれないことが指摘されている（加藤・山田・川端，2014；柳井・繁枡・前川・市川，2001）．また，その他にも Hattie（1985）によってさまざまな統計的方法が提案されているものの，「十分な一次元性」の判断基準は厳密にいえば明確ではない（Embretson & Reise, 2000）．言い換えれば，IRT を適用する際には，項目反応データが持つ情報を IRT モデルという 1 心理学的なモデルを通してある程度限定的に表現することが前提となる．つまり，複雑な心理学的特性を測定している，あるいは測定することが意図されているテストでは，1 次元性を仮定する IRT モデルでは測定論的に重要な情報を落としてしまう危険性がある．

事実として，わが国の測定・評価研究においてはこのような構成概念の精緻かつ正確な測定の重要性について再認識が求められている（石井，2014）．したがって，将来的なテスト開発を考慮したとき，測定評価技術としての IRT を単にテストデータに適用するのではなく，そのテストが何を測定しようと意図され，たしかに測定できているかというテストの妥当性研究はより一層不断に行なわれる必要があると判断できる．

問題と目的

このように時代の要請にともないテストが測定すべき心理学的特性が多様化するにしたがって，それに対応する心理測定技術も発展してきた．特に，多様な心理学的特性を捉えるための手法としてわが国でも注目されつつあるのが多次元 IRT（multidimensional IRT）

である. 多次元 IRT は通常の IRT モデルでは 1 次元性の仮定を置いていたものを多次元へ拡張した IRT モデルである. 歴史的には, その理論的基盤は 1970 年代後半から 1980 年代前半にかけて行われてきた. しかしながら，実際のテストへの適用はそれほど進んでいるとはいえないというのが現状である．その理由として，多次元 IRT の等化（equating）に関する研究が不十分であること（Min, 2007；Simon, 2008）や，多次元 IRT のテスト情報量が行列形式となるために利用しにくいことなどが指摘されている（星野，2001）．これに加えて，テストの実用場面を想定すれば，複数の測定内容に同時に関連するような項目の作成技術が確立されていないこと，もともと 1 次元の測定内容を想定した作成した項目を多次元 IRT で分析した結果を個人のスコア算出に利用してよいのか，もし仮によいとすればどのような場合かはっきりしないことなどが挙げられる（沖・前川，2014）．

通常，学力テストにおいてはそれを構成する下位領域（subscales）が存在することが一般的である．通常の IRT 分析ではテスト全体として 1 つの構成概念を測定することが仮定されているため，下位領域別の得点にはそれほど関心がない場合が多い．それでも下位領域別の得点を求められる場合には，現実的な対処として下位領域別に平均正答率を正答数得点ベースで計算することが一般的である．しかしながら，このように下位領域を別々に分析した場合，項目レベルでの次元間の関係性に関する測定論的な情報を得ることができない．つまり，ある項目が，テスト全体が測定する構成概念を反映しているのか，また下位領域特有の構成概念を反映しているのかについては判断できないことになる．このような測定論的な観点からの学力テストにおける下位領域特有の影響に関する研究や多次元 IRT を学力テストに適用した研究はほとんど行なわれておらず，今後そのような研究の蓄積が期待されているところである（沖・前川，2014）．

その中でも， Ekmecki（2013）は PISA，Mckinley & Way（1992）は TOEFL，Zwick（1987）は NAEP に対して多次元 IRT や構造方程式モデリング（structural equation modeling, SEM）を使って，下位領域によるテストの次元性に対する影響について検討している．具体的には，Ekmecki（2013）では，PISA2003，2006，2009 の数学的リテラシーにおける「内容（content）」,「過程（processes）」,「文脈（context）」を構成するそれぞれ 3～4 つの下位領域に対して， 1 因子のモデル，下位領域ごとに因子を想定した上で因子間相関を許容した確認的因子分析モデルを仮説として設定した．その結果として，仮説モデルのあてはまりの程度や下位領域間の因子間相関が 0.8 以上あることから，テストの次元性に下位領域の影響はそれほどなく，1 次元性を満たしていると結論付けている．また, Zwick（1987）では 1983～84 年に行われた NAEP の読解力データに対して，多次元 IRT を使って 1 因子モデルと 2 因子モデルを仮説として設定し，あてはまりのよいモデルを特定することで次元性検証を行った. その結果，1 因子構造のあてはまりがよく，NAEP の読解力項目は 1 つの構成概念を測定していることを明らかにした．一方で Mckinley & Way（1992）では，1987～88 年の TOEFL データを使って，多次元 IRT を用いた 1 因子モデルから 4 因子モデルまでの仮説モデルを設定し，尤度比検定や情報量基準によるモデル間比較を行った．その結果，2 因子構造が支持され，事前に置かれている測定領域がテストの次元性に何らかの影響を持っていることも確認されている．このように下位領域がテスト全体の次元性に及ぼす影響はテストデータごとに異なり，その決着は未だに確定していないことも事実である．しかしながらその事実に加えて，いずれの研究においてもテスト全体としての次元性検証にとどまり，テスト全体が測定している構成概念と下位領域が測定している構成概念との項目単位での検証はされていない．言い換えれば，テスト全体が測定する構成概念をコントロールした上でもなお，どれだけ項目レベルで下位領域特有の特徴が残っているかどうかについては検討されてこなかった.

これを統計的に検証するためには，テスト項目全体に影響を持つ一般因子（general factor）と下位領域特有の影響を示すグループ因子（group factor）を設定した多次元 IRT の確認的なモデリングの一種となる双因子モデル（bifactor model；Gibbons & Hedeker, 1992）が有効である．その利点として，一般因子とグループ因子を反映した項目パラメータを同一モデル内で推定することができるため，テスト全体が測定する構成概念と，下位領域が測定する構成概念に関する情報量が得られることなどが指摘されている（Reise, Moore & Haviland ,2010）. これを活用した先行研究としてたとえば，Reise, Morizot & Hays（2007），Reise, et al.（2010）, Reise, Ventura, Keefe, Baade, Gold, Greem, Kern, Gately, Nuechterlein, Seidman, & Bilder（2011）などがある．これらの研究では，一貫して双

因子モデルを含む多次元 IRT モデルを医療分野のデータに適用し，項目識別力パラメータを項目レベルで参照することによって，一般因子の影響を統制した上でのグループ因子の影響を相対的に比較している．特に，Reise, et al.（2007）では CAHPS2.0（Consumer Assessment of Healthcare Providers and Systems）のデータ（N=1,000）における 16 の質問項目について，1 次元の IRT 分析，2 因子を想定した確認的な多次元 IRT 分析，それら全項目に影響を与える一般因子と下位領域となるグループ因子を想定した双因子モデルによって分析を行った．その結果，双因子モデルがもっともあてはまりがよく，確認的な多次元 IRT 分析によって推定された 2 因子間の相関が 0.74 という比較的相関の高いデータにおいても，双因子モデル内での一般因子とグループ因子それぞれにおける心理学的な情報量を規定する項目識別力パラメータから，項目が測定する構成概念について定量的なエビデンスが得られることが明らかになっている．しかしながら，このような因子間相関が比較的高い場合においても，双因子モデルでのグループ因子の項目識別力パラメータの値が一般因子のそれよりも大きい項目が 16 項目中 2 項目存在していた．これは，その項目がテスト全体の測定する構成概念を測定しているというよりもむしろ，下位領域特有の影響を相対的に強く反映していることを意味する．この事実から，テスト全体の次元性検証に加えて，双因子モデルを含む多次元 IRT を使った項目レベルでの測定論的な検証を行うことは，テストの妥当性研究において重要であると判断できる．

このように多次元 IRT を用いた研究は見られるものの，学力テストの下位領域に関する多次元 IRT を用いた項目レベルでの測定論的な検証は行なわれていないのが現状である．学力テストにおいてもより一層複雑な心理学的特性を測定することが求められてきている中で，テストの内容的な区切りとなる下位領域による次元性への影響，またそれ特有の測定論的な情報量を項目レベルで検討することは重要である．そこで本研究では，Reise, et al. (2007) による手法を用いながら，これまで試みられることのなかった学力テストの下位領域に関する定量的な情報を多次元 IRT によって得ることを目的とする．具体的には，後述する仮説モデルにしたがって項目パラメータを推定する．情報量基準によりモデル間比較を行ったあと，各項目が測定領域に対して持つ心理学的な情報量を項目情報量の観点から確認する．最後に多次元 IRT を用いた際の潜在能力尺度値への影響を考察する．

方法

3.1. 使用データ

2011 年国際数学・理科教育動向調査（TIMSS2011）における数学を受検したわが国の中学校 2 年生データ（N=4,414）を使用する．項目数は 215 である(1)が，分析過程において項目識別力パラメータが異常に大きく推定されるヘイウッドのケース（heywood case）が見られた M042114A の 1 項目のみ分析から除外し，結果として 214 項目が分析の対象となっている．また，元データでは正答/誤答に加え部分点（部分正答）を与える多値型項目が 15 項目存在するが，本研究では部分正答以下をすべて誤答として処理し，すべての項目を正答/誤答の 2 値反応データに変換している．

TIMSS2011 中学校 2 年生数学では，「認知的領域（cognitive domain）」ならびに「内容的領域（content domain）」という 2 つの領域が設定されている．認知的領域は，知識（knowledge），推論（reasoning），応用（applying）の 3 領域から構成されており，出題される問題の割合はそれぞれおよそ 35%，40%，25%となっている．知識とは数学的な事実，概念，道具，手順を基にした知識に関すること，応用とは知識や概念的理解を問題場面に応用すること，そして推論とは見慣れない場面や複雑な文脈の問題や多段階の問題を解くことであると定義されている．次に内容的領域には，数（number），代数（algebra），図形（geometry），資料と確からしさ（data and chance）の 4 領域に分かれており，それぞれ 30%，30%，20%，20%の割合でテストに含まれている．（Martin & Mullis , 2012；国立教育政策研究所編，2013）．本研究では，これらの領域のうち認知的領域に焦点を絞り，TIMSS2011 データの「構造的な側面」について多次元 IRT を用いて検証することにする．テスト項目がどの領域を測定しているのかについてはすべて web 上で公開されている（IEA, 2009）ため，確認的なモデリングを行うことでテスト項目が意図された領域を測定している下位領域を検証することになる．ただし，TIMSS2011 では PISA や NAEP でも採用されている釣り合い型不完備ブロックデザイン（balanced incomplete bock designs；BIBD）(3)に基づくマトリックスサンプリング（matrix sampling）であるため，各項目に解答する受検者数が異なる場合があることには注意が必要である．

表１　わが国の TIMSS2011 数学データにおける基本統計量

itemID	人数	通過率	P.BIS	BIS
M032166	625	0.768	0.373	0.515
M032626	625	0.374	0.210	0.268
M032673	625	0.752	0.534	0.729
M052216	632	0.862	0.482	0.755
M052231	632	0.809	0.400	0.578
M052214	632	0.525	0.383	0.481
M052302	632	0.864	0.429	0.674
M042032	636	0.805	0.344	0.495
M042059	636	0.583	0.588	0.743
M042236	636	0.871	0.417	0.665
M042226	636	0.728	0.596	0.799
M032721	625	0.560	0.275	0.347
M032757	625	0.872	0.340	0.542
M032760A	625	0.662	0.552	0.715
M032760B	625	0.507	0.598	0.749
M032760C	625	0.384	0.523	0.666
M032761	625	0.173	0.481	0.711
M032692	625	0.722	0.481	0.642
M052362	632	0.853	0.441	0.678
M052408	632	0.864	0.393	0.617
M052206	632	0.585	0.472	0.597
M052429	632	0.709	0.452	0.599
M032595	625	0.798	0.432	0.616
M052061	632	0.748	0.442	0.602
M052228	632	0.661	0.453	0.586
M052173	632	0.282	0.395	0.526
M052002	632	0.169	0.404	0.601
M052084	632	0.669	0.528	0.686
M042031	636	0.668	0.549	0.712
M042086	636	0.442	0.568	0.714
M042245	636	0.489	0.551	0.691
M042270	636	0.854	0.398	0.614
M042201	636	0.789	0.554	0.783

3.2.　分析枠組みとモデル

多次元 IRT は一般に補償型（compensatory）モデルと非補償型（noncompensatory）モデルに大別される．補償型モデルは，複数の能力を測定するテストにおいて，ある能力が低い場合でも他の能力が十分高ければ当該の項目には正答しやすいという仮定を置くモデルである．つまり，数学的にはそれぞれの次元同士は和の関係にある．一方，非補償型モデルは当該の項目に正答するにはある能力だけが高いだけでは達成されないことをモデル化しており，数学的にはそれぞれの次元同士の積によって正答確率を定義していることに特徴がある．本研究では，補償型多次元２値 2PL モデル（以下，多次元 2PL モデル）を採用し，数式で表現すれば

$$P\left(u_{ij}=1\middle|\boldsymbol{\theta_i},\boldsymbol{a_j},d_j\right)=\frac{exp\left(\boldsymbol{a_j}\boldsymbol{\theta'}_{\boldsymbol{i}}+d_j\right)}{1+exp\left(\boldsymbol{a_j}\boldsymbol{\theta'}_{\boldsymbol{i}}+d_j\right)} \quad (1)$$

と表される．このときu_{ij}は受検者iの項目jに対する反応を示し，また次元数をmとすると$\boldsymbol{a_j}$は$1\times m$の項目jの識別力パラメータベクトル，$\boldsymbol{\theta_i}$は$1\times m$の受検者iの潜在特性尺度値ベクトル，d_jは困難度に関連するパラメータ（スカラー）を示している．このとき，d_jは 1 次元性を仮定した項目困難度パラメータb_jと同じ解釈はできない．多次元 IRT の場合には多次元困難度（multidimensional difficulty, MDIFF）

$$MDIFF_j=-\frac{d_j}{\sqrt{\Sigma_{v=1}^{m}a_{jv}^2}} \quad (2)$$

によって同じ解釈が可能となる（Reckase, 2009）．ただし，補償型の多次元 IRT はカテゴリカル因子分析と同値であり（荘島, 2003；Takane & De Leeuw, 1987），探索的な場合には項目識別力パラメータの単純構造への回転が可能となる．

また，測定領域に関する事前の仮説がある場合，確認的な多次元 IRT 分析を実行することも可能である．その代表例としての双因子モデルにおける項目識別力パラメータを行列で表記すると

$$\boldsymbol{a}=\begin{bmatrix}a_{11} & a_{12} & 0\\ a_{21} & a_{22} & 0\\ a_{31} & 0 & a_{33}\\ a_{41} & 0 & a_{43}\end{bmatrix}$$

となる．これから明らかなように，双因子モデルは全項目へ影響を与える一般因子と，それぞれの下位領域からの影響としてグループ因子を定めるモデルである．

このとき，一般因子を統制することによってグループ因子による影響を検討できる点がこのモデルの特徴である．

また，多次元 IRT でも通常の IRT と同様に項目情報量を定義することができる．項目jに正答する確率と誤答する確率をそれぞれ$P_j(\theta)$，$Q_j(\theta)$とすると，通常の 2PL モデルにおける項目情報量関数は，

$$I_j(\theta) = \frac{\{a_j P_j(\theta)Q_j(\theta)\}^2}{P_j(\theta)Q_j(\theta)} = a_j^2 P_j(\theta)Q_j(\theta) \quad (3)$$

と定義される．多次元 IRT においてはこれを多次元空間に拡張し，空間上の一点である$\boldsymbol{\theta_i}$とv次元目のθとのなす角をα_{iv}とすると

$$I_j(\boldsymbol{\theta}) = \frac{\{\boldsymbol{a_j} P_j(\boldsymbol{\theta})Q_j(\boldsymbol{\theta})\}^2}{P_j(\boldsymbol{\theta})Q_j(\boldsymbol{\theta})} = \frac{(P_j(\boldsymbol{\theta})Q_j(\boldsymbol{\theta})\sum_{v=1}^{m} a_{jv}cos\alpha_{iv})^2}{P_j(\boldsymbol{\theta})Q_j(\boldsymbol{\theta})} = P_j(\boldsymbol{\theta})Q_j(\boldsymbol{\theta})(\sum_{v=1}^{m} a_{jv}cos\alpha_{iv})^2 \quad (4)$$

として定義できる（Reckase & Mckinley, 1991；Reckase, 2009）．当然，（4）式で定義される項目情報量関数も，(1) 式より項目識別力パラメータベクトル$\boldsymbol{a_j}$，困難度に関するパラメータd_j，潜在特性尺度値ベクトル$\boldsymbol{\theta_i}$の関数となっており，多次元の場合でも項目識別力パラメータの値によって項目情報量が規定されることがわかる．この性質を利用することで，特定の項目について同一モデル内での次元間の情報量を比較することができる．特に，双因子モデルの場合には，テスト項目全域に渡り測定される一般因子と下位領域が測定するグループ因子の測定論的な情報量の比較が可能となることが理論的に導ける（Reise, et al., 2007）．

そこで本研究では，Reise, et al.（2007），Rindskopf & Rose（1988）を参考にして以下の 4 つのモデルを仮説として設定した．まずモデル A として 1 次元性を仮定した通常の 2PL モデルを設定する．このとき，テスト全体は 1 つの「数学力」を測定していることを仮定することになる．次に，特定の下位領域が該当しない項目についてはすべての項目識別力パラメータを 0 とする確認的な多次元 IRT モデルをモデル B として設定する．このとき，テストの下位領域間には何らかの相関関係があると考えるのが自然であることから，下位領域間の因子間相関を認めることにする(4)．次に，テスト全体に「数学力」因子，それに加えて下位領域特有の影響を認める双因子モデルを設定するが，このときモデル B と同様に下位領域間の関係を考慮することが可能であるため，因子間相関を許容しないモデルと許容するモデルをそれぞれモデル C とモデル D とする．

なお，分析には R3.1.0 におけるパッケージ mirt1.7（Chalmers, 2015）を使用し，項目パラメータの推定には Metropolis-Hastings Robbins-Monro 法（MH-RM 法，Cai, 2010）(5)を用いた．

結果と考察

仮説モデルにおける推定された項目識別力パラメータを整理したものが表 2 である．ただし，紙面の都合上，各測定領域の分冊順に上位 10 項目の推定値のみを掲載している．

本分析における統計的な情報量基準として，赤池情報量基準（Akaike's Information Criterion, AIC；Akaike, 1974），ベイジアン情報量基準（Bayesian Information Criterion, BIC；Bozdogan, 1987），サンプルサイズ調整済み赤池情報量基準（corrected AIC, AICc；Sugiura, 1978），サンプルサイズ調整済みベイジアン情報量基準（sample size adjusted BIC, SABIC；Sclove, 1987），偏差情報量基準（Deviance Information Criterion, DIC；Spiegelhalter, Best, Carlin & van der Linde, 2002）の 5 つの指標を使用することとする．これらの情報量基準は多変量解析一般において用いられている指標であるが，IRT 分析にも利用できることがわかっている（Kang & Cohen, 2007）..

表２　仮説モデルの項目パラメータ推定値と情報量基準の値

	モデルA	モデルB				モデルC				モデルD			
	a1	知識	推論	応用	数学力	知識	推論	応用	数学力	知識	推論	応用	
M032166	0.692	0.696			0.690	0.082			0.657	0.246			
M032626	0.315	0.309			0.320	0.314			0.303	0.105			
M032673	1.451	1.257			1.259	0.176			1.249	0.196			
M052216	1.336	1.387			1.680	0.808			1.436	0.905			
M052231	0.804	0.809			0.800	0.098			0.771	0.259			
M052214	0.658	0.667			0.682	0.283			1.028	0.283			
M052302	0.984	0.989			1.006	0.254			0.965	0.528			
M042032	0.828	0.781			1.022	0.150			0.587	0.257			
M042059	1.445	1.257			1.498	0.133			1.140	0.385			
M042236	0.921	0.901			0.953	-0.197			0.932	-0.241			
M032721	0.398		0.398		0.371		0.169		0.338		0.210		
M032757	0.780		0.806		0.741		0.661		0.993		0.805		
M032760A	1.281		1.298		1.511		1.339		1.328		1.640		
M032760B	1.373		1.566		2.424		2.053		1.843		2.076		
M032760C	1.161		1.271		1.504		1.145		1.284		1.432		
M032761	1.594		1.536		1.704		-0.470		1.640		0.258		
M032692	0.935		0.874		1.011		-0.096		0.976		0.083		
M052362	0.981		0.972		0.987		0.065		0.994		0.086		
M052408	0.962		0.935		0.976		-0.353		0.969		0.216		
M052206	0.875		0.847		0.916		-0.200		0.935		-0.247		
M052429	0.795		0.774		0.899		-0.323		0.872		-0.078		
M032595	0.890			0.936	0.964			0.429	0.861			0.817	
M052061	0.855			0.854	0.865			0.423	0.844			0.130	
M052228	0.934			0.927	0.958			0.276	0.891			0.424	
M052173	0.874			0.875	0.887			0.246	0.928			0.159	
M052002	1.266			1.367	1.516			0.757	1.420			0.136	
M052084	1.146			1.114	1.146			0.157	1.086			0.373	
M042031	1.065			1.036	1.072			0.546	1.018			0.442	
M042086	1.751			1.733	1.259			0.370	1.545			0.311	
M042245	1.103			1.083	1.096			0.115	1.057			0.311	
M042270	0.788			0.757	0.796			-0.499	0.751			0.271	
M042201	1.727			1.316	1.704			0.429	1.586			0.456	

AIC	130512.314	130545.044	129941.376	129932.435
AICc	130604.605	130638.702	130160.662	130153.950
SABIC	131887.725	131930.096	132004.491	132005.191
BIC	133247.738	133299.641	134044.511	134054.744
DIC	130512.314	130545.044	129941.376	129932.435
logLik	-64828.157	-64841.522	-64328.688	-64321.217

因子間相関	知識	推論	応用
知識	1.000	0.908	0.944
推論		1.000	0.913
応用			1.000

因子間相関	知識	推論	応用
知識	1.000	0.602	0.720
推論		1.000	0.665
応用			1.000

表 2 を参照すると，AIC，AICc，DIC はモデル D，SABIC と BIC はモデル A を支持していることがわかる．したがってこの結果から，相対的にはモデル A あるいは D があてはまりの良いモデルだと判断できるが，1 つには絞りきることができない．そこで，まず「数学力」としての一般因子を想定していないモデル B を参照すると知識・推論，知識・応用，推論・応用でそれぞれ 0.908，0.944，0.913 という非常に高い因子間相関があることが確認できる．次に，一般因子による影響を統制したモデル D においては知識・推論，知識・応用，推論・応用でそれぞれ 0.602，0.720，0.665 となっており，下位領域間での相関が弱まっていることが確認できる．これらの事実のみで，今回のテストデータが「1 次元性を満たす」とは断定できないものの，テストを構成する下位領域特有の影響よりもテスト全体として 1 つの構成概念の影響を受けている可能性が高いと推察できる．

しかしながら，モデル D において推定された項目識別力パラメータ全体を参照すると，「数学力」因子よりも下位領域特有の影響を受けている項目，つまり下位領域の項目識別力パラメータの値が「数学力」のそれより大きい項目は 214 項目中 23 項目存在した．つまり，これらの項目については通常の 1 次元性を仮定する IRT 分析では表現できていなかった下位領域特有の要素を強く反映している可能性があると示唆されることになる．表 2 中に示した項目のうち，そのような項目は「推論」を測定するとされた M032760A～C の 3 項目が該当する．これらについて，まず通常の 1 次元性を仮定した場合のモデル A における項目パラメータを使って項目特性曲線（ICC）を描けば

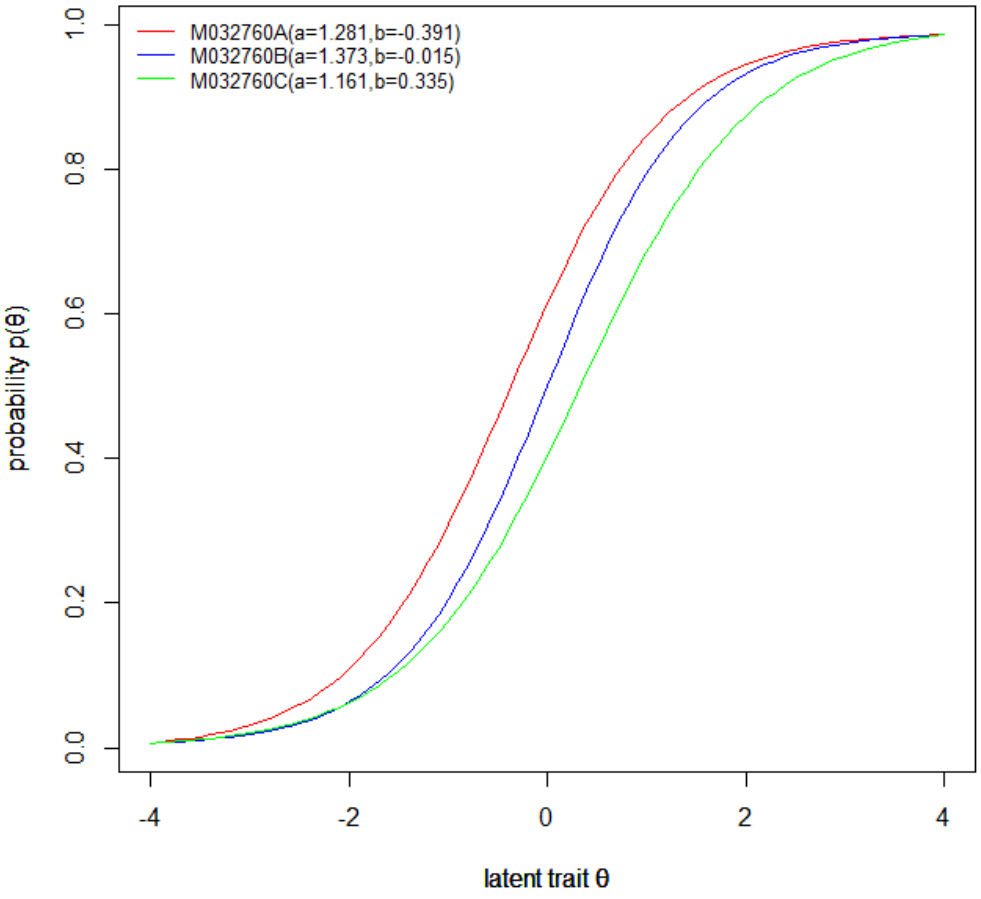
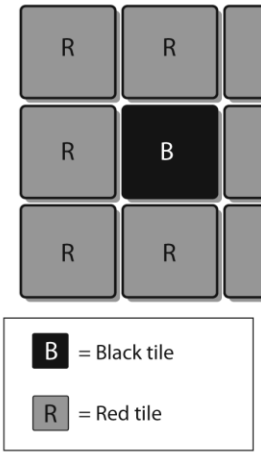


図 1　M032760A～C における ICC

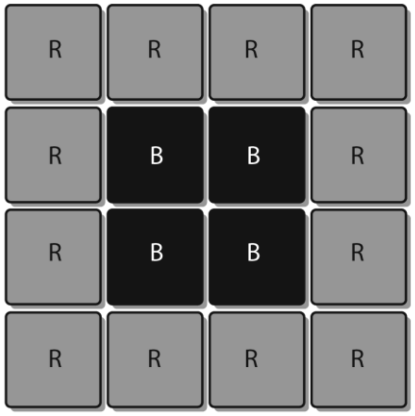
となる．表 1 で示した基本統計量（通過率，P.BIS，BIS の値はそれぞれ M032760A では 0.662，0.552，0.715，M032760B では 0.507，0.598，0.749，そして M032760C では 0.384，0.523，0.666）と図 1 の ICC を参照しても，これらの情報からテストの次元性に関して特徴的な項目であるとは判断できない．むしろ，相対的に高い項目識別力を保持していることから，測定論的には「いい項目」と判断される項目の典型例であろう．次に，項目が測定領域に対して持つ情報量は，(4) 式より項目識別力パラメータの 2 乗値に依存するという性質を理論的に導くことができる．そこで表 2 で整理した「推論」と「数学力」の項目識別力パラメータから，その 2 乗値を比較することにより，M032760A〜C において下位領域としての「推論」は「数学力」のそれぞれ 1.527 倍，1.270 倍，1.243 倍の情報量を持っていることがわかる．実際の項目内容は

Pat has red tiles and black tiles. Pat uses the tiles to make square shapes.

The 3 × 3 shape has
1 black tile and 8 red tiles.



The 4 × 4 shape has 4 black
tiles and 12 red tiles.



The table below shows the number of tiles for the first three shapes Pat made. Pat continued making shapes using this pattern. Complete the table for the 6 × 6 and 7 × 7 shapes.

Shape	Number of Black Tiles	Number of Red Tiles	Total Number of Tiles
3 × 3	1	8	9
4 × 4	4	12	16
5 × 5	9	16	25
6 × 6	16		
7 × 7	25		

Use the patterns in the previous table to answer the following questions.

A. Pat made a shape with a **total** of 64 tiles, how many were black and how many were red?

Answer: _____ black tiles _____ red tiles

B. Pat made a shape that used 49 **black** tiles.
How many **red** tiles did Pat use in that shape?

Answer: _____ red tiles

C. Next, Pat made a shape using 44 of the **red** tiles. How many black tiles would Pat need to complete the black part of the shape?

Answer: _____ black tiles

となっており（IEA, 2009），単に数学的な知識等を活用するだけではなく，図や表を読み取り，必要な計算する過程を乗り越えなければ正答にはたどり着かない項目であることがわかる．つまり，多次元 IRT を基盤とした双因子モデルを使った分析によって，基本統計量やICC では判断できない下位領域特有の影響が定量的に明らかになり，それは定性的に判断しても妥当なものであることが確認できた．

また，それぞれのモデルにおける潜在能力尺度値を推定した．本研究では，潜在能力尺度値を EAP 推定法（expected a posteriori）によって行った．その際の事前分布には標準多変量正規分布を用い，推定した項目識別力パラメータと反応データをもとに潜在特性尺度値を推定した．今回推定するモデルは情報量基準によって支持されているモデル A と D とし，その次元ごとの潜在能力尺度値の箱ひげ図が図 2，その平均値と標準偏差を表 3 に整理した．図 2 より，すべての潜在能力尺度値における中央値はそれほど変化していないことがわかる．通常の IRT 分析から推定される潜在能力尺度値と双因子モデルの潜在能力尺度値（「数学力」）は裾が少し短くなっているものの，ほぼ同じ分布することが確認できる．一方，双因子モデルにおける「数学力」を統制した後の下位領域別の潜在能力尺度値は標準偏差も小さく，その分布は狭くなっていることが確認できる．これに加えて，推定された「数学力」における潜在特性尺度値と下位領域「知識」「推論」「応用」のそれとの相関はそれぞれ 0.230，0.243，0.207 であった．この結果から，テスト全体が測定する「数学力」と下位領域間の相関は低程度であり，「数学力」を統制することによって下位領域特有の構成概念についての潜在特性尺度値が得られたと判断できる．

さらに，TIMSS は受検者の個人スコアよりも集団統計量に関心があり，テストデザインとしては BIBD を採用しているため，すべての受検者が同一の冊子に解答することはなかった．そこで，推定した潜在能力尺度値から推算値（plausible values）(6) を発生させ，その平均と標準偏差について確認した（表 4）．推算値とは，受検者における潜在能力尺度値の事後分布からの無作為標本のことである．PISA や TIMSS 等の国際的な学力調査において，その目的は個々人の潜在能力尺度値の推定よりも，集団の学力分布を性格に把握することにある．EAP 推定法によって潜在能力尺度値を推定すると，その平均については能力分布の母平均の不偏推定量であるものの，分散については母分散を過小

評価してしまうことが知られている．また，教育社会学的な二次分析においては推算値を用いることで正確な統計的な分析が可能になると期待できる．そこで本研究においても，多次元 IRT による潜在能力尺度値の推定値から推算値を発生させ，その理論的な性質得られるかを確かめる．なお，推算値は国際的な学力調査において 5 つ発生させることが通例となっている（von Davier, Gonzalez, & Mislevy, 2009；Little & Rubin, 1987）ことから，本研究でもその数を 5 とした．それら 5 つの推算値の解釈については，それらの平均と標準偏差が一致しているほど良いということではなく，あくまで EAP 推定法によって過小推定された潜在能力尺度値の標準偏差が，推算値を発生させることによって補正されるのを確認することが重要である．

そこで表 4 におけるすべてのモデルにおける推算値の標準偏差を参照すると，表 3 における潜在能力尺度値の標準偏差よりも大きくなっていることが確認できる．さらに，平均値についてはそれほど大きな差は見られなかった．いずれの結果も推算値における理論的性質を持ち合わせていることがわかる．

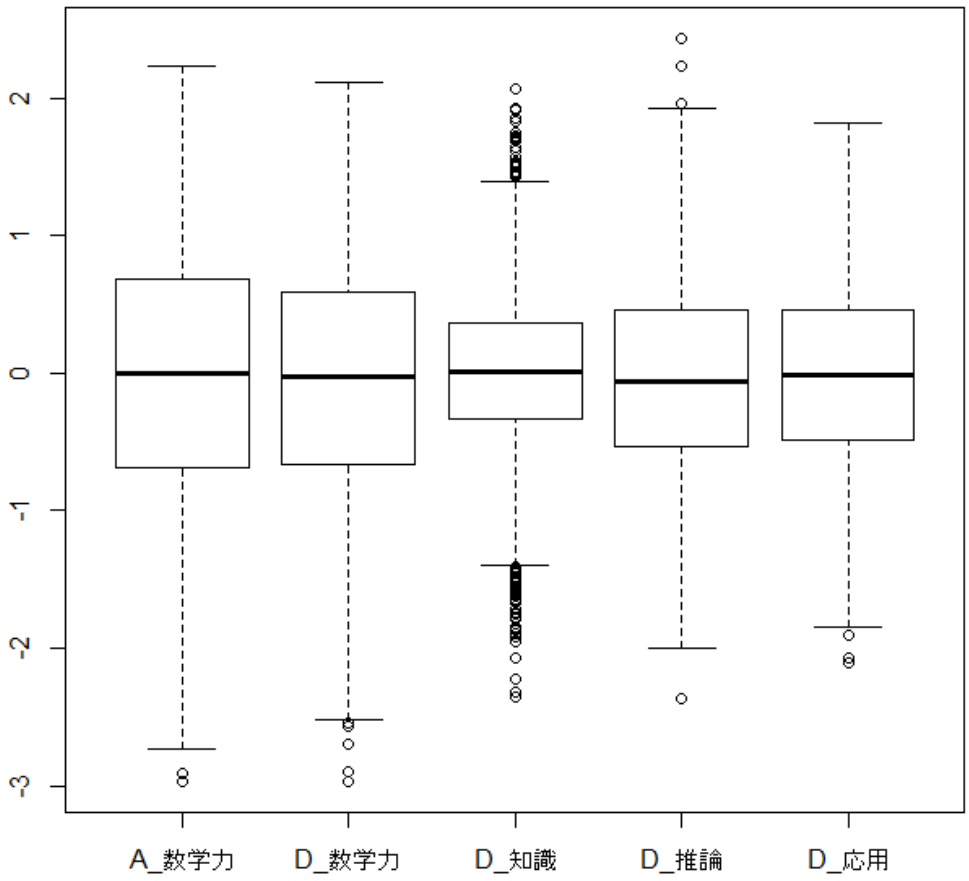


図 2　次元ごとの潜在能力尺度値の箱ひげ図

表 3　次元ごとの潜在能力尺度値の平均と標準偏差

	A_数学力	D_数学力	D_知識	D_知識	D_推論
平均	-0.017	-0.043	-0.002	-0.030	-0.025
標準偏差	0.945	0.930	0.566	0.638	0.639

表 4　次元ごとの潜在能力尺度値から発生させた推算値の平均と標準偏差

	pv_A_ 数学力 1	pv_A_ 数学力 2	pv_A_ 数学力 3	pv_A_ 数学力 4	pv_A_ 数学力 5
平均	-0.017	-0.022	-0.026	-0.013	-0.020
標準偏差	1.001	1.000	0.995	1.005	1.004

	pv_D 数学力 1	pv_D_ 知識 1	pv_D_ 推論 1	pv_D_ 応用 1	pv_D 数学力 2	pv_D_ 知識 2	pv_D_ 推論 2	pv_D_ 応用 2
平均	-0.033	-0.020	-0.018	-0.029	-0.043	-0.005	0.000	-0.008
標準偏差	1.057	1.016	1.048	1.028	1.052	1.040	1.038	1.042

pv_D 数学力3	pv_D_ 知識 3	pv_D_ 推論 3	pv_D_ 応用 3	pv_D 数学力4	pv_D_ 知識 4	pv_D_ 推論 4	pv_D_ 応用 4	pv_D 数学力5	pv_D_ 知識 5	pv_D_ 推論 5	pv_D_ 応用 5
-0.038	0.011	0.014	0.000	-0.040	-0.021	-0.015	-0.037	-0.033	-0.010	-0.018	-0.026
1.053	1.024	1.042	1.036	1.062	1.031	1.053	1.036	1.053	1.032	1.034	1.033

本研究の成果と今後の課題

本研究では TIMSS2011 のわが国における中学校 2 年生数学データを対象に，テストの妥当性研究の一環として，これまで試みられることのなかった学力テストの下位領域に関する定量的な情報を多次元 IRT によって得ることを目的とした．その結果，推定された項目パラメータから，テスト全体が測定している「数学力」よりも下位領域特有の影響が大きい項目が 214 項目中 23 項目存在することが確認できた．さらに，その中から M032760A～C に着目し，1 次元性を仮定する IRT 分析では拾うことのできなかった下位領域特有の情報を仮説モデルで推定された項目識別力パラメータの値と実際の項目内容から定量的・定性的に検証することができた．これらの結果から，多次元 IRT によって通常の IRT 分析では確認できなかったテストの測定内容に関する詳細な情報を得ることができたことは特筆に値する．

このように，テストが測定する一般因子よりも下位領域特有のグループ因子の情報量が大きいような項目に対して，実際の対処としては，定量的には本研究で示したように項目の平均正答率や点双列相関係数，双列相関係数等の基本統計量を改めて参照する必要があろう．また，通常の一次元性を仮定する IRT 分析によって ICC を描き，極端に大きい/低い識別力あるいは困難度を保有する項目ではないかを確認する必要がある．その上で，教科の専門家による項目内容の定性的な判断を通し，当該の項目をテストから除く，あるいは分析から除外する等の対処が考えられる．そのテストが現場での教育実践で用いられる用途である場合には，教師による受検者へのフィードバックは項目内容に注意しながら慎重に行う必要がある．

また，仮説モデルにしたがって潜在能力尺度値を推定し，その分布の中央値にそれほど相違はないものの，分布の広がりには違いが見られた．特に，通常の IRT 分析で推定される潜在能力尺度値と，双因子モデルにおける一般因子の潜在能力尺度値の分布は似通ったものであったが，モデル D において一般因子「数学力」を統制したとき，グループ因子のうち「知識」のみ相

対的に分布の広がりが極端に狭くなっていることは注目に値する．この現象は，今回のテスト全体が測定する「数学力」は知識的な要素を強く反映しており，それを統制することによって起きているものであると推察できる．さらに，推算値を発生させ，各モデルにおける潜在能力尺度値の標準偏差より，推算値のそれのほうが大きくなり，平均値はそれほど変わらないという推算値の理論的な性質を担保していることが確認できた．この結果から，多次元 IRT を通して分析された潜在特性尺度値を元に，教育社会学/経済学的な二次分析での活用が可能であることも明らかになった．

ハイステイクス場面で運用されるテストや，全国規模の学力調査等の結果が広範囲の教育現場へフィードバックされるようなテストであればあるほど，「テストが何を測っているのか」という妥当性に関する議論は慎重であるべきである．しかしながら，わが国における教育を巡る議論は十分な科学的根拠を基盤としたものとは言えない．とりわけ，テストに関する議論は，テストを通して算出されるテスト得点が前年に比べて「上がったか下がったか」という短略的なものが中心であり，テストの測定論的な品質についてはそれほど関心が高くない領域であった．そのような状況のなかで，これまでわが国では試みられることが少なかった多次元 IRT を用いて，テストの「構造的な側面」からの妥当性研究の一環として，テストの下位領域に関する定量的な情報を明らかにできたことは，本研究で初めて得られた知見である．

しかしながら本研究にも課題が残る．TIMSS2011 は本来多値型データであるが，理論的な配慮として多次元の 2PL モデルを用いるためにデータを 2 値反応へと加工した．今後は，多値型の多次元 IRT モデルである多次元段階反応モデル（Muraki & Carlson, 1995）や多次元一般化部分採点モデル（Yao & Schwarz, 2006）などを適用することが必要である．さらに，多次元 IRT を基盤にした潜在特性尺度値から推算値まで発生させることが可能であることが明らかになり， これらを従属変数にした教育社会学/経済学的な分析も今後期待される．

付記・謝辞

なお，本研究の一部は日本教育心理学会第 57 回総会（2015 年）にて報告されました．

また，査読の先生方からは大変有意義な御指摘を賜りましたことを心から感謝申し上げます．

注釈

1)　中には，導入設問としてA，Bを回答し，その結果を用いてZを回答させる項目がM042300Z，M042229Zの2項目のみ存在した．これらに関してはZのみを分析の対象としていることに注意されたい．

2)　ただし，紙面の都合上，掲載する項目数は各下位領域につき10項目とする．itemIDがM032116からM042226まで「知識」，M032721からM052429までが「推論」，M032595からM042201までが「応用」を測定する項目である．なお，表2おいても同様である．

3)　従来のようなすべての受検者が同一の項目へ一斉に解答する方式では，テストが測定したい領域をカバーするには多くの項目が必要となり，現実的ではなかった．BIBDでは複数の冊子（booklet）を隔たりなく実施することでその限界を超えるテストデザインである．日本語による詳しい解説は柴山・熊谷・佐藤・足立・志水（2013）を参照されたい．

4)　実際に因子間相関を認めないモデルについて項目パラメータを推定した．その結果AIC, AICc, SABIC, BIC，DICがそれぞれ130545.044，130638.702，131930.096，133299.641，137428.385であり，仮説モデルと比較しても相対的にあてはまりの悪い結果となっていた．

5)　最近ではIRTにおける項目パラメータの推定には周辺最尤推定法（marginal maximum likelihood estimation）（Bock & Aitkin, 1981）を使用するのが一般的である．しかしながら，多次元IRT分析ではその次元数が高くなるにしたがって，周辺最尤推定法では収束が遅くなってしまう等の課題が指摘されてきた（Cai, 2010）．MH-RM法では項目数が多く，さらに本分析で仮説として設定したモデルのようにテストの次元数が高い場合に有効な推定法であるため採用した．詳しくはCai（2010）を参照されたい．

6)　推算値の数学的な説明のため，受検者の反応パターンを$\boldsymbol{x}$，潜在特性尺度値をθ，尤度関数を$f(\boldsymbol{x}|\theta)$とし，潜在特性尺度値をベイズ推定するため，事前分布として$g(\theta)\sim N(\mu,\sigma^2)$を仮定する．このとき事後分布$h(\boldsymbol{x}|\theta)$は

$$h(\theta|\boldsymbol{x})=\frac{f(\boldsymbol{x}|\theta)g(\theta)}{\int f(\boldsymbol{x}|\theta\,)g(\theta)}$$

と表される．これからの無作為標本が項目反応パター

ン$\boldsymbol{x}$をもつ受検者の推算値となる. 推算値の理論的な詳細はWu（2004），日本語による説明は柴山他（2013）を参照されたい.

参考文献

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Contro*, **19**, 716-723.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC：American Psychological Association.

Arai, S., & Mayekawa, S. (2005). The characteristics of large-scale examinations administered by public institutions in Japan -from the viewpoint of standardization-. *Japanese Journal for Research on Testing*, **1**, 82- 92.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**, 443-459.

Bock, R. D., Gibbons, R. D., & Muraki. E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, **12**, 261-280.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC)：The general theory and its analytical extensions. *Psychometrika*, **52**, 345-370.

Cai, L. (2010). Metropolis-Hasting Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, **35**, 307-335.

Chalmers, P. (2015). Package 'mirt'. (https://cran.r-project.org/web/packages/mirt/mirt.pdf)（2015 年 7 月 26 日）

中央教育審議会（2014a）．中央教育審議会第 15 回高大接続特別部会　議事録.

中央教育審議会（2014b）．新しい時代にふさわしい高大接続の実現に向けた高等学校教育，大学教育，大学入学者選抜の一体的改革について（答申）．

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologist*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gibbons, R.D. & Hedeker, D.R. (1992). Full-information item bifactor analysis, *Psychometrika*, **57**, 423-436.

Hattie, J. (1985). Methodology review: Assessing undimensionality of tests and items. *Applied Psychological Measurement*, **5**, 139-164.

星野崇宏（2001）．多次元項目反応理論での相関のある特性値の線形結合に関するテスト情報関数，教育心理学研究，**49**, 491-499.

IEA（2009）. TIMSS 2007 Assessment. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.（http://timssandpirls.bc.edu/timss2011/international-released-items.html　平成 28 年 2 月 27 日アクセス）

池田央（1970）．テストの科学　井上健治（編）　テストの話　中公新書.

石井秀宗（2014）．本邦における測定・評価研究の動向―構成概念を精緻に測定することの重要性の再認識を目指して―　教育心理学年報, **53**, 70-82.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, **31**, 331-358.

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.

国立教育政策研究所編（2013）．算数・数学教育の国際比較　明石書店.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: J. Wiley & Sons.

Lord, F.M. (1952). A theory of test scores. *Psychometrics Monograph*, No.7.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McKinley, R. L., & Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models. *TOEFL technical report TR-5*, Princeton, NJ; Educational Testing Service, Februry.

Messick, S. (1989). Validity. In R. L. Linn (Ed.)

Educational Measurement (3rd) (pp.12-104) Washington, DC : American Council on Education & Macmillan. （メシック，S. 池田央・柳井晴夫・藤田恵璽・繁枡算男(監訳)(1992). 教育測定学（上巻）(pp.19-145)　みくに出版）.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, **50**, 741-749.

Min, K. (2007). Evaluation of linking methods for multidimensional IRT calibrations. *Asia Pacific Review*, **8**, 41-55.

文部科学省 (2015). 高大連携システム開発会議　中間まとめ.

Muraki, E., & Carlson, E. B. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, **19**, 73-90.

中室牧子（2015).「学力」の経済学　ディスカバートゥエンティーワン.

沖嘉訓・前川眞一 (2014). 多次元項目反応モデルによるテストデータの分析　日本テスト学会第 12 回大会発表抄録集（於：帝京大学).

Parry, C. D., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, **15**, 35-46.

Reckase, M.D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reckase, M.D., & Mckinley, R.L. (1991). The discriminating power of item that measure more than one dimension. *Applied Psychological Measurement*, **15**, 361-373.

Reise, S.P., Moore,T.M., & Haviland, M.K. (2010). Bifactor Models and Rotation: Exploring the extent to which multidimensional data yield univocal scale scores. *J Per Assess*, **92**, 544-559.

Reise, S.P., Morizot, J. & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures, *Quality of Life Research*, **16**, 19-31.

Reise, S.P., Ventura, J., Keefe, R,S.E., Baade, L.E., Gold, J.M., Greem, M.F., Kern, R.S., Gately, R.M., Nuechterlein, K.H., Seidman, L.J, & Bilder, R. (2011). Bifactor and Item Response Theory Analyses of Interviewer report Scales of Cognitive Impairment in Schizophrenia, *Psychological Assessment*, **23**, 245-261.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, **23**, 51-67.

Sclove, S. L. (1987). Application of Model-Selection Criteria to Some Problem in Multivariate Analysis. *Psychometrika*, **52**, 333-343.

柴山直 (2008). 日本のテスト文化について　人事試験研究, **208**, 2-13.

柴山直・熊谷龍一・佐藤喜一・足立幸子・志水宏吉(2013). 全国規模の学力調査におけるマトリックス・サンプリングに基づく集団統計量の推定について 平成 24 年度文部科学省委託研究「学力調査を活用した専門的課題分析に関する調査研究」研究成果報告書.

Simon, M. K. (2008). Comparison of concurrent and separate multidimensional IRT linking of item parameters. PhD thesis, University of Minnesota.

Spiegelhalter, D. J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: SeriesB*, **64**, 583.639

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics–Theory and Methods*, **7**, 13-26.

荘島宏二郎 (2003). 複数の項目反応モデルの母数の同時推定　豊田秀樹 (編) 共分散構造分析 (技術編) 朝倉書店.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, **52**, 393-408.

von Davier, M. , Gonzalez, E., & Mislevy, R.J (2009). What are plausible values and why are they useful? IERI Monograph Series, **2**, 9-36.

Wu, M. (2004). Plausible Values. *Rasch Measurement Transactions*, **18**, 976-978.

柳井晴夫・繁枡算男・前川眞一・市川雅教 (2001). 因子分析―その理論と方法―　朝倉書店.

Yao, L., & Schwarz, R. D. (2006). A multidimensional

partial credit model with associated item and test statistics: An application to mixed-format test. *Applied Psychological Measurement*, **30**, 469-492.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, **24**, 293-308.