秋田県立大学総合科学研究彙報　第18号（2017）：１－８

PISA 2012における特異項目機能の分析

渡部　　諭[1]・澁谷　泰秀[2]・鈴木　康弘[2]

わが国おける教科に関するジェンダー差、特に理系の教科に関する男女差についての研究は多い。ところが、わが国においてテスト項目の作成における特異項目機能（differential item functioning、以下 DIF）に対する認識は高いとは言えず、そのためにジェンダー差を論じる場面でも DIF に対する考慮がなかったり、ジェンダー差と DIF とを混同している研究が見られる。たとえば、北條（2013）は国際数学・理科教育動向調査（TIMSS）および OECD 生徒の学習到達度調査（PISA）の日米のデータの分析を行い数学の学力分布の差と数学学習態度の相違について検討しているが、DIF については全く言及がない。また、伊佐・知念（2014）は小学校 3 年生から 6 年生までの 4 年分のデータ及び中学校 3 年間のデータについて国語と算数・数学の学力の男女別の分析を行ったものであるが、ここでも DIF について検討した形跡は全く見られない。

DIF とは、テストの受験群の相違によってあるテスト項目が有利または不利に働く結果、群間の本来の相違とテスト項目による効果とが混同される現象をさす。もう少し厳密に定義するならば、DIF とは群の能力をそろえた上で見られる両群における統計的な特徴の相違を指さす（Angoff, 1993）。DIF の判定に用いられる統計技法は多種あるが、項目反応理論（item response theory、以下 IRT）を用いた DIF の判定が最近多くみられる。De Beer（2004）では、IRT を利用した DIF 分析を行う際には DIF の大きさが focal group と reference group の 2 本の項目特性曲線（item characteristic curve、以下 ICC）の間の面積によって表されるとしている。また Zieky（1993）によれば、ETS（Educational Testing Service）では DIF の大きさが数値によってカテゴリー化され、それに基づいて DIF が生じた項目の除外や修正を行っている。

ところで、既述したように、群間に存在する本来の相違と DIF の結果生じる相違とは厳密に区別すべきであるが、先行研究の中にはこの両者を混同していると思われるものが散見される。あるいは、DIF の可能性があるにもかかわらずその対応をせずに群間差を求めて、それをそのまま受験者群間の差に帰している研究も見られる。たとえば、Adedoyin（2010）では2004年に実施されたボツワナ中学校卒業資格試験（The Botswana Junior Certificate test）の受験生から抽出された4000名の回答に対して、3 パラメタ・ロジスティックモデルの ICC の比較による DIF 分析を行った結果、5 項目において当て推量パラメータが男子生徒と女子生徒間で有意に異なることが明らかにされた。しかし、IRT を用いた DIF 分析においては、ICC の形態やパラメータの値の相違は DIF が存在する証拠であり、それをそのまま受験者群のジェンダー差に帰することは妥当な方法ではなく、これは DIF とジェンダー差とを混同していることになる（de Beer, 2004）。

本研究の目的は、PISA の科学的リテラシー項目におけるジェンダー差の分析を行う前提として、科学的リテラシー項目の DIF 分析を行うことである。その際に、後述するように IRT を用いた DIF 分析を用いる。

[1]総合科学教育研究センター
[2]青森大学　社会学部

方　法

分析データ

OECD 生徒の学習到達度調査（PISA）2012年調査（以下、PISA2012）データは、R のパッケージ intsvy（Caro and Biecek, 2016）に含まれるデータパッケージ PISA2012lite を用いた。このうち、科学的リテラシー項目53項目のデータはこのパッケージの中の scoredItem2012に含まれる。ところが、このデータには生徒の性別データは含まれておらず、性別データは生徒の属性データである student2012に含まれる。そこで、student2012と scoredItem2012に共通に含まれる生徒の ID を表す STIDSTD をこの両者で確認したところ、完全に1対1対応をしていることがわった。したがって、student2012から抽出した国籍が日本であるデータに含まれる性別データと、scoredItem2012から抽出した国籍が日本であるデータに含まれる科学的リテラシー項目53項目を合併し 1 個のデータを作成し以後の分析に用いた。以後の分析はすべて R version 3.3.1（R Core Team, 2016）および RStudio version 0.99.902（RStudio Team, 2015）を用いて行われた。

分析方法

後に実施する Raju の方法（Raju, 1988）のために、まず科学的リテラシー項目をすべて 2 値によって表す。PISA2012における回答は 4 択であるが、このうち"Score 1"を 1 に、"Score 0"及び"N/A"、"Not reached"をすべて 0 に変換する。

続いて DIF 分析を行う。Özdemir（2015）は TIMSS 2011データに対して IRT に基づいた DIF 分析の 3 方法（Lord のカイ二乗検定（Lord, 1980)、Raju の方法、尤度比検定（Thisses, Steinberg and Wainer, 1988)）を用いて2PL モデルを作成し、item purification がある場合とない場合についてそれぞれ分析を行った結果、3 方法の中で他の方法と最も大きな相違を示した方法は Lord のカイ二乗検定であること、3 方法において item purification がある場合には DIF 項目が増えること、尤度比検定には item purification の影響はないことを明らかにした。また、item purification の有無に関わらず 3 方法に共通の DIF 項目は 2 項目であった。あくまでも分析に用いたデータとの関係で見なければいけないことであるが、Özdemir（2015）によれば Lord のカイ二乗検定は他の方法による分析結果との相違が大きいことが明らかになったために本研究の分析では用いないことにする。また、尤度比検定については、現在のところ Rasch モデルのみが利用可能であるため uniform DIF 項目の発見のみが可能である（Magis et al., 2015, p.53)。そのため、尤度比検定を uniform DIF 項目と nonuniform DIF 項目の両方の発見が可能である Raju の方法と比較することは適切ではないと思われる。したがって、本研究では、IRT に基づいた DIF 分析として Raju の方法のみを採用し、item purification がある場合と item purification がない場合についてそれぞれ2PL モデルを作成し分析を行うことにする。IRT のモデル作成には R のパッケージ irtoys を、また DIF の分析には R のパッケージ difR を用いる。

次に DIF の存在が確認された項目について男性群と女性群に分け、四分相関係数を用いてそれぞれの群の因子数を決定し因子分析を行う。もし項目群が一次元であることが確認される場合には、これらの項目群全体の2PL モデルを作成するが、因子分析の結果 2 因子以上の存在が確認される場合には、Yen（1993）にしたがい因子毎に項目をまとめた後に、それぞれの因子に属す項目群について2PL モデルを作成する。因子分析は R のパッケージ psych を用いる。

各因子に属する項目群について作成された2PL モデルについては、最初にモデルの適合度の検討を行う。モデルの適合度は個人適合度、項目適合度を求め、その他にモデルによる予測値と観測値との比較を行う。続いて、項目情報関数、テスト情報関数、テスト特性関数により2PL モデルの測定の精度を検討する。

最後に項目パラメータ及び能力パラメータの推定値について検討を行う。項目パラメータについて

は、項目特性曲線によるグラフの形状の検討及び項目識別力パラメータと項目困難度パラメータの散布図によるパラメータ間の散布状況を見る。能力パラメータについては、男性群と女性群における能力パラメータの分布の比較を行い、両群における相違について検討する（加藤・山田・川端, 2014）。

結　果

Raju の方法による DIF 分析の結果、item purification がない場合には14個（全体の26.4%）、item purification がある場合には22個（全体の41.5%）の DIF 項目が得られた。item purification がある場合の結果を Fig 1 に示す。なおこの図中、項目番号は全53項目の中での番号を表し、PISA2012で用いられた項目番号とは異なる。この結果は、item purification がある場合には DIF を示す項目が増えることを指摘したÖzdemir（2015）と一致する。さらに、item purification がない場合の14個の DIF 項目はすべて item purification がある場合の22個の DIF 項目に含まれていた。したがって、以後の分析には item purification がある場合に得られた22個の項目を用いることにする。

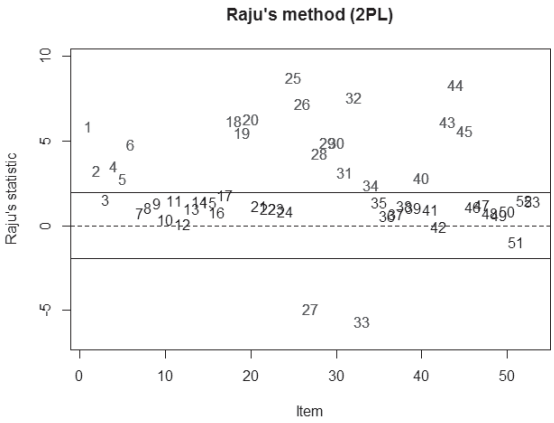

Fig. 1　Raju の DIF 分析の結果（item purification がある場合）

次にこれらの22個の DIF 項目について男性回答者群と女性回答者群に分けた。その結果、男性群は3,330名、女性群は3,021名であった。以後、性別による DIF が観察された22項目について、男女群それぞれに関してどのような相違が見られるかを IRT 分析によって検討する。

最初に四分相関係数を用いて男性群と女性群それぞれの群の因子数を決定した。その結果、スクリーテストと並行分析共に両群においてそれぞれ１因子が確認された。両群のスクリーテストと並行分析の結果を Fig 2 〜 5 に示す。

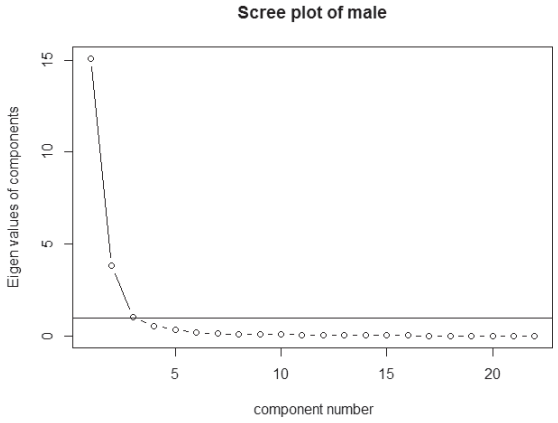

Fig. 2　スクリーテスト結果（男性群）

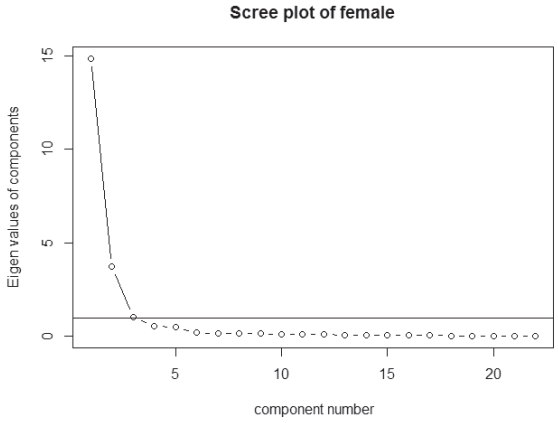

Fig. 3　スクリーテスト結果（女性群）

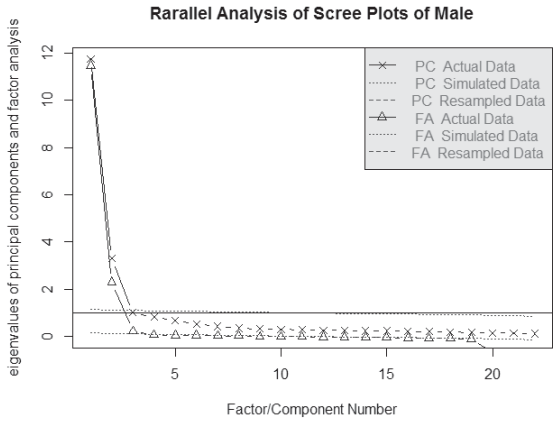

Fig. 4　並行分析結果（男性群）

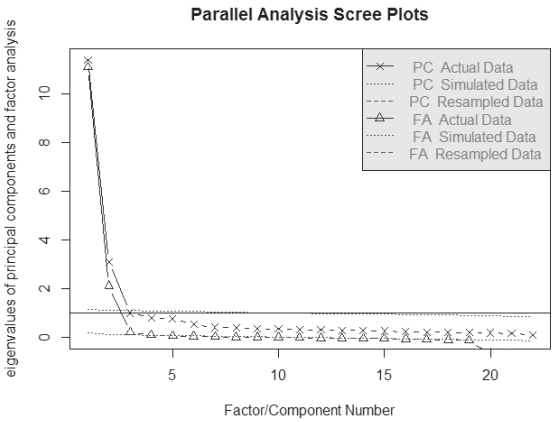

Fig.5　並行分析結果（女性群）

両群共に 1 次元性が確認されたことにより、両群において22項目を用いて2PL モデルを作成する。得られた2PL モデルについて、最初に個人適合度と項目適合度によるモデルの適合度の検討を行う。

個人適合度は irtoys で採用されている z_3 統計量（Drasgow, Levine and Williams, 1985）を用いた。分析の結果、男性群の場合 z_3 統計量は平均が0.27で標準偏差が1.10であった。また z_3 統計量の絶対値が平均値から標準偏差の 1 倍より大きな者は270名で全体の8.1%、2 倍より大きな者は185名で全体の5.6%であった。一方、女性群の場合 z_3 統計量は平均が0.21で標準偏差が1.35であった。また z_3 統計量の絶対値が平均値から標準偏差の 1 倍より大きな者は196名で全体の6.5%、2 倍より大きな者は130名で全体の4.3%であった。以上より、z_3 統計量より極端な反応をする回答者は少なく個人適合度は良好であると結論づけられる。両群の z_3 統計量のヒストグラムを Fig 6 及び 7 に示す。両群の z_3 の頻度分布は極めて類似していることがわかる。

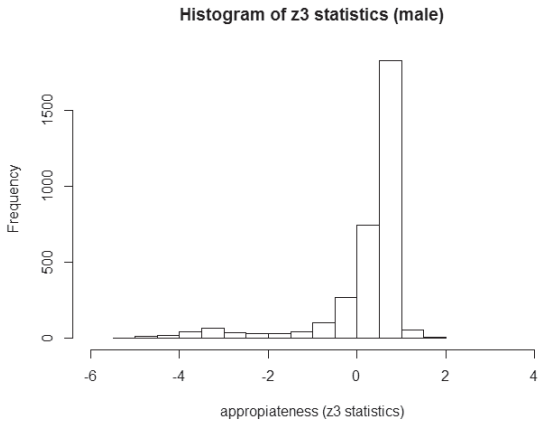

Fig. 6　z_3 統計量のヒストグラム（男性群）

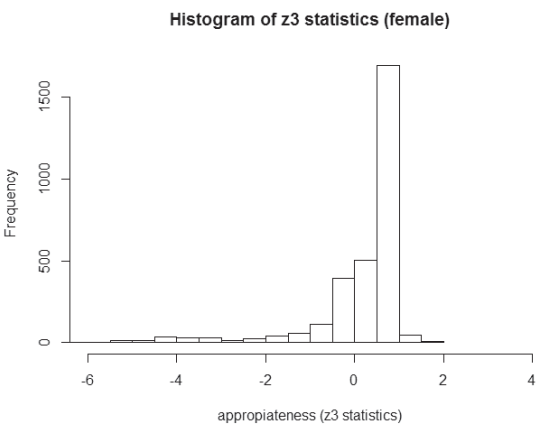

Fig. 7　z_3 統計量のヒストグラム（女性群）

項目適合度については、irtoys の関数 itf を用いて理論正答率である項目特性曲線と予め能力値に基づいて分けられた回答者群の正答率との比較を行う方法、関数 scp を用いてテスト特性関数の 1 標準誤差を用いた信頼区間に各群の回答者のそれぞれの回答がどの程度含まれるかを見る方法、項目情報関数及びテスト情報関数を用いる方法の 3 方法による検討を行った。

関数 itf を用いた方法においては、それぞれの回答群において全22項目について項目特性曲線と予め分けられた回答者群の回答が同一のグラフに示されるが、男性群においては能力値の－ 2 から 0 の範囲における適合度は良いが 0 から 2 の範囲において適合度が悪い項目が17項目、逆に能力値の－ 2 から 0 の範囲における適合度は悪いが 0 から 2 の範囲において適合度が良い項目が 4 項目、－ 2 から

2 の範囲で適合度が良かった項目が 1 項目見られた。一方、女性群においては能力値の－ 2 から 0 の範囲における適合度は良いが 0 から 2 の範囲において適合度が悪い項目が17項目、－ 2 から 2 の範囲で適合度が良かった項目が 5 項目見られた。念のために1PL モデルを作成し同様の分析を行っても、これらの傾向は改善が見られなかったのでこのまま続行することにした。

次に関数 scp を用いて得られた結果を Fig 8 及び 9 に示す。これらの図は、両群におけるテスト特性曲線とその 1 標準誤差を用いた信頼区間に各回答者の回答をプロットしたものである。これらの図より、能力値が負の領域においてテスト特性曲線が減少関数であること、各回答者の回答を見ると両群において能力値が 0 から 1 の領域で信頼区間からはずれる回答が多いこと、男性群においては能力値が－ 1 から 0 の領域で、また女性群においては能力値が－ 4 から－ 2 の領域で回答が信頼区間外にある者が目立つことがわかる。

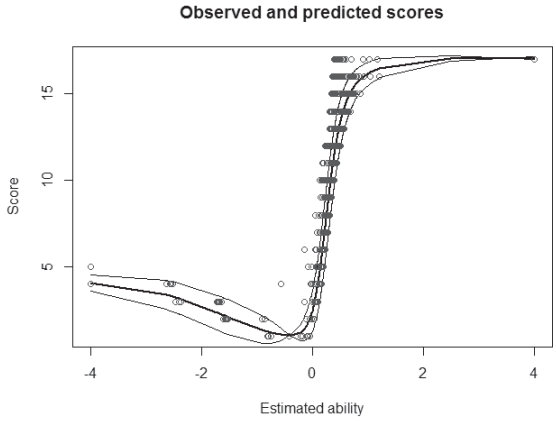

Fig. 8　関数 scp の分析結果（男性群）

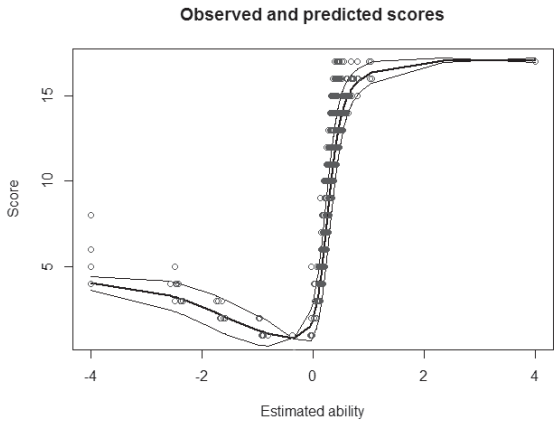

Fig. 9　関数 scp の分析結果（女性群）

また、情報関数については次のようなことが明らかになった。まず、テスト情報関数より両群共に 0 から 1 の領域における情報量が多いことがわかる。実際各項目の情報関数を見ると、能力値の 0 から 1 の領域で情報関数が最大である項目が両群共に15項目見られ、これらの項目は両群で同一であった。その他の項目は、能力値が－ 4 から 0 の領域で情報関数が最大である項目が 4 項目、能力値が 0 から 3 の領域で最大である項目が 2 項目、能力値が－ 4 から 4 の領域で最大である項目が 1 項目であり、これらの項目はすべて両群で同一であった。

そこで次に項目パラメータの推定を行い項目特性曲線を描いたところ、以下の 3 種類のグループが得られた（Fig10及び11）。第 1 のグループは能力値の増加に伴って正答率も増加する曲線で17項目、第 2 のグループは能力値の増加に伴って正答率が減少する曲線で 4 項目、第 3 のグループは能力値の

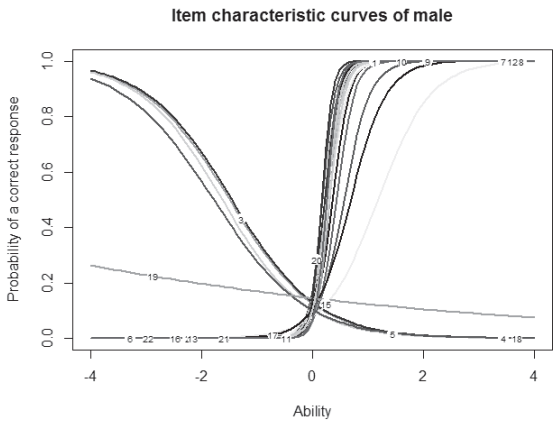

Fig.10　項目特性曲線（男性群）

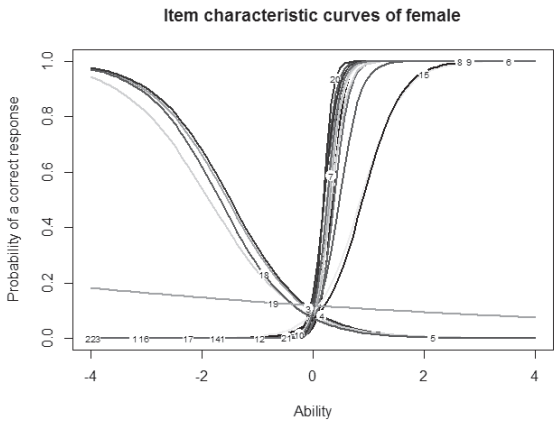

Fig.11　項目特性曲線（女性群）

全範囲に渡って正答率があまり変化しない曲線で 1 項目である。これらのグループは項目特性曲線の項目識別力パラメータと項目困難度パラメータの値に依存して生じるものであるから、これらの 2 個のパラメータの散布図においても同様の 3 グループが得られるはずである。そこで、項目識別力パラメータと項目困難度パラメータの散布図を Fig12及び13に示す。Fig10とFig12を比較してみると、男性群においては項目19だけのグループ、項目 3 、 4 、 5 、18のグループ、残りの項目から構成されるグループの 3 個のグループが存在することがわかる。同様のことは Fig11とFig13を比較することによって女性群についてもいえることがわかる。

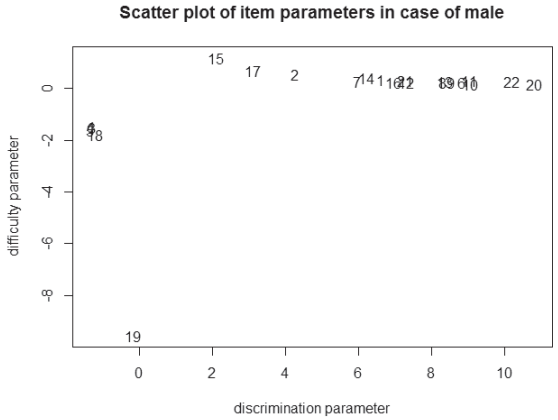

Fig.12　項目識別力パラメータと項目困難度パラメータの散布図（男性群）

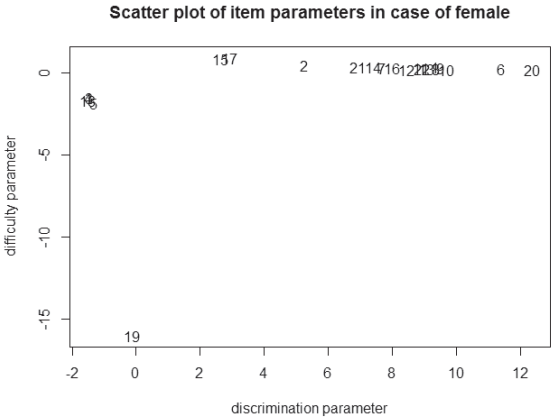

Fig.13　項目識別力パラメータと項目困難度パラメータの散布図（女性群）

最後に能力パラメータの推定について述べる。能力パラメータの推定は関数 mlebme を用いた最尤推定法と MAP（maximum a posteriori）推定法及び関数 eap を用いた EAP（expected a posteriori）推定法に基づいて行う。スペースの関係で最尤推定法による両群の能力パラメータのヒストグラムをそれぞれ Fig14及びFig15に示す。Fig14とFig15の比較より、能力値が－0.5から 0 の間の値をとる度数が男性群の方が多いこと、能力値が3.5から 4 の間を示す回答者が男性群において見られることより、女性群に比較して男性群の方が能力値が高い回答者が多いことがわかる。

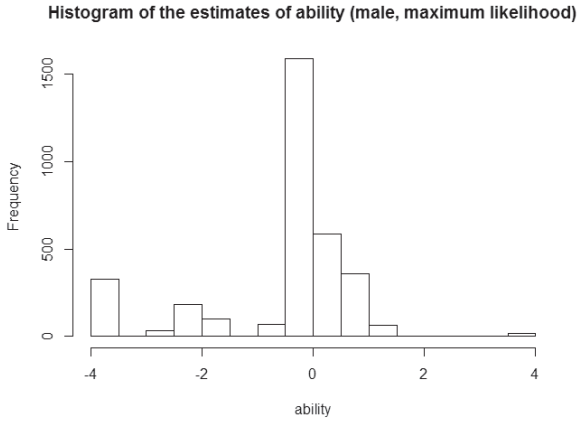

Fig.14　能力パラメータのヒストグラム（男性群）

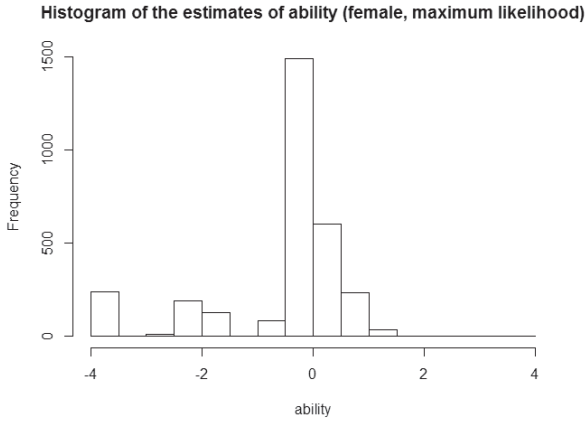

Fig.15　能力パラメータのヒストグラム（女性群）

考　察

PISA2012の日本のデータについて性別群に関する DIF 分析を行ったところ、DIF が観察された

項目が22個得られた。続いて、これら22個の DIF 項目について2PL モデルを用いた分析の結果、女性群より男性群の方が能力値が高い者が多かった。ただし、両群においてテスト特性曲線が能力値の負の領域において減少関数である項目が見つかるなど、今後の検討課題も見つかった。

本研究において分析の対象としたデータは科学的リテラシー項目であるが、このデータの DIF 項目において男性群の方が能力値が高い結果が得られたことは、科学的リテラシーに関する最近の先行研究と一致する。

まず、PISA の日本のデータを分析しまとめた報告によれば、PISA2000のデータでは科学的リテラシー得点が女子が男子より 7 点高いが、統計的な有意差はないことが報告されている（文部科学省, 2001)。また、PISA2003のデータでは男子が女子より 4 点高いが、統計的な有意差はないことが報告されており（文部科学省, 2004)、PISA2006のデータでは男子が女子より 3 点高いが、統計的な有意差はないこと（国立教育政策研究所, 2007)、PISA2009のデータでは女子が男子より12点高いが、統計的な有意差はないことが報告されている（国立教育政策研究所, 2010)。ところが、PISA2012では男子が女子より11点高く、統計的な有意差があると報告している（国立教育政策研究所, 2013)。以上より、過去 5 回の PISA のデータにおいて初期の 4 回のデータでは性差が見られなかったのに対して、2012年実施の PISA のデータにおいて性差が見られたことは、わが国において科学的リテラシーの性差の拡大を示す証左と言える。

PISA のデータ以外にも科学的リテラシーの性差を明らかにした研究は見られる。内田・守（2016）理科における性差が中学校の段階で現れ、それは数学における性差と類似していることを指摘した。また、河野・池上・中澤・藤原・村松・髙橋（2004）では、中学生における理科に対する意識や態度に及ぼす家庭的背景の影響について論じており、親が持っている理科の重要性に関する認識が高い程理科に対する好感度が高いことを明らかにしている。リベルタス・コンサルティング（2014）では、平成24 年度全国学力・学習状況調査で明らかになった中学生の理科嫌いの増加を踏まえ、理科に関する意欲・関心等が低下する要因分析を行った。その結果、理科がわからないことや理科の実験・体験が少ないこと、理科の授業の内容を普段の生活と関連づけられないことが、中学校における理科の関心・意欲の低下の主な要因であることが明らかにされた。即ち、学校の理科の授業における学習内容と生活体験との乖離が理科嫌いの原因の一つであると言える。

以上より、わが国における科学的リテラシーや理科における性差の存在が明らかにされたような状況であるが、本研究で指摘した男女群間に存在する本来の性差と DIF の結果生じる群間の相違との混同という面から考察すると、更に厳密な検討と議論が求められると思われる。即ち、本来能力差に基づいて定義される変数である科学的リテラシーと、課された問題の持つ属性に起因して群間に現れる相違である DIF は区別されるべきであるが、先行研究では果たしてこの点がどのように扱われたのか必ずしも明確ではないことが気がかりである。本研究で明らかにされたように、男女群間において22個の DIF 項目が得られ、それらの項目を用いた IRT モデルによる分析の結果男女群間で能力差が見られたことは、先行研究において PISA2012において科学的リテラシーの性差が見られたことに対する疑念を提出するものである。本研究の最初で述べたように、両群において DIF を除去した上で現れる両群間の相違が本来の意味での群間差であるならば、科学的リテラシーの性差を明らかにするためには、最初に DIF を示す項目群を除去し、次に残った項目群において科学的リテラシーの得点が両群で異なるかどうかを検討すべきである。

参考文献

Caro,D. and Biecek,P.（2016）intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software*.

Drasgow,F., Levine,M.V. and Williams,E.A.（1985）Appropriateness measurement with

polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.
Falissard, B. (2015) Package 'psy'. https://cran.r-project.org/web/packages/psy/index.html.
北條雅一（2013）数学学習の男女差に関する日米比較. Discussion Paper No.1301, Kyoto Institute of Economic Research.
伊佐夏実・知念渉（2014）理系科目における学力と意欲のジェンダー差. 日本労働雑誌, No.648, 84-93.
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
河野銀子・池上徹・中澤知惠・藤原千賀・村松泰子・髙橋道子（2004）. ジェンダーと階層からみた「理科離れ」 －中学生調査から－. 東京学芸大学紀要, 第1部門, 教育科学, 55, 353-364.
国立教育政策研究所（2007）. OECD 生徒の学習到達度調査～2006年調査国際結果の要約～.
国立教育政策研究所（2010）. OECD 生徒の学習到達度調査～2009年調査国際結果の要約～.
国立教育政策研究所（2013）. OECD 生徒の学習到達度調査～2012年調査国際結果の要約～.
Lord,F.M. (1980) *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.
Magis,D., Beland,S. and Raiche,G. (2015) difR: Collection of methods to detect dichotomous differential item functioning (DIF).
Magis,D., Beland,S., Tuerlinckx,F. and de Boeck,P. (2010) A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42,* 847-862.
文部科学省（2001）. OECD 生徒の学習到達度調査 ～2000年調査国際結果の要約～.
文部科学省（2004）. OECD 生徒の学習到達度調査 ～2003年調査国際結果の要約～.
Özdemir,B. (2015) A comparison of IRT-based methods for examining different item functioning in TIMMS 2011 mathematics subject. *Procedia - Social and Behavioral Sciences, 174,* 2075-2083.
Partchev,I. (2015) Package 'irtoys'. https://cran.r-project.org/web/packages/irtoys/index.html.
Raju,N.S. (1988) The area between two item characteristic curves. *Psychometrika, 53,* 495-502.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
リベルタス・コンサルティング（2014）. 「全国学力・学習状況調査の結果を用いた理科に対する意欲・関心等が中学校段階で低下する要因に関する調査研究」調査報告書.
Rizopoulos,D. (2006) ltm: An R Package for Latent Variable Modeling and Item Response *Analysis. Journal of Statistical Software, 17,* 1-25.
RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.
Thissesn,D., Steinberg,L. and Wainer,H. (1988) Use of item response theory on the study of group difference in trace lines. In H.Wainer and H.Braun (Eds.), *Test Validity* (pp.147-170), Hilladale, NJ: Erlbaum.
内田昭利・守一雄（2016）. 女子中学生は理科で躓く －中学校 3 年間の教科ごとの成績推移の分析－. 信州大学教育学部研究論集, 第 9 号, 95-111.
Yen,W.M. (1993) Scaling performance assessments: Strategies for mapping local item dependence. *Journal of Educational Measurement, 30,* 187-213.
Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P.W. Holland and H.waimer (Eds.), *Differential Item Functioning* (pp.337-347).