広島工業大学紀要教育編
第15巻（2016）39–43
論　　文

異なる集団間での項目反応理論パラメータ

――数学プレースメントテスト成績の大学間比較の一例――

廣瀬　英雄*

（平成27年９月24日受付）

Parameters for the Item Response Theory between Different Groups

—— An Example of Comparison of the Mathematics Placement Tests between Two Universities ——

Hideo HIROSE

(Received Sep. 24, 2015)

Abstract

The item response theory (IRT) is often used in modern educational evaluation scene. However, if the two groups are different from each other, the estimated IRT parameters such as item difficulty parameters, discriminant parameters, and students' ability parameters in each group may differ from each other. We have investigated this phenomenon using the two universities' testing results (high school level mathematics placement tests). The universities are both specified to science and technology. The number of students in investigation in a university is 145, and 179 in another university. As a result, we have found that 1) the estimates for the item difficulty parameters show no clear differences between the two groups, 2) the estimates for the students' ability parameters differ from each other if the two groups are different. To compare the difference of ability among groups, use of combined data is strongly required.

Key Words: different groups, placement test, item response theory, classical test theory

1　はじめに

大学でよく行われている成績評価法は，教員が問題毎にあらかじめ配点を決め，総合点により学生の成績評価としている方法である。これは古典的テスト理論に沿った方法である。学生の能力値のランクは，平均，標準偏差を求めた後で，その集団での（正規分布に従うと仮定したときの）確率的な位置から決定できる。しかし，成績は問題の配点によって変わるので，適切な配点であれば問題はないが，教員が恣意的に配点を変えれば評価を変えることができるため，一般に公平性が保たれているとはいえない。

これに対して，問題の難易度も評価に加えて公平性と公正性を目指した方法が，現代テスト理論の一つ，項目反応理論（Item Response Theory, IRT）[1-3] である。ここではIRTを用いた評価法を用いて評価を行うことを試みる。IRTによれば，学生の能力値に加えた問題毎の評価も同時に可能になり，それを使うことによってアダプティブなテスト[4]も可能になる。また，オンラインのテストも可能になる。IRTの理論についてはその概略を付録に示す。ただし，どちらの方法でも，学生がある集団内では相対的にどの位置にいるかを決めることによって評価値を決定している。従って，集団が異なれば評価値が異なる可能性がある。評価値を絶対評価に近くなるようにするには，集団の規模が大きいことが望ましい。ここでは日常よく見られる比較的小さい

* 広島工業大学環境学部環境デザイン学科

大きさの集団間での検討を行った結果について述べる。

異なった集団間で成績評価の比較を行うには絶対評価に近い方法が好ましい。両集団を含む更に大きな集団の中での評価を行う必要がある。例えば多くの受験者をかかえるTOEIC 試験では絶対評価に近い評価が可能になる。ただし，そのような比較は日常なかなか難しい。そのため，あらかじめ決められた同じ配点で総合点の比較を行うことがよく行われる。絶対評価に近いはずの IRT でも異質な集団間比較は難しい。そこで，ここでは，IRT を使った集団間の比較は可能なのか，可能ならどの程度可能なのかを，ある程度集団の性格が異なる二つの大学で行ったプレースメントテストのサンプリング結果を用いて行ってみた。ここでは，大学入学直後の学生に高校数学の試験問題を課し，評価結果の比較を行った結果について述べる。

2　二つの集団で実施された試験の方法について

異なった集団でも同じ時期に同年代の学生に試験を課すことが望ましいが，いろいろ制約があるので，ここでは違う時期ではあるが学生はH大学とK大学（どちらも工科系大学）の二つの大学で，入学直後のサンプリング学生を対象とした。集団のサイズは同程度が好ましいと思われるので，サイズ効果が入らないように同定度の大きさにした。H大学では，30人クラスと60人クラスについての１年生の解析学のコース（必修），60人クラスについての１年生の線形代数学のコース（選択），および２年生の解析学のコース（15人，20人程度の選択科目）で合計145人を対象とした。K大学では，86人クラスと93人クラスの合計179人を対象とした。

プレースメントテストについては，数Ⅰ，数A，数Ⅱや数B，および数Ⅲや数Cすべてを調べることにした。2014年度以前の古い学習指導要領に従う学生もいることを考慮して，新しい指導要領内容には従わず以前のものに従うテスト内容とした。そのため，行列の問題も１部含まれている。具体的には，数Ⅰが７問，数Aが３問，数Ⅱが６問，数Bが２問，数Ⅲが９問，数Cが７問の合計34問である。H大学での回答法はマークシート方式（９択）で，K大学での回答法は記述式であるが，マークシートの選択肢が多いので評価結果にそれほど差異は出ないものと考える。

3　IRT による能力評価と問題の困難度評価

これまでプレースメントテストの成績評価については，あらかじめ決められた素点について正解した点数を合計した総合点で行っていることが多い。いわゆる古典的テスト評価法である。例えば，H大での成績は図１のように総合的に表わすことができる。また，問題の難易度は，個々の問題に関する正答の割合から求めることができる。例えば，H大でのテスト結果からは図２のように問題の難易度を求

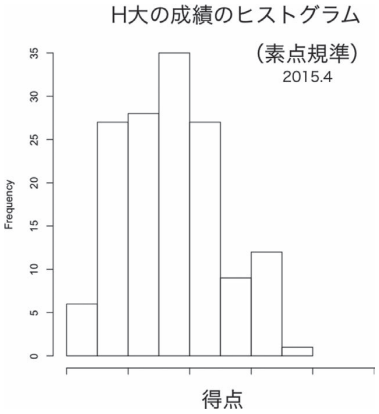


図１　H大での総合成績のヒストグラム

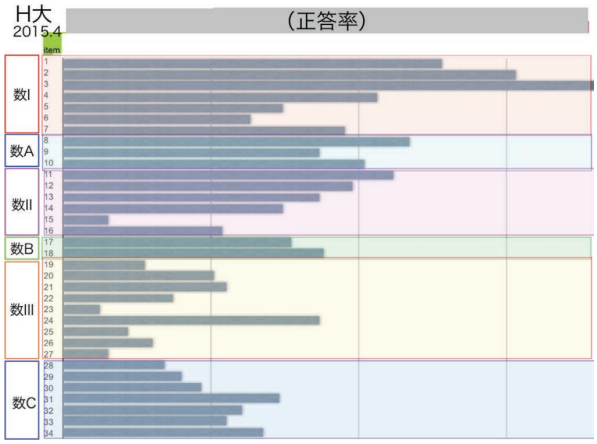


図２　古典的方法による問題の難易度（H大）

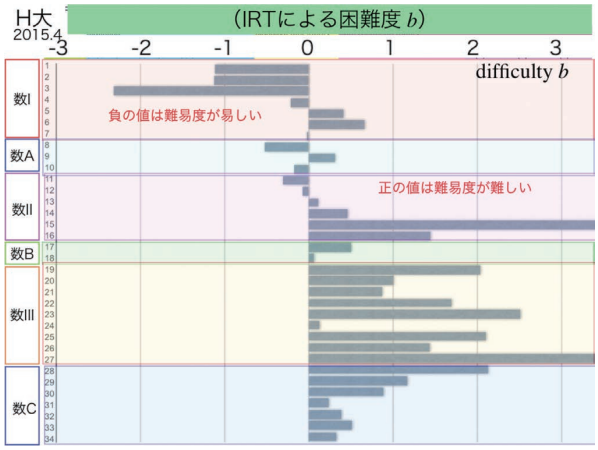


図３　IRT による問題の難易度（H大）

めることができる。数Ⅰ，Aはある程度易しいが，数Ⅲ，Cについては難しいことが表されている。

ただし，古典的方法では問題間の関連性を使った問題の難易度については計算されてはいない。IRT は，学生の習熟度評価に加えて問題の難易度も考慮して評価を行うことができる。例えば，H大での問題の難易度は図３のように表される。負の値は相対的に易しい問題を，正の値は難しい問題を表している。古典的評価法と同様，数Ⅰ，Aはあ

る程度易しいが，数Ⅲ，C については難しいことが表されているが，それに加えて数Ⅱ，B の難しさも表されている。IRT ではこのパラメータの他に識別度を表す指標も設けている。識別度は問題が学生の習熟度判定にどのように貢献するかを表したもので，いわゆる良問，悪問の指標として表現することができる。

IRT は問題の難易度と学生の習熟度とを同時に評価できる。図 4 はH大の学生の習熟度を表したものである。習熟度はロジスティック関数で表されているが，近似的には標準正規分布と同じように表現されたものと考えてよい。図で B11，B21，C11 は 1 年生，C21 は 2 年生である。

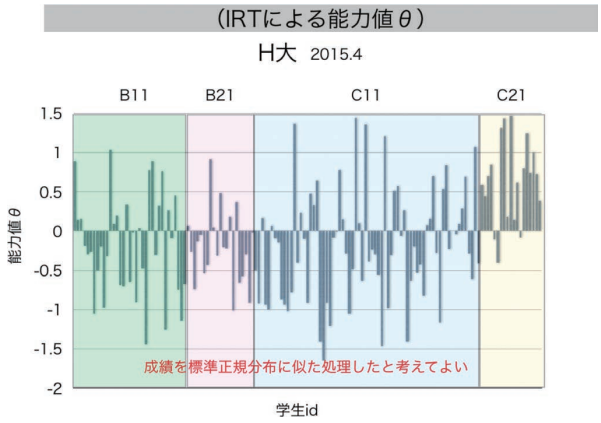


図 4　IRT による学生の習熟度（H大）

4　2 大学の情報を同時に用いた IRT 評価結果

これまでは一つの大学内だけでの評価について述べた。ここでは，二つの大学でのプレースメントテスト結果を同時に用いた結果について述べ，それぞれの大学内で求めたパラメータ，二つの大学を同時に用いたパラメータの推定結果について比較を行う。

4.1　困難度パラメータ *b* の比較

困難度パラメータは調べた集団によって求めることができる。ここではH大単独，K大単独，H大とK大の両方で求めた困難度パラメータを比較してみる。図 5 は34問の問題についての困難度をH大単独での結果，K大単独での結果で比較したものである。一つの組織であれば困難度パラメータは（例え成績が異なっても）どちらも似たような値になるようにも思われるが，図を見るとかなり明確に異なっていることが分かる。ほぼすべての問題が点線の上側に位置している。

図 6 は問題毎の困難度を，1）H大単独，2）K大単独，3）H大とK大の両方で求めた結果について棒グラフで比較してみた。棒グラフでは分かりにくいので（折線で表すことに意味はないが）比較結果を明瞭に区別するために折

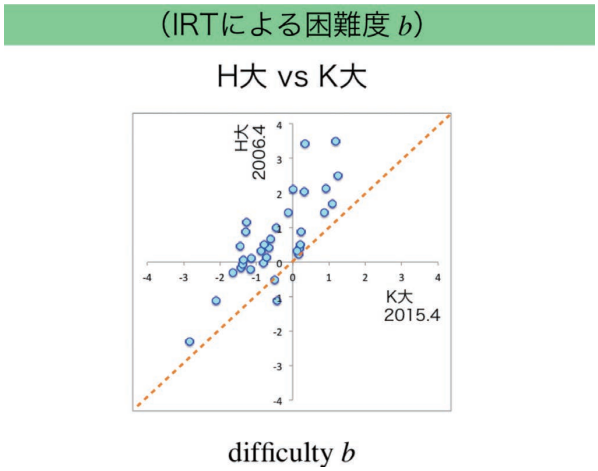


図 5　困難度パラメータの比較（H, K）

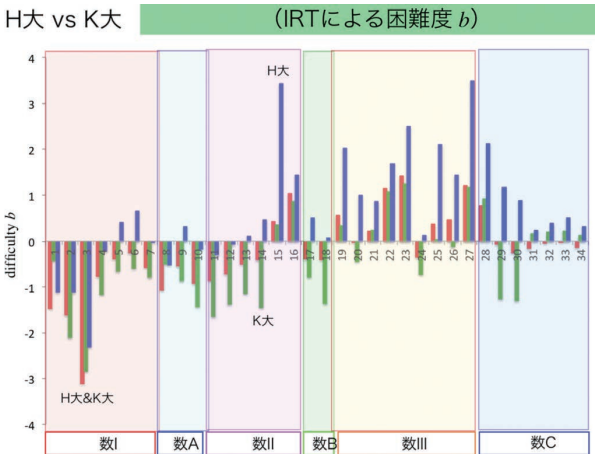


図 6　困難度パラメータの比較 1（H, K, H&K）

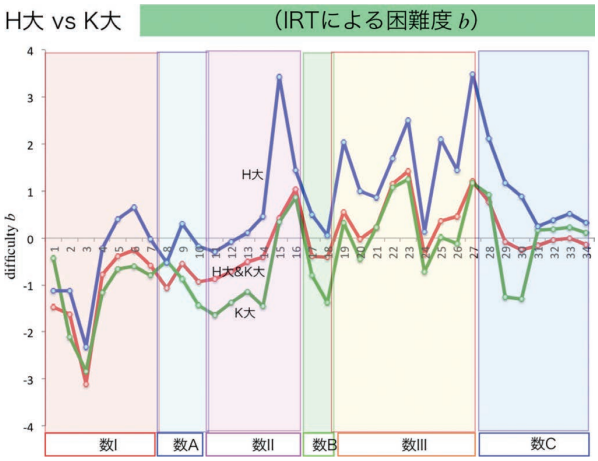


図 7　困難度パラメータの比較 2（H, K, H&K）

線でも表してみたものが図 7 である。H大とK大の両方で求めた結果は，それぞれH大単独，K大単独で求めた結果の間に位置している。このことは，一つ一つの集団で求めた IRT の困難度パラメータは，その集団内でしか通用しないように思われるが，実はそうではなく相対的に区別できるように求められていることを示している。つまり，異

なった集団毎に求められたパラメータは，ある程度公平性を保てるような結果になっていると考えられる。

4.2　識別度パラメータ a の比較

図８に識別度パラメータの比較を示す。数Ⅲ，Ｃを除くほぼすべての問題で識別度は0.5から１の間に分布しているので特に問題はなくほぼ良問と言える。際立っているのはＫ大での数Ｃの最後の４問である。極めて識別度が高い。この問題は離散型の確率分布を求める問題である。

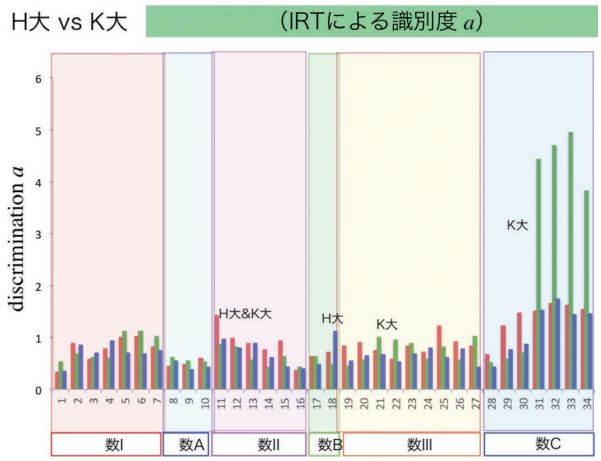


図８　識別度パラメータの比較（H, K, H&K）

4.3　能力パラメータ θ の比較

集団が変われば能力推定値も変わる。ここではそれがどの程度なのかを調べてみる。調査する集団はＨ大生145人，Ｋ大生93人クラスと86人クラス（合計179人）の二つである。比較はＨ大145人とＫ大生 86人との間で行った。

１）図９は調査したすべての学生についてその能力推定値を示したものである。おおよそ３つの区分に分かれている。一番左はＨ大とＫ大の両方のデータを用いて求めたＫ大生93人クラスの推定値，中央は86人クラスの推定値を（1）Ｈ大とＫ大の両方のデータを用いて，（2）Ｋ大だけのデータを用いて求めたＫ大生86人の推定値，一番右は（1）Ｈ大とＫ大の両方のデータを用いて，（2）Ｈ大だけのデータを用いて求めたＨ大145人推定値を表している。少し分かりにくいので次にＨ大，Ｋ大それぞれについて，Ｈ大とＫ大の両方のデータを用いた場合の比較を示す。

２）図10はＫ大の学生についてその能力推定値を示したもので，（1）Ｈ大とＫ大の両方のデータを用いて，（2）Ｋ大だけのデータを用いて求めたＫ大生86人の推定値を表している。

３）図10はＫ大の学生についてその能力推定値を示したもので，（1）Ｈ大とＫ大の両方のデータを用いて，（2）Ｈ大だけのデータを用いて求めたＨ大生145人の推定値を表している。

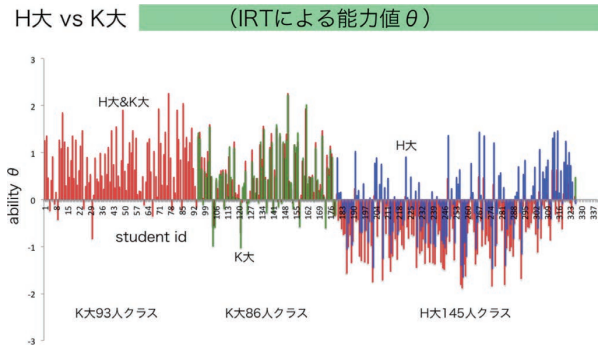


図９　能力パラメータの比較（H, K, H&K）

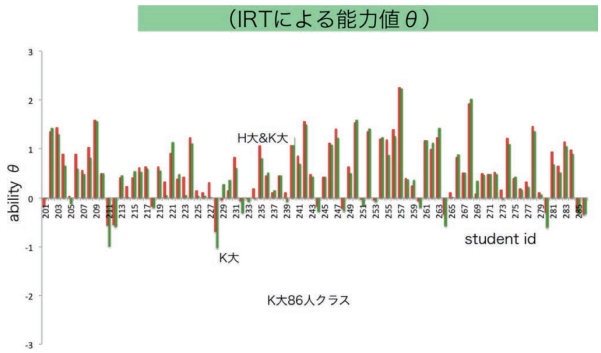


図10　能力パラメータの比較（K, H&K）

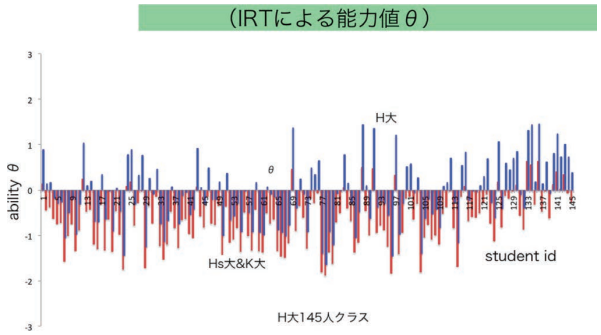


図11　能力パラメータの比較（H, H&K）

図10からは，Ｈ大とＫ大の両方のデータを用いた場合の結果とＫ大だけのデータを用いて求めた場合の能力値の推定値にそれほどの違いが見られない。しかしながら，図11からは，Ｈ大とＫ大の両方のデータを用いた場合の結果とＨ大だけのデータを用いて求めた場合の能力値の推定値に明らかな違いが見られる。Ｈ大だけでの推定値は正の値と負の値がバランスよく配置されているのに対し，Ｈ大とＫ大の両方のデータを用いた場合ではほとんどすべての学生について負の値が示されている。

図12に，Ｈ大とＫ大の両方のデータを用いた場合を横軸に，Ｈ大またはＫ大のみのデータを用いた場合を縦軸に配置した推定値を示す。図10，11では詳細には見えなかった違いがここでははっきり見える。つまり，Ｋ大の推定値は，

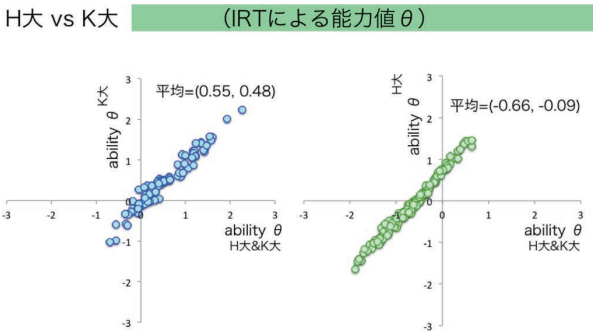


図12　能力パラメータの比較（H, K, H&K）

両方および単独のデータを用いても同じように正の方向に割合が多く，H大ではその逆である。また，K大では平均値の差は小さいが，H大ではそれが大きい。集団内での個々の学生の比較は，両方のデータを用いても単独のデータを用いても可能であるが，集団間の学生の比較は両方のデータを用いたときにしか行えない。このことは問題の難易度の場合と異なっている。

5　まとめ

異なる集団間での項目反応理論（Item Response Theory, IRT）のパラメータの推定値の比較を，二つの大学で行った高校数学のプレースメントテストの結果を用いて具体的に行った。その結果，次のようなことが分かった。1）IRTを使えば古典的な比較法よりもより詳しい分析が可能になる，2）問題の難易度については，異なった集団間の差異はそれほど際立っては見えない，つまり，二つの集団のデータを同時に用いて求めた難易度の推定値と一つの集団で求めた推定値の間にはそれほどの差は見られない，3）しかしながら，学生の能力推定値になると，二つの集団のデータを同時に用いて求めた難易度の推定値と一つの集団で求めた推定値の間には大きな差異が認められる。異質な集団間での能力推定値を比較する際にはすべてのデータを用いた土俵の上で行わなければならない。

文　献

1）R. K. Hambleton and H. Swaminathan, Item Response Theory: Principles and Applications. Springer, 1984.
2）R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory. Sage Publications, 1991.
3）W. J. D. Linden and R. K. Hambleton, Handbook of Modern Item Response Theory. Springer, 1996.
4）H. Hirose and T. Sakumura, Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2012, pp. 8-12, 2012.

付録　項目反応理論（Item Response Theory, IRT）

これまでの評価法では，各問題にはあらかじめ配点が与えられ，それぞれの問題の得点を合計した総得点が評価値であった。同じ試験を多くの人に課せば全員の総得点が得られる。そこから平均や標準偏差を算出すれば，自分の相対的な評価値を偏差値という形で求めることができる。しかし，問題の配点を変えれば総得点が違ってくる場合がある。配点によって評価値が変わるのは公正な評価法とはいえないかもしれない。そこで，問題の難易度と各受験者の学習習熟度とを同時に求めながら，公正で公平な評価法が提案された。これがIRTによる評価法である。この理論は，これまでにTOEFLや情報処理検定など多くの公的な場面で適用されている。ここではこの評価法を用いている。

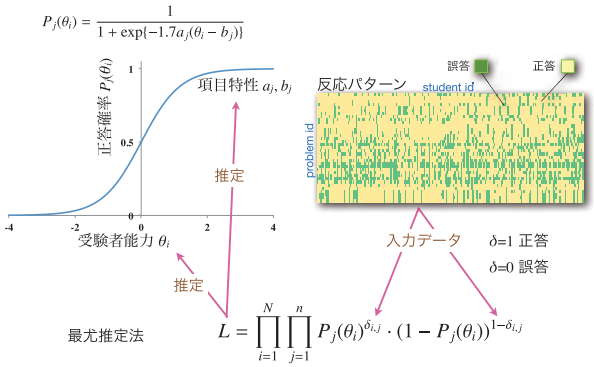
IRTでは，各問題jに対する受験者iの評価確率$P_j(\theta_i)$が2パラメータロジスティック分布，

$$P_j(\theta_i)=\frac{1}{1+\exp(-1.7a_i(\theta_i-b_i))}$$

に従っていると仮定する。a_j, b_jは問題jの識別力（簡単にいうと，問題の良し悪しを表す）と困難度（文字どおり，問題の難易度を表す）を，θ_iは受験者iの習熟度を表している。受験者$i=i,\cdots,N$が項目$j=1,\cdots,n$に対して取り組んだ結果，その解答が正答なら，$\delta_{i,j}=1$，誤答なら$\delta_{i,j}=0$と書き表すと，すべての受験者がすべての問題に挑戦した結果（これを反応パターンという）の確率は，独立事象を仮定すれば，

$$L=\prod_{i=1}^{N}\prod_{j=1}^{n}P_j(\theta_i)^{\delta_{i,j}}(1-P_j(\theta_i))^{1-\delta_{i,j}}$$

と表される。これを尤度関数という。付録図に，IRTによる評価の過程のイメージを示す。誤答0と正答1からなる$\delta_{i,j}$を上の尤度関数Lに代入し，それを最大にするようなa_j, b_j, θ_iを同時に求めるのがIRTによる評価法である。



付録図　項目反応理論の概念図