広島工業大学紀要研究編
第54巻（2020）149-154
論　　文

テストの点数のゆらぎについて

廣瀬　英雄*

（令和元年9月18日受付）

On Fluctuation of Score of Examination

Hideo HIROSE

（Received Sep. 18, 2019）

Abstract

Comparing the results of placement tests and final exams performed in a class, it has been common to see the whole aspect and investigate the correlation between them. It seems that little has been studied whether useful information can be obtained. In other words, the scores x_i and y_i of the student i in the class are only the simultaneous samples of the random variables X and Y, and the relationship between x_i and y_i was not deeply investigated. However, a score of student A is governed by A's ability and proficiency level, and it is normal to think that there will be bounded fluctuations. In other words, the range of variation of x_i is not defined in the domain of X, but is more limited. In this paper we pursue such a matter.

By doing this, it is possible to observe fluctuations in test scores under various conditions, which cannot be eliminated due to the form of the test (descriptive or multiple choice) among teachers. It is also possible to extract variations due to teaching manners of teachers. Here, the basic idea to deal with such things is described.

Key Words: fluctuation, learning check testing, computer based testing, multiple-choice paper questions, description type questions, item response theory, end-term examination.

1　はじめに

実に奇妙なことに、ほとんど誰もが、テストを受けた点数はさも絶対かのように、テストの結果に従って異議を唱えない。ただ一度しか受験しないテストで、その点数が60点以上は合格、59点以下は不合格、というように。それは、これ以上に公平感を持たせる説得性のある方法論がないからである。ただし、失敗したとき、ほとんど皆、今回はチャンスがなかっただけだと気がついており、結果については合理的にあきらめて受け入れている。しかし、その程度について、どのくらいの人たちが理解してこのことを受け入れているのかはなはだ疑問である。

例えば、全く同じようなテストを続けて2回受けるとする。このとき、1回目で合格したとき、2回目でも合格するチャンスはどの程度なのだろうか。2回目では失敗することも、合格することももちろんある。つまり、テストの点数はゆらいでいると考えられる。その原因は説明変数（共変量）によって説明されるものもあるだろうし、純粋に確率的変動だけに由来するものもあるだろう。いろいろ考えられる。こういうことを理論的に把握していないと、2回目でも合格するチャンスはどの程度なのだろうか、というような問いにはなかなか応えられないのではないだろうか。

* 広島工業大学データサイエンス研究センター & 環境学部建築デザイン学科

あるいは、数学Ⅰのテストの点数と数学Ⅱのテストの成績の相関はどうなっているのだろうか、ということはよく取り上げられているが、その実態はきちんと把握されているのだろうか。相関係数だけで議論が進められていないだろうか。

ここでは、このような、テストの点数のゆらぎについて考えてみたい。このことを理解することによって、日常の行動がテストの点数に支配されることがなくなり、更には、より適切なテストの結果の使い方までを考えることができると思われる。

2　TOEIC を 2 回受験したときの点数

図 1 は、大学に入学してすぐに入学生全員が受験した TOEIC の点数と、同じ学生の約 1 年後の TOEIC の点数を比較したものである。図の左は2006年の 1 年生と2007年の 2 年生、右は2007年の 1 年生と2008年の 2 年生である。大学に入って英語の授業を受けているはずなので 1 年後には TOEIC の点数も上がっていることが期待される。

しかし、実際には、2006年入学生の平均は19点上がっているものの、個々の学生の点数はかなり変動しており一概に 1 年後点数が上がっているとは言えない。個々の学生の両者間の差の標準偏差は72点程度にもなる（2006年での標準偏差は82点、2007年での標準偏差は85点）。相関係数は 0. 63程度である。2007年入学生の場合の同様である。なお、当時の ETS に確認したところ、同一人物が実力の上下がない状態で複数回受験する場合の点数のぶれはおよそ50点と見込んでいるとのことであった。

つまり、TOEIC のような、社会的には評価値に比較的絶対的な信頼度が持たれているものでも、テストの点数にはゆらぎが発生しており、その大きさまで観測されている。

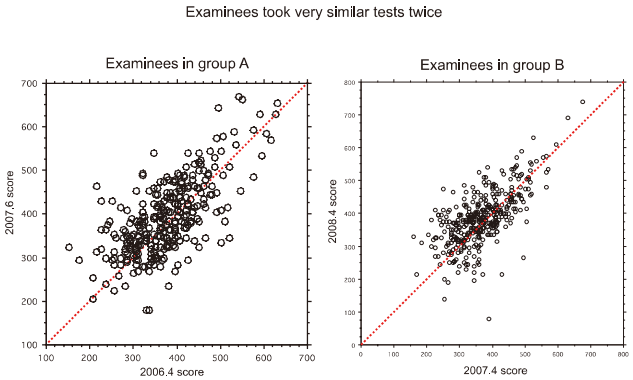

図 1　同一人物が 2 回 TOEIC を受験したときの 1 回目と 2 回目のスコアの関係
（1 回目は大学入学直後、2 回目は 1 年後）

ちなみに、大学入学後 3 年経った時点でどうなっているかを調べた結果が図 2 である。点数が大きく伸びている学生が幾人かみられるものの大半はあまり変わっていない。大きく伸びた学生にインタビューしたところ、海外への留学を契機に勉強の意欲が湧いたとか、さまざまな動機による個人的な取り組みの姿勢が大きく影響しているようであった。逆に、スコアを下げた学生のその理由は、入学試験のための勉強のインセンティブがなくなった、というケースが少なくなかった。

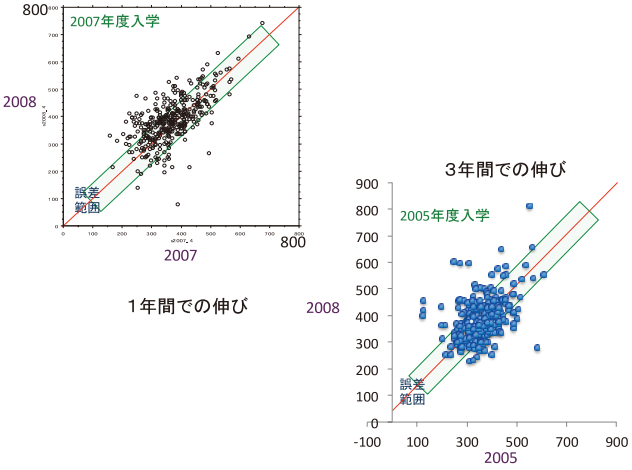

図 2　同一人物が 2 回 TOEIC を受験したときの 1 回目と 2 回目のスコアの関係
（1 回目は大学入学直後、2 回目は 3 年後）

3　2 つのテストの点数の確率変数の取り扱い

相関を持つ 2 変量の確率変数を表すとき、確率変数 X と確率変数 Y が同時分布関数 $F(x,\ y)$ に従い、同時密度関数 $f(x,\ y)$ を持つ、というように考えることが多い。例えば、数学Ⅰのテストの点数を X、数学Ⅱのテストの点数を Y とするように。数学Ⅰのテストの点数が高ければ数学Ⅱのテストの点数もある程度高いことが予想され、相関係数は例えば0. 7というような関係を想定することだろう。

図 3 は、大学入学直後に入学してきた学生全員に受けてもらったプレースメントテストで、数学 IA を中心とした問題（PTA と呼ぶ）に対応したテストの点数と数学 IIB および数学 III までを中心とした問題（PTB と呼ぶ）に対応したテストの点数の関係をプロットしたものである。学生数はおよそ1100人。数学 IA の問題は暗記型の問題ではなくよく考えないと解けない問題になっており、数学 IIB および数学 III の問題はその分野の知識を問うどちらかといえば暗記型の問題になっている。両者の間には正の相関があることは期待される。相関係数を計算すると0. 72となっている。

通常、このような前提に立って議論が進められることが多い。さて、ここで、X、Y のある観測値 x_i y_i の値は何を表しているかと言えば、もちろん i 番目の学生の数学 IA のテストの点数 x_i、数学 IIB および数学 III の点数 y_i である。

A君のテストの点数は、仮定されている X、Y の同時分布に従う分布から抜き出されたものとみなす、と考える。ここで、現実のデータからモデル（同時分布確率モデル）に移行する間には捨象化あるいは抽象化が入って、両者は異なるものであるというのはわざわざ断らない。

さて、現実的な場面ではA君の実力というものを想定することがある。例えば、A君は普段よくできる、あるいはあまりできない、といった具合に。このとき、A君のサンプルを先の同時分布からどうやって自然に拾うことができるだろうか。x_i y_i は分布の頻度にそってランダムに選ばれるだけで、特にA君の個性を感じて選ばれている訳ではない。何か足りない、と感じる。

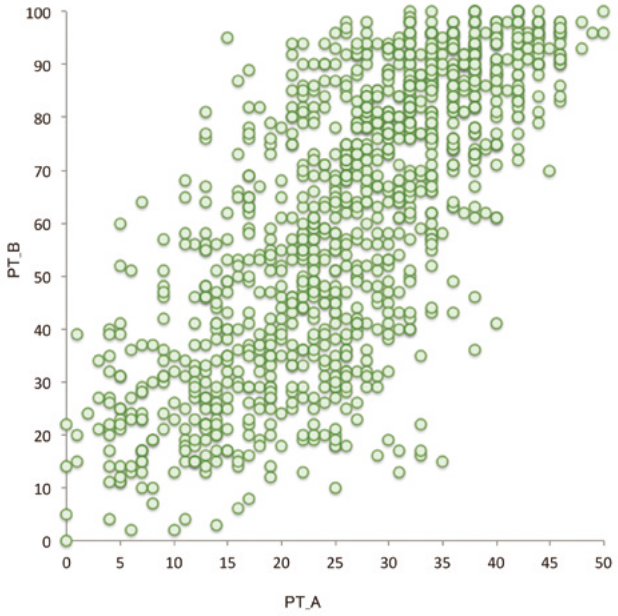

図3　2つのプレースメントテスト結果のスコアの関係
（PTA：数学IAで主に考える問題を出題、
PTB：数学IIBと数IIIで主に知識を問う問題を出題）

ここで、1つの仮想データを考えてみよう。図4は、図3で用いたPTAの22問の回答結果で、奇数番目の問題を集めたときの成績（項目反応理論、IRT、を用いている）と偶数番目の問題を集めたときの成績を比較したものである。同じ人が同じような難易度の問題を同じ条件下で受験したときでも、両者の間には大きな開きがあることがわかる。これは、図1、2で示したような2つのテストの点数の比較とは異なり、点数のばらつきへの影響条件はかなり削減されていると考えられるが、やはり点数にゆらぎがある。

図5は、図4と同様な比較を行ったものであるが、対象をPTAのみから、PTAとPTBの両方を使っているので、問題数はおよそ3-4倍になっている。図1-3とは異なり、$X = Y$ の線の上を中心に同じような幅でばらついた能力値（IRTのability値）が観測される。つまり、2変量正規分布のような形にはなっていない。図4ではこのことは明瞭には見えていないがサンプル数が少ないためで、実際には同様な現象が起こっていると思われる。このことは何を示しているのだろうか。

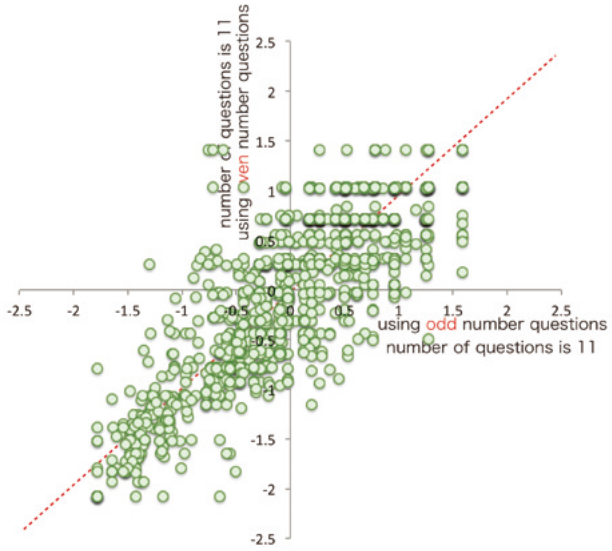

図4　PTAの22問の回答結果で、奇数番目の問題（1-21）11個を集めたときの成績と偶数番目の問題（2-22）11個を集めたときの成績の比較

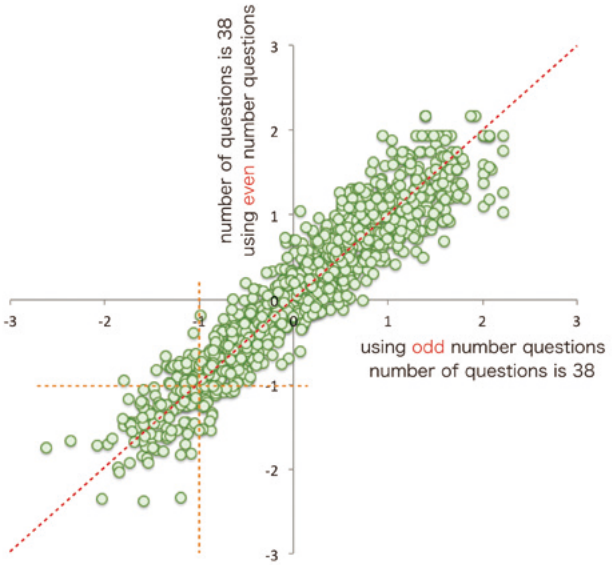

図5　PTAとPTBの77問の回答結果で、奇数番目の問題（1-75）36個を集めたときの成績と偶数番目の問題（2-76）36個を集めたときの成績の比較

A君は普段よくできる、あるいはあまりできない、ということについて、もう一度考えてみる。A君は、仮定した確率分布（例えば2変量正規分布）から抽出されたサンプルというよりは、A君の能力値（習熟度）というのはある程度固定されていて、観測される値は、固定された

値を中心に、そのまわりを一定のランダムな確率変数の挙動に従って動く値に加算したもの、と考えることはできないだろうか。

図6は、そのような仮説に従ったシミュレーションデータから作られた2つのテストの成績を比較したものである。ここで､学生の能力を p(p=0.1, 0.2, ,, , 0.9)とし(言い換えると、学生が問題を解ける確率が p である)、問題数を100問与えたときの正答数を成績と解釈している。

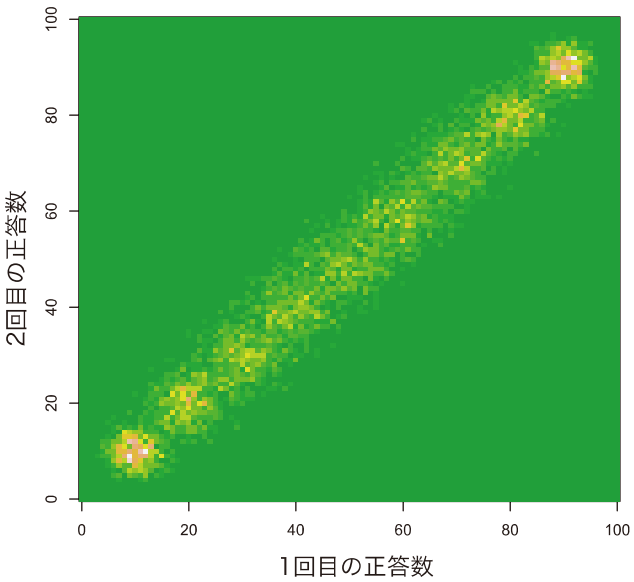

図6　学生の能力を p（p=0.1, 0.2, ,, , 0.9）と仮定し、シミュレーションデータから作られた2回のテストの成績の比較

図5、6は非常によく似ており、仮説の妥当性を示唆している。このことから、学生の ability というのはある確率分布に従っていると考えてよいが、個々の学生の ability は異なった平均を持っており、学生個々のテストの点数はそこを中心に確率的にゆらいでいる、つまり、X_1, , , , X_n のちらばり具合はある確率分布（例えば一様分布）に従い、各 X_i が観測された値 x_i には、ある値に一定の確率変動（例えば2項確率誤差）が加わっていると考えることができる。クラス全体を見る確率変数としては、X=(X_1, , , , , X_n)、といった多変量の分布を考える方が適切であるかもしれない。

一方、現代テスト理論では、はじめからテスト問題の難易度と受験者の能力の両方を未知数とし、尤度方程式を解いてそれらを求める、という形をとっている。尤度原理を用いているので、求められた推定値の信頼度も求められる。つまり、A 君の能力値はある程度固定された点 θ_i として固定され、その値のばらつきは尤度方程式を解く際に出てくる Fisher の情報量から求められる。つまり、正規分布近似としたときの標準偏差から求められる。従って、ability 値 θ_i のゆらぎが求められることになる。図4、5でドットで示されているものは、θ_i のまわりをゆらぐ点の観測値として考えることができる。この場合、ability を推定する際には、いろいろな共変量の影響を受けているものが自然に含まれていると考えてよい。ただし、θ_i の分布には普通標準正規分布が仮定されている。

それに対して、図4、5のように考えた場合のゆらぎは、共変量の影響を取り払った確率的変動だけが観測されていると考えることもできる。

そうすると、2つの異なったテストの間で観測されるゆらぎと、2つのほぼ同様なテストを単独に続けて2回行なったと仮定したときのゆらぎから、2回のテストの間にどのような発展（あるいは衰退）があったのかを推測することができる可能性がある。

4　教員の評価バイアスとその除去

これまでの例では、テストの採点はすべてコンピュータが行なっており、教員などのヒトの手は借りていない。しかし、200人くらいのクラスでは、クラスを分割して同じ内容を教える場合があり、例えば3人の教員が担当する際に、教員間で、教えることの優劣の差異、評価の差異が発生し、それが（図4－6のような確率的変動を受けない）学生が本来持っている能力値を変えて観測しているかもしれない。

例えば、図7は、200人程度のクラスを、100点満点のプレースメントテストの結果によって、上位2クラス（プレースメントテスト50点以上)、下位2クラス（プレースメントテスト50点未満）の4つのクラスに分割したときの上位2クラスでの、1）プレースメントテストの成績、2）LCT の ability 値、3）期末試験の成績（@、A、B、C、D の5段階）の関係を見たものである。

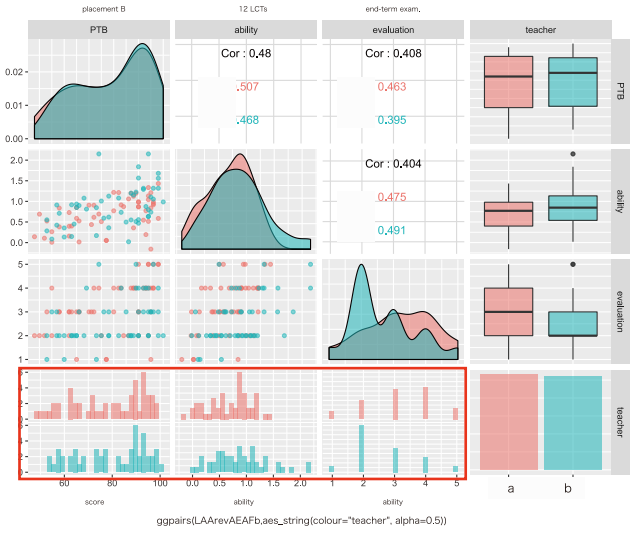

図7　プレースメントテストの成績、LCT の ability 値、期末試験の成績（@、A、B、C、D の5段階）の関係
（期末試験は筆記型で教員がそれぞれ採点）

図を見ると、プレースメントテストの成績と LCT の ability 値の分布は 2 人の教員間でそれほど変わらないが、最後の期末試験の成績は大きく異なった分布になっていることが分かる。これが、教えることの優劣の差異、評価の差異にあたる。少なくともヒトの手による評価値についてはバイアスを除去したい。

そこで、期末試験も多肢選択型の問題にしてすべてをコンピュータによって処理してみた。その結果が図 8 である。LCT と期末試験では IRT を用いている。比較のために同じ内容を用いたいが、同じ内容について 2 回の試みを行うことはできないので、別の試験になっている。200人程度のクラスを、プレースメントテストの成績をもとに、同じような分布を持つクラスを 3 つ編成し、そこで、1）

図 8　プレースメントテストの成績、LCT の ability 値、期末試験の成績（ability 値）の関係
（期末試験は多肢選択型でコンピュータが一括採点）

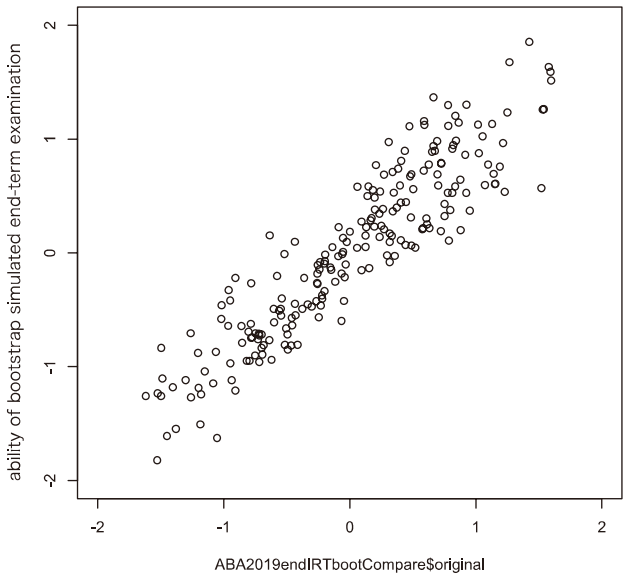

図 9　期末試験の成績（ability 値）の関係
（横軸は期末試験の結果そのもの、縦軸は bootstrap シミュレーションによって生成された結果）

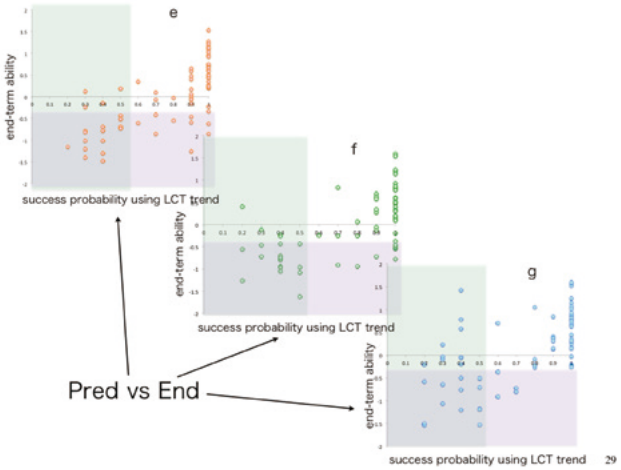

図10　LCT トレンドをもとにした学期半ばでの期末試験合格予測結果と期末試験の成績（ability 値）の関係
（横軸は合格予測結果、期末試験の成績）

プレースメントテストの成績、2）LCT の ability 値、3）期末試験の成績（@、A、B、C、D の 5 段階）の関係を見てみた。

3 人の成績評価は分布上あまり異なっているようには見えない。つまり、教えることの優劣の差異、評価の差異ともに、それらのバイアスが除去されているように見える。

同じような期末試験を 2 回受験したと仮定して両者を比較することもできる。1 回目は期末試験の結果そのもの、2 回目は仮想的に bootstrap シミュレーションによって生成された結果を用いることができる。その結果を図 9 に示す。

図 8 の分布から図 9 の分布を差し引いたものが教員による教えることの優劣の差異によるゆらぎになる可能性がある。実際に、3 教員それぞれについて、LCT トレンドをもとにした学期半ばでの期末試験合格予測結果と期末試験の結果を比較すると 3 教員の間には相違が見られる。図10にそれを示す。これを一括して見ると確率的変動以外のゆらぎにつながっているように思われる。

5　まとめ

これまで、クラスで行なったプレースメントテストと期末試験の結果を比較する際、全体をまとめて表示して両者間の相関を見ることはよくあったが、それ以上立ち入って、両者間から何か有用な情報が得られるかどうかについてはあまり調べられていなかったと思われる。つまり、クラスの学生 i の成績 x_i と y_i は確率変数 X と Y の同時分布というだけで、x_i と y_i の関係についてはそれらの性格を問うことはしなかったということである。しかし、A 君の成績は、A 君の能力や習熟度に支配されており、大きく変動することはないと考えるのが普通である。つまり、x_i の変動範囲は、X の定義域ではなく、それよりもっと限られている。

ここではそのことについての問題提起を行なっている。

そうすることで、さまざまな条件下でのテストの点数のゆらぎを観測する可能性ができて、除去できない確率的変動、試験の形態（記述式か多肢選択式か）による変動、教員間の教え方による変動を抽出することもできる。ここではそのことができる基本的な考え方について述べた。

文　献

1）廣瀬、ラーニングアナリティクス：LCT 成績と期末試験成績の関係、広島工業大学紀要教育編、pp. 59–63, Vol. 18, 2019.

2）廣瀬、大規模オンラインテスティングから得られるラーニングアナリティクス、広島工業大学紀要研究編、pp. 159–166, Vol. 53, 2019.

3）廣瀬、大規模授業支援テスティングシステムとそのラーニングアナリティクス、統計数理、Vol. 66, No. 1 , pp. 79–96, 2018.

4）廣瀬、ラーニングアナリティクス指向学習支援、コンピュータ＆エデュケーション（CIEC）、Vol. 45, pp. 23–30, 2018.

5）Hideo Hirose, Meticulous Learning Follow-up Systems for Undergraduate Students Using the Online Item Response Theory, 5th International Conference on Learning Technologies and Learning Environments (LTLE2016), pp. 427–432, 2016.

6）Hideo Hirose, Difference Between Successful and Failed Students Learned from Analytics of Weekly Learning Check Testing, Information Engineering Express, Vol 4, No 1, pp. 11–21, 2018.

7）Hideo Hirose, Prediction of Success or Failure for Examination using Nearest Neighbor Method to the Trend of Weekly Online Testing, International Journal of Learning Technologies and Learning Environments (IJLTLE), Vol 2, No 1, pp. 19–34, May 31, 2019.

8）Hideo Hirose, Key Factor Not to Drop Out is to Attend Lectures, Information Engineering Express, Vol 5, No 1, pp. 11–21, May 31, 2019.