広島工業大学紀要研究編
第53巻（2019）159–166
論　　文

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

廣瀬　英雄*

（平成30年8月6日受付）

Learning Analytics Obtained from the Large-scale Online Testing

Hideo HIROSE

(Received Aug. 6, 2018)

Abstract

The FP (follow-up program) system, a large-scale online testing system, is principally an assistance system for learning students, where the FP system was aimed at helping students who need basic learning and aimed at assisting teachers who have to engage in teaching a variety of educational students. However, all the teachers and stuffs do not always know such a system deeply. The author had a chance to explain the system in public at the research conference held in June 22, 2018, and thus he used this opportunity to do this. This paper describes the details of the lecture in June 22, 2018. Firstly, he mentions the item response theory comparing to the classical testing theory. Then, he describes one of the methods for detecting drop-out students, as a learning analytics.

Key Words: online testing, learning check testing, item response theory, drop-out

1　はじめに

2016年度から始まったフォローアッププログラム（follow-up program, FP）も3年半が経った。FPは，授業時間でのオンラインテスト（LCT, learning check testing），フォローアップクラス（FPC）でのCWT（collaborative work testing），FPT（follow-up program testing）の3つのオンラインテスティングから構成されているユニークなラーニングシステムである。しかし，FPシステムは他大学でも類をみない全学生を対象にしている大規模ラーニングシステムであっても，基本的には学生支援プログラムであるため，運用や評価法の詳細についての本質的な側面が学内でもあまり知られていないように思われる。そこで，ここでは，最近経営システム学会主催の研究会で招待講演「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の方向性」[1)] のテーマで実施された内容に沿ってFPを解説する。講演では，1）項目反応理論（IRT, item response theory）に基づく成績評価，2）FPにおけるオンラインテスティング，3）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，の3部に分けて説明がなされたが，このうち，2）については他の文献[20)] で紹介されているのでそちらを参照されることとして，ここでは，1）と3）について概説する。3）についても多方面からの視点があるが[21,22)]，ここでは，オンラインテスティングを使ってドロップアウトリスクを抱えた学生を早期に発見しアラートを発するような一つの方法について焦点を絞って説明する。

2　古典的テスト理論と現代テスト理論

2.1　古典的テスト理論

古典的テスト理論では，各得点は，受験者の潜在的な能力θと受験時のゆらぎTとの和で表現される。ここで，θ

* 広島工業大学環境学部建築デザイン学科

古典的テスト

問1	8	9	5	7	10
問2	6	10	4	4	6
問3	5	9	7	5	7
問4	5	9	5	4	8
問5	10	9	2	10	10
受験者	A	B	C	D	E
得点	34	46	23	30	40

配点は各問10点

平均=34.6，標準偏差=8.88

正規分布フィッティング

	標準得点	偏差値
C	-1.31	36.9
E	0.61	56.1

図 1　古典的テスト理論での評価法

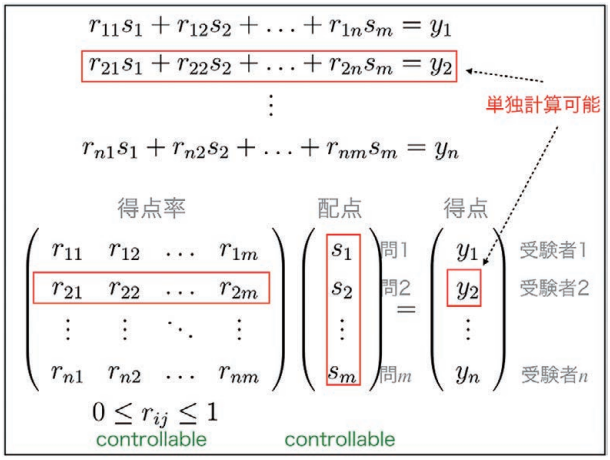

図 2　古典的テスト理論での配点と得点と得点

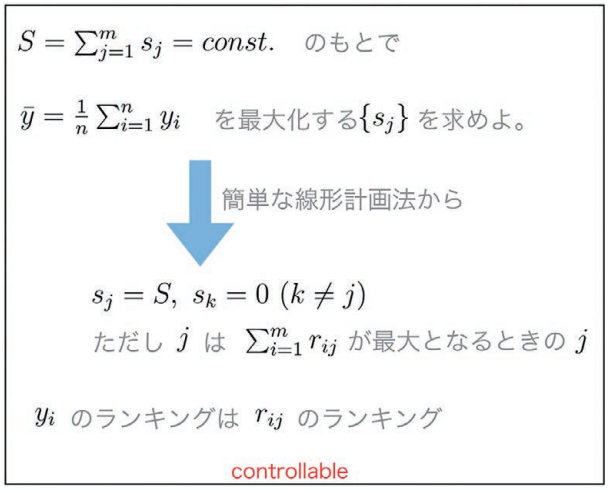

図 3　古典的テスト理論での配点の調整

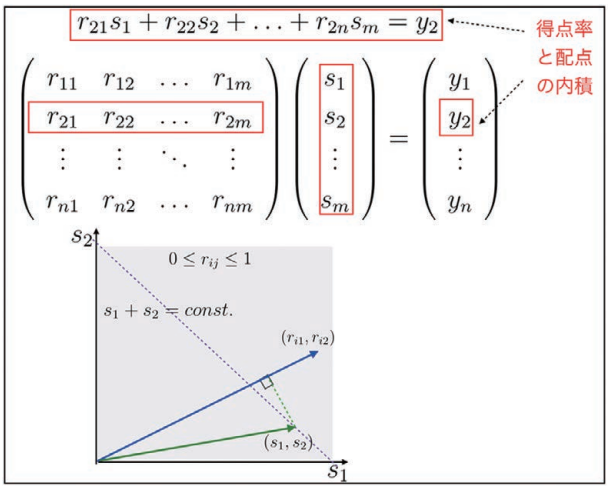

図 4　古典的テスト理論での配点と得点率の内積

は受験集団が持つ分布の確率変数を，T は，普通，正規分布を背後に持つ確率変数を表す。

しかし，通常は，図 1 にみられるような表現がよく用いられ，潜在能力とゆらぎが区別されず一緒のものとして取り扱われることが多い。つまり，ある母集団の中での能力値のみが正規分布で表されると考えられているように思われる。例えば，図では，受験者 E は，母集団の中でたまたま40点をとり，母集団の中での位置は偏差値56.1として取り扱われている確率変数の一つの実現値を表したものであるとの解釈である。

これは，古典的テスト理論での表現法とは本質的に異なっていることに注意されたい。ただ，上のように解釈されることも多いので，ここではこのような解釈で議論を進める。そのような背景で，普通のテストでよく用いられる配点と得点について考えてみる。

図 2 に，古典的テスト理論での配点と得点率および得点の関係を行列の計算の形で表してみた。もし，配点が事前に与えられおり，受験生の個々の問題への得点率が得られれば，受験生個々に採点が可能である。ここで，配点を変えると得点も変わってくるという意味で，図には controllable という書き方で示した。広い意味では，一般に記述式の採点の場合を考えると，得点率も controllable ではあるが，ここでは配点のみが controllable であると考えよう。

配点をうまく変えれば，平均点をあげることができたり，ある特定の問題を厚遇できたりして，結果的には受験者のランキングが変わることにつながってしまう。これは，極端な場合になるが，平均点をあげるには，得点率の最も高い問題にだけ配点して他は配点を与えないようにすることで達成される（図 3）。つまり，一番易しい問題だけに配点するという方法である。このとき，受験者のランキングはその問題の出来だけで決まってしまう。

このことは，図 4 に示すように，配点ベクトルと得点率ベクトルの内積が得点になっていることから当然のことといえる。図では 2 問の場合を例示しているが，そこでは，2 問目の得点率が高いので 2 問目だけに配点するとベクトルの内積の大きさが最大になることが示されている。

このことは，配点は試験前に明らかにされていれば採点前，あるいは明らかにされていなければ採点後でも，得点

の調整を採点者の意思によって変えることが可能になっていることを示している。次に述べる項目反応理論では，一般的にはそのような余地が入らず，ある意味で公平な評価法になっていることを示そう。

2.2　現代テスト理論

項目反応理論（IRT, item response theory）では，ある受験者 i が問題 j に取り組んだときに反応した解答（正答なら1，誤答なら0）のマトリクス全体で受験生の評価を行う。マトリクスの例を図5に示す。また，2値の応答に対する確率分布として図6に示すようなロジスティック分布を仮定している。

応答マトリクスが与えられた時，最尤推定法によって受験生の能力（習熟度）θ と問題の困難度（a, b）を同時に求めるのが項目反応理論である（図7）。直接的に尤度を最大にすると推定での計算過程に不安定性が起こるので，一般に，周辺化を行って θ の確率変数を消去した後に困難度パラメータを求めることを行っている。その際，ベイズ推定，EM アルゴリズム，MCMC（マルコフチェーンモンテカルロ）法などの計算法が使われている。

応答マトリクスはすべての要素に数値が入っている完全マトリクスで対応するのが普通であるが，場合によっては図8に示すようにマトリクスが不完全になることがある。この場合にも対応できるような方法[19]が示されているので，FP のシステムではこの方法も取り入れている。

IRT で評価する際，多様な習熟度を持つ学生に対して公平な問題設定をどのように行えばよいのだろうか。情報量の観点からは，受験生の習熟度と問題の困難度が一致した場合に得られる情報量が最大になり，従って推定誤差が最小になることがわかっている。そこで，多様な学生の習熟度を背景に，すべての学生が最適推定に対応できる問題を含むように設定した場合の模式図に示したものが図9である。図10はそのときの予測誤差である。最適でない問題にあたったときの予測誤差の減少はそれほど大きくはないので図9の配置で効率的に予測できる可能性がある。FP システムでの問題設定はこのようなことに配慮して行われている。つまり，問題が7問あった場合，難易度を標準正規分布での z 値の問題を，−3，−2，−1，0，1，2，3のように一問ずつ配置する方法である。受験生の習熟度が−1の学生は問題の困難度が−1の問題を50%の確率で解く

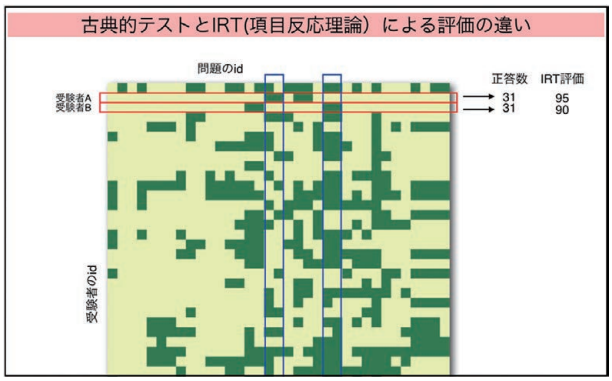

図5　項目反応理論での応答マトリクス

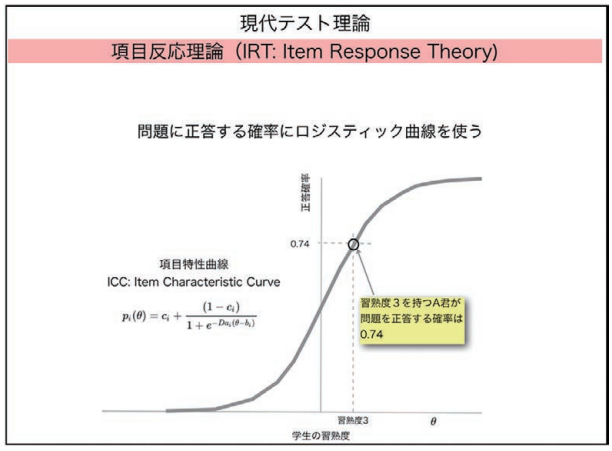

図6　項目反応理論での応答確率分布

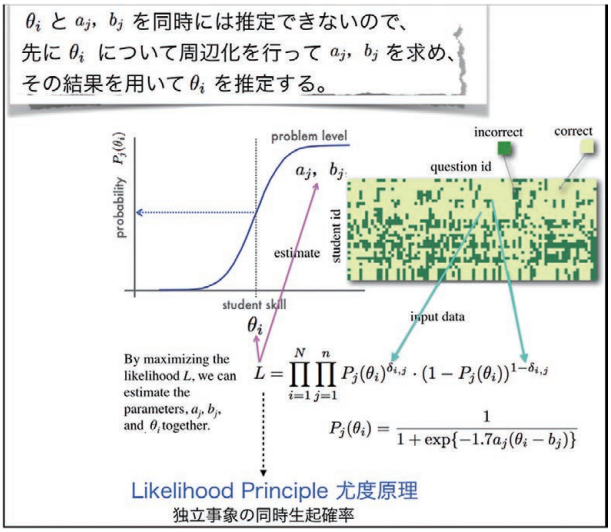

図7　項目反応理論でのパラメータ推定法

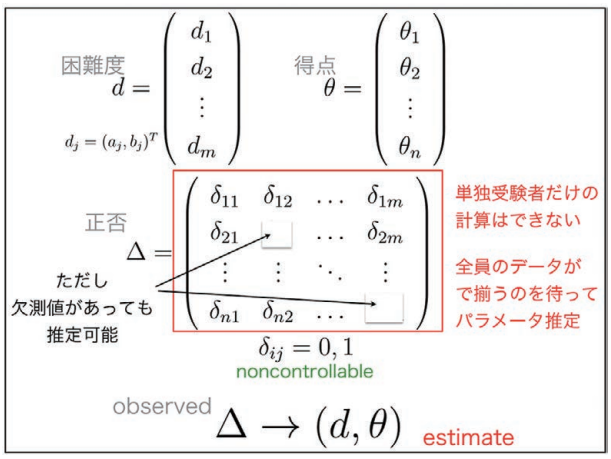

図8　項目反応理論での不完全マトリクスへの対応

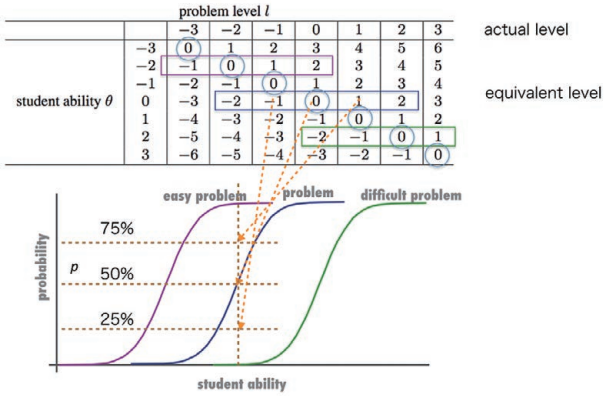

図9　項目反応理論での最適問題設定模式図

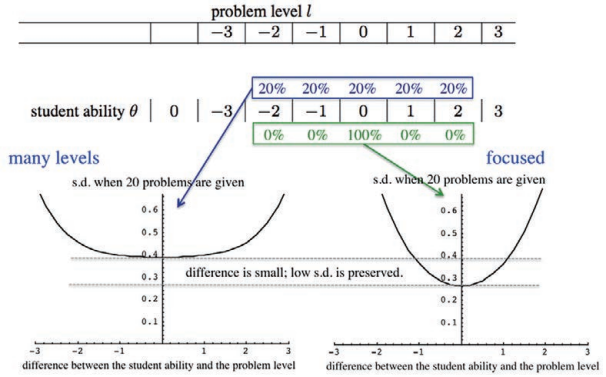

図10　項目反応理論最適問題設定での予測精度

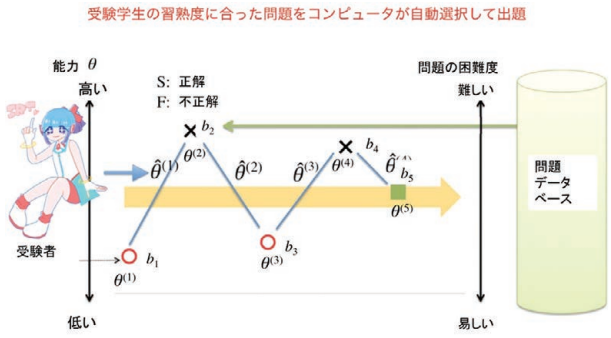

図11　アダプティブテスティングシステム

ことができるというように考えている。

このように，問題の困難度も推定できるようになれば，受験生に最適な問題をシステムが自動的に判断して出題してくれるアダプティブシステム（図11）の設計も可能になってくる。FP システムでは，CWT と FPT にこのアダプティブテスティングを用いている。

項目反応理論を古典的なテスト理論と比較してまとめたものを図12に示す。また，最近の IRT の研究の方向性を図13に示す。

3　ドロップアウトリスク学生の早期発見

ここでは，ラーニングアナリティクスの中でも FP にとって特に重要な，ドロップアウトのリスクを持つ学生を早期に発見して早めにアラートを発するような一つの方法について述べる。

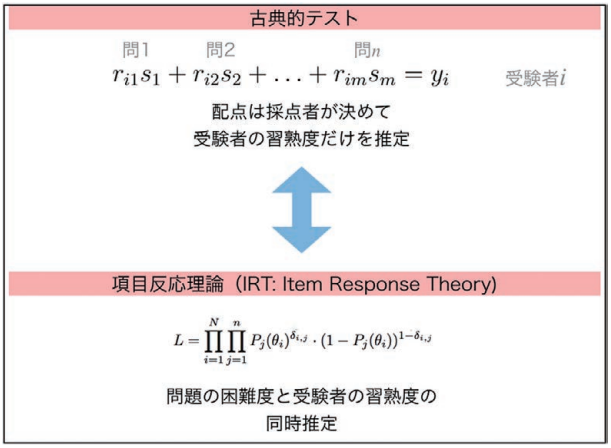

図12　項目反応理論と古典的テスト理論の比較

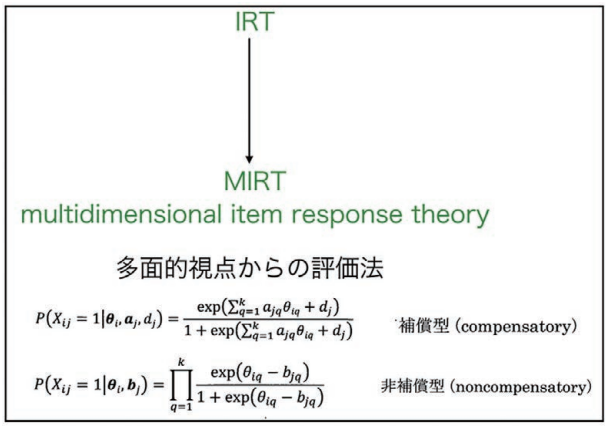

図13　項目反応理論と古典的テスト理論の比較

プレースメントテストを行うと受験生の習熟度の分布がある程度把握できる。この分布そのものがそのまま動かずに継続するという錯覚を持つことが多いと思われるが，実際にはテスト結果は確率的な変動を伴うものであり，一度の受験機会の結果がそのまま継続するものでは決してない。

図14には，潜在的な習熟度に対して，テストによって得られたスコアの分布が複雑にからみあっていることを模式的に示してみた。極端な言い方をすれば，非常に優秀なグループとその対照にあたるグループの変化は少ないと考えられるが，中間層はその大きさが大きいことと変化の幅が大きいことからなかなかとらえどころがない。従って，中間層に入っているドロップアウトのリスクを持つ学生を早期に特定することは極めて困難である。

それは，プレースメントテストや LCT の結果から期末試験の合否を予測する困難さに直結している。例えば，図15は，２種類のプレースメントテストのスコアのヒストグラムを示しているが，この分布のどこを閾値に設定すれば期

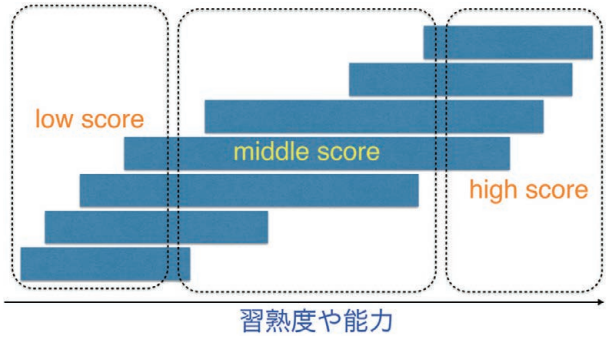

図14　潜在的な習熟度とテストスコアの分布の模式図

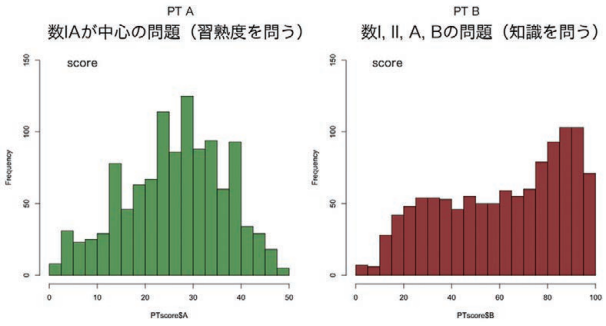

図15　プレースメントテストのスコアのヒストグラム

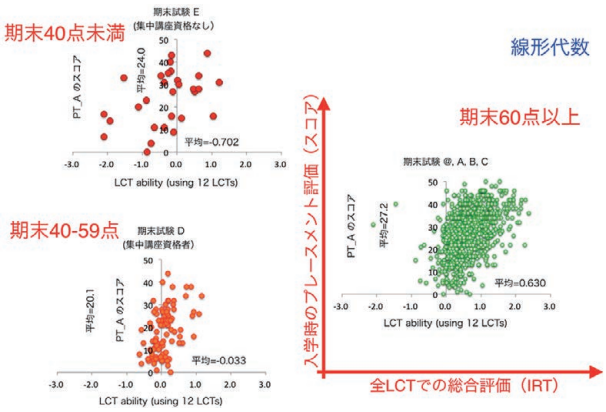

図16　プレースメントテストに LCT の結果を加えた散布図

末試験の合否を誤差最小にして予測できるかは難しい問題になっている。

図16では，プレースメントテストに LCT の結果を加えても，期末試験の結果が60点以上になる学生と59点以下になる学生を識別することは難しいことが示されている。ここで，LCT の結果は，全授業で実施された LCT の累積結果を用いている。図17に示すように，期末試験合格と不合格の2グループの LCT の習熟度のヒストグラムを見ると，合格グループに中に不合格グループがすっぽりと埋め込まれているように見える。

そこで，まず，LCT の合格回数と FPC の欠席回数の頻度分布と期末試験の合否の間に対応関係がないか調べてみた。図18にそれを示す。図から，LCT の合格回数が10回以上の学生は期末試験に失敗することはないように思われる。逆

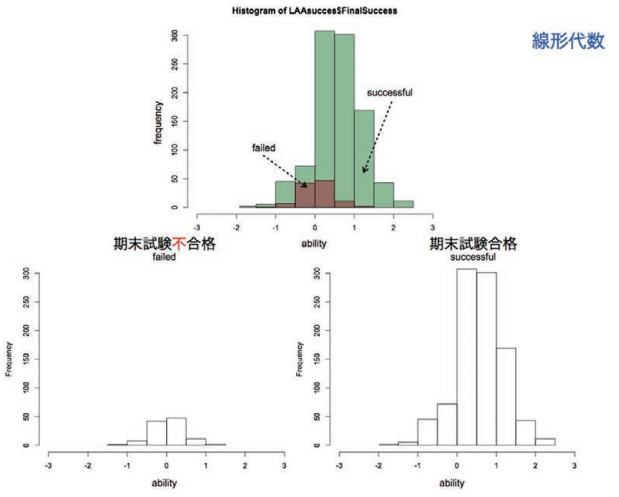

図17　期末試験合格と不合格の2グループの LCT の習熟度のヒストグラム

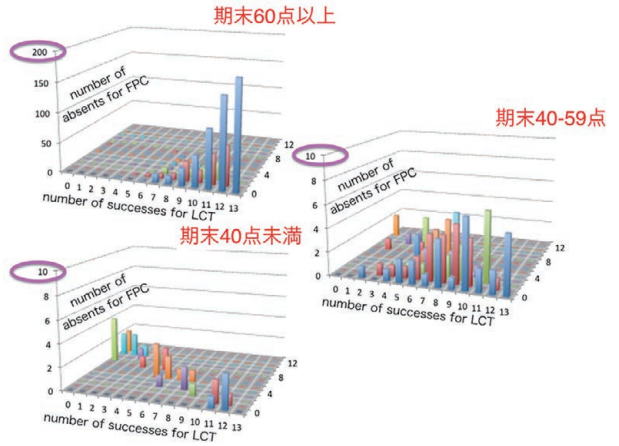

図18　LCT の合格回数と FPC の欠席回数の頻度分布

の言い方をすると，LCT の合格回数が9回以下の学生にはアラートが有効かもしれない。

上の結果は，学期が終了してはじめてわかるので，期末試験の合否予測には間に合わない。そこで，次に，LCT の ability のトレンドを毎授業で調べることで予測につながる結果が得られないかと考えてみた。

図19は，毎授業で得られた問題から推定された各学生の各単元に対する ability 値のトレンドをみたものである。ゆらぎが大きすぎてトレンドがつかめない。

そこで，当該授業までに得られた LCT の応答すべてを使った応答マトリクスの累積値から ability 値を求めてそのトレンドをみてみた。図20，21は，期末試験合格者の当該 LCT 単元までの LCT 累積値から推定した ability 値のトレンド，期末試験不合格者の当該 LCT 単元までの LCT 累積値から推定した ability 値のトレンドを表している。

両図には明らかな違いがありそうなので，期末試験に失敗するリスクを抱えた学生を早期に特定するための基準を最近傍（Nearest Neighbor）で求めてみた。図22に，トレ

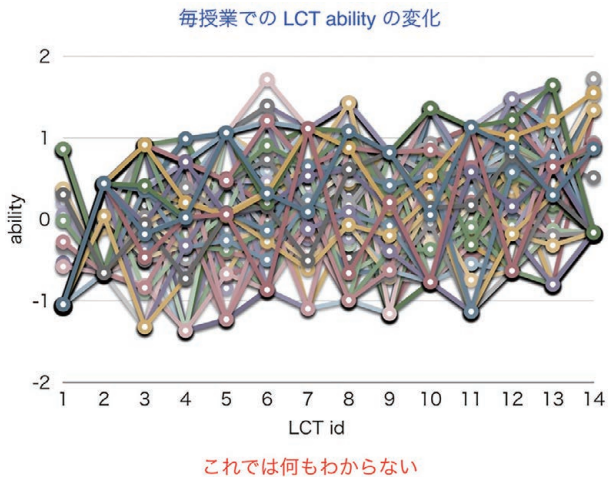

図19　LCT 単元に対する LCT の ability 値のトレンド

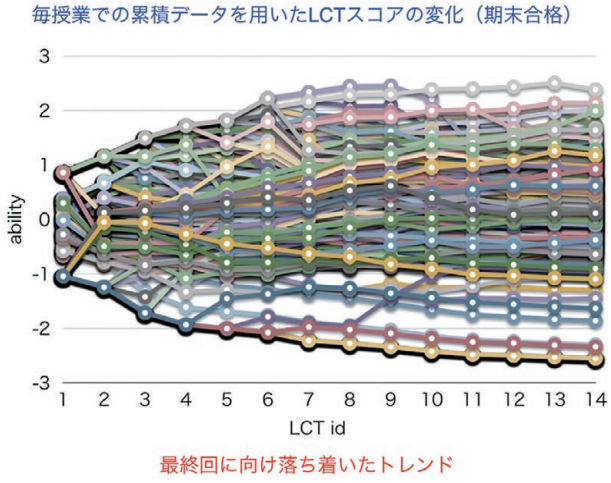

図20　当該 LCT 単元までの LCT 累積値から推定した ability 値のトレンド（期末試験合格者）

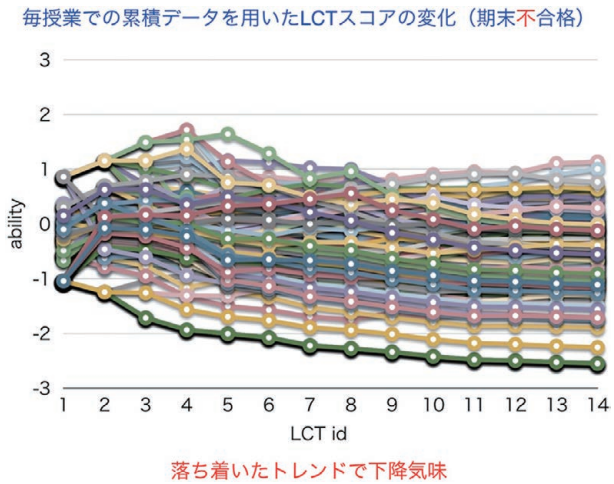

図21　当該 LCT 単元までの LCT 累積値から推定した ability 値のトレンド（期末試験不合格者）

ンド間の距離を用いて類似度を求める考え方を図示した。図23には，距離を求める数式を示した。

この方法を用いると，ある学生に最も似た学生を10人集

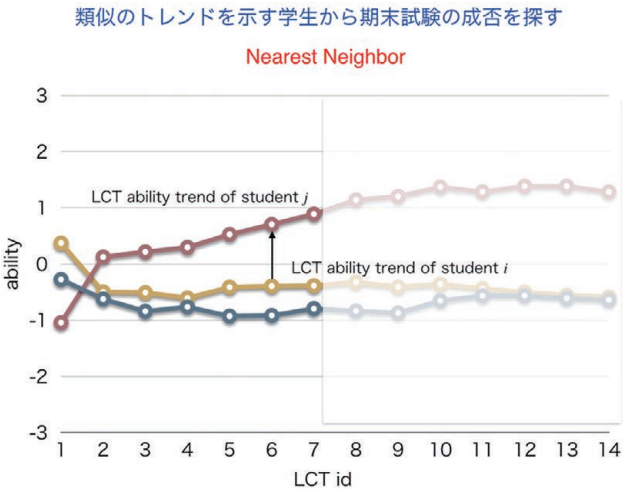

図22　トレンド間の距離を用いた類似度

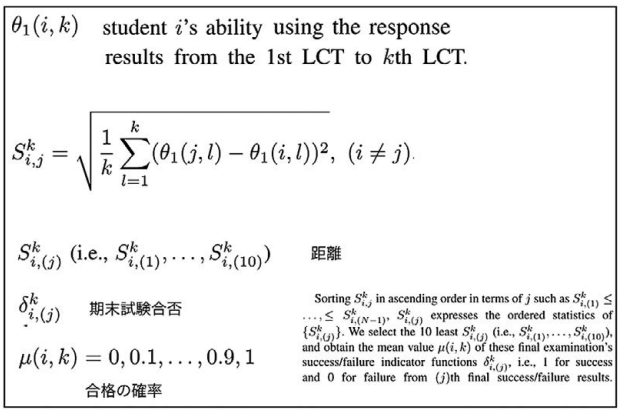

図23　距離を用いた類似度の求め方

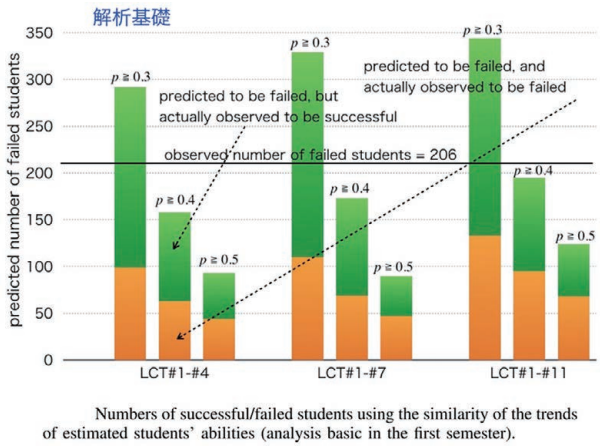

図24　期末試験の合格者予測数と実際の合格者と不合格

めてそれらの学生の動向（期末試験の合否）から当該学生の期末試験の合否確率を求めることができる。この場合，確率は0，0.1，0.2，…，1.0の11パターンが得られる。そこで，リスク管理に重要になりそうな確率として不合格率0.3から0.5の間について予測してみた結果が図24である。図では，LCT の１回目から４回目まで，７回目まで，11回

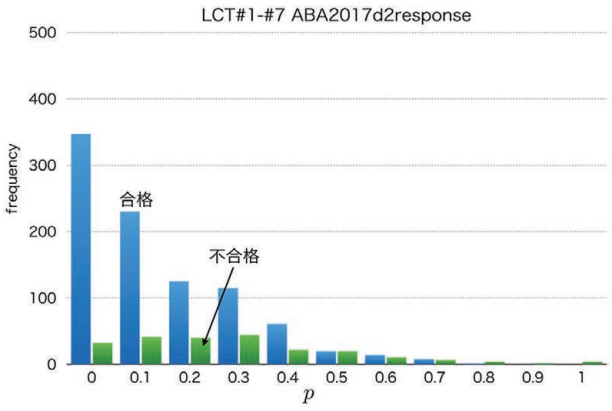

図25　LCT 受験状況と FPT 受験状況についての関係

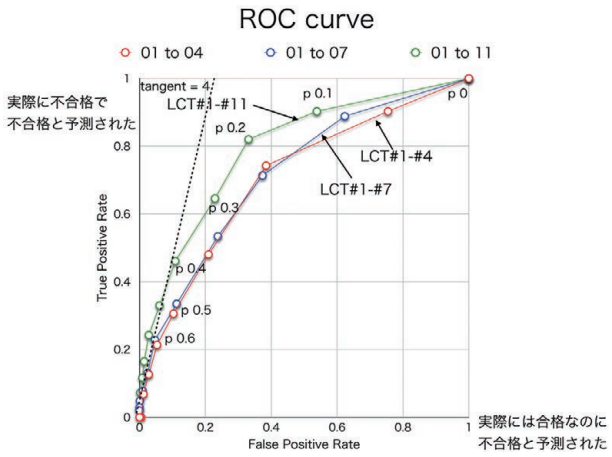

図26　ROC 曲線

目までの 3 ケースを示している。例えば，LCT の 1 回目から11回目までの結果を使った場合，期末試験の不合格率が0.4以上になる人数を約200人と予測しており，実際にその半数の100人が期末試験に失敗していることがわかる。

図25には，LCT の 1 回目から 7 回目までの結果を使った場合の，期末試験の各不合格率に対する実際の合格者数と不合格者数を示している。不合格率の設定を低くすれば不合格者を多く捕捉することができるが，一方で，合格者も捉えてしまう。両者はトレードオフの関係になっている。

図26は，2 値分類によく用いられる ROC 曲線を表したものである。最適な合格率の設定値を求めたいのであるが，ここでは合格者ではなく，不合格者に興味があるのでコストにはそちらに多くのウエイトを置いた。ここでは，不合格と合格に対するウエイトの比を 4 ： 1 とした。そのときの最適な合格率の設定値は0.4となった。

図27は，Recall Precision 曲線である。これがトレードオフの関係を表していることになる。

これまで，オンラインテスティングの結果から得られたデータからラーニングアナリティクスを行った結果，大まかに図28にまとめられるようなことがわかった。

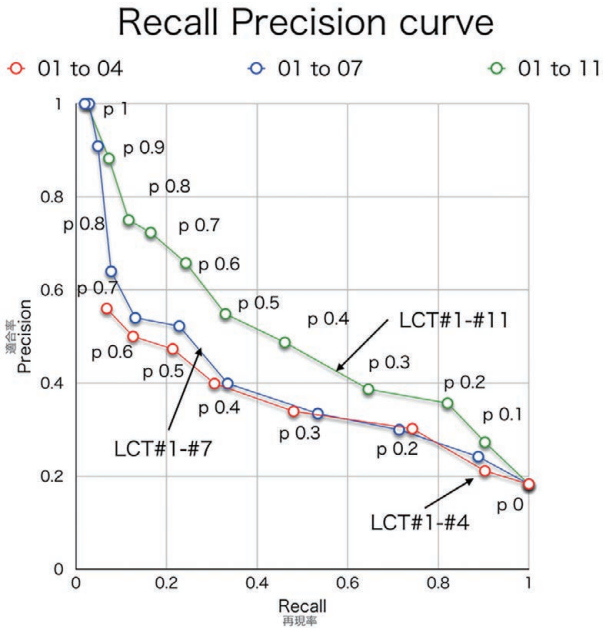

図27　Recall Precision 曲線

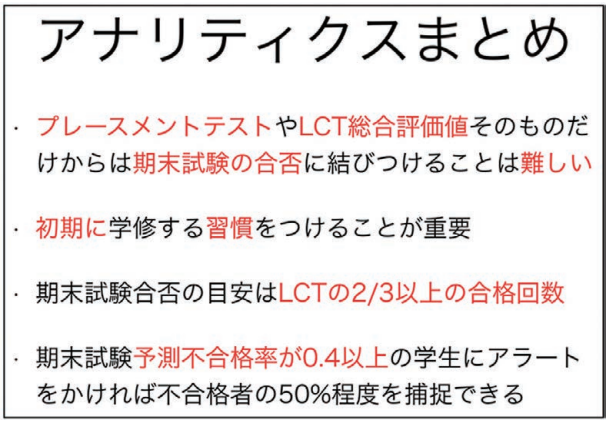

図28　ラーニングアナリティクスから得られた要点

4　まとめ

ここでは，項目反応理論（IRT, item response theory）に基づく成績評価について，古典的テスト理論との比較を行った後，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のうち，特に，ドロップアウトのリスクを抱えた学生の早期発見法の一つを紹介した。それは，当該授業までに蓄積された LCT の結果をまとめて評価して，ability 値の各単元に対するトレンドを求めた上で，トレンド間の類似度に最近傍を用いて，リスク学生の早期予測を行うものである。この方法によれば，リスクを抱えた学生の半数は予測によって捕捉することが可能であることがわかった。また，LCT の不合格回数の割合が2/3を下回ると，不合格リスクが高いことがわかった。

文　献

1）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の方向性，日本システム経営学会イノベーション試行データ分析研究会招待講演，(2018.6.22)

2）廣瀬，ラーニングアナリティクス：フォローアップ演習（CWT）の場合，広島工業大学紀要教育編，pp. 149–155, Vol. 51, 2017.

3）廣瀬，新入生全員を対象としたオンラインテストの実際，広島工業大学紀要教育編，pp. 27–35, Vol. 16, 2017.

4）廣瀬，フォローアップクラスにおける授業設計について，広島工業大学紀要教育編，pp. 37–41, Vol. 16, 2017.

5）廣瀬，LCT（習熟度確認テスト）と FPT（フォローアップテスト）の受験状況と期末試験の関係，広島工業大学紀要研究編，pp. 93–101, Vol. 52, 2018.

6）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.

7）廣瀬，多様な学生集団から固有集団を早期に分類する方法について，広島工業大学紀要教育編，pp. 131–135, Vol. 51, 2017.

8）廣瀬，ラーニングアナリティクス：授業確認テスト（LCT）の場合，広島工業大学紀要教育編，pp. 137–147, Vol. 51, 2017.

9）廣瀬，ラーニングアナリティクス：フォローアップ演習（CWT）の場合，広島工業大学紀要教育編，pp. 149–155, Vol. 51, 2017.

10）廣瀬，新入生全員を対象としたオンラインテストの実際，広島工業大学紀要教育編，pp. 27–35, Vol. 16, 2017.

11）廣瀬，フォローアップクラスにおける授業設計について，広島工業大学紀要教育編，pp. 37–41, Vol. 16, 2017.

12）廣瀬，フォローアップクラス参加による学習効果の確認法について，広島工業大学紀要教育編，pp. 43–47, Vol. 16, 2017.

13）廣瀬，フォローアッププログラムにおけるオンラインテストの学生の受け止め方，広島工業大学紀要教育編，pp. 49–53, Vol. 16, 2017.

14）廣瀬，ラーニングアナリティクス：授業確認テストとフォローアップ確認テストの受験トレンド，広島工業大学紀要教育編，pp. 55–60, Vol. 16, 2017.

15）廣瀬，アダプティブテスティングにおける IRT 困難度の推定：LCT の結果を用いた支援推定法，広島工業大学紀要研究編，pp. 103–108, Vol. 52, 2018.

16）廣瀬，ラーニングアナリティクス：LCT と FPT の受験状況トレンド2017 vs 2016，広島工業大学紀要教育編，pp. 65–70, Vol. 17, 2018.

17）廣瀬，テスト問題の配点と得点調整に関する一考察：項目反応理論との比較，広島工業大学紀要教育編，pp. 71–77, Vol. 17, 2018.

18）廣瀬，LCT（習熟度確認テスト）と FPT（フォローアップテスト）の受験状況と期末試験の関係，広島工業大学紀要研究編，pp. 93–101, Vol. 52, 2018.

19）作村，徳永，廣瀬，EM タイプ IRT による不完全マトリクスの完全化とその応用，情報処理学会論文誌，数理モデル化と応用 Vol. 7, No. 2, pp. 17–26, 2014.

20）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.

21）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.

22）Hideo Hirose, Success/Failure Prediction for Final Examination using the Trend of Online Testing Result, 7th International Conference on Learning TecS̆ologies and Learning Environments (LTLE2018), 2018. to appear.