広島工業大学紀要研究編
第54巻（2020）141-147
論　　文

LCTの評価結果を用いたCBTによる期末試験の合否予測

廣瀬　英雄*

（令和元年8月18日受付）

Success/Failure Prediction for CBT Based End-term Examination using LCT Evaluation

Hideo HIROSE

（Received Aug. 18, 2019）

Abstract

Using the LCT (learning check testing) results in the follow-up (FP) program system, we have shown that we can predict the students' future results of success/failure to the end-term examination with a probability to some extent. However, we also anticipate that we may not be able to enhance a certain level of prediction accuracy for evaluation because the description type questions and their evaluations to the final examinations depend on two or more teachers, resulting unfair scores although the LCT provides the multiple-choice questions and the evaluations are performed automatically using computers. However, we have had a chance to perform such a CBT using the multiple-choice questions and the automatic evaluations using item response theory. Here, in this paper, we have investigated whether the accuracy for success/failure prediction to the end-term examination using CBT can be improved than that using the description-type testing. We have found that the CBT based on the end-term examination using the LCT improves the prediction accuracy for success/failure than that based on the description-type testing using the LCT.

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

1　はじめに

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には、毎授業で全学生に実施されるLCT（learning check testing）の蓄積結果から項目反応理論によってabilityの推定値を求め、そのトレンドが学生毎でどう似ているかをsimilarityによって測れば、リスクのある学生を早期に発見することが可能であることを示した。また、その方法として最近傍（Nearest Neighbor）法が適していることを示した[1-10]。この方法によれば、期末試験の正否を、すべてのLCTを一括して用いて求めたabilityから正否を分ける最適なしきい値を求めて正否を予測する方法よりも優れていることがわかっている。

しかしながら、最近傍法を使っても、期末試験の正否を誤って判断する誤分類率は20%以下とかなり低くなってはいるものの、学期半ばでの「このまま何も手を打たなければ期末試験に失敗する確率は0.4以上である」という予測した該当者に対して、それが実際に的中するのは約40-50%程度となり、アラートを出した学生の半数は実際に期末試験に合格している結果となっている。つまり、誤分類

* 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

率は低くても、的中率は高くない、という指摘がなされている[11]。履修者1000人中150人が不合格と仮定したとき300人にアラートを出しておけば不合格者への的中率は上がるものの、1000人中300人へのアラートは、３人に１人という割合となるため、予測精度をもっと上げたい。

ところが、予測精度を上げるには限界がある。LCT は、全学生に同じ問題を出してコンピュータが解答の正誤を判定するため公正で公平な評価が行なわれていると考えられるが、期末試験は一般に記述式の試験で実施され、クラス毎に全く同じ問題が出される訳ではなく、また、採点はクラス担当の教員が個別に行なうため、クラス間で評価にバイアスが発生する可能性があると考えられるからである。そこで、期末試験そのものにも公正で公平な評価を求めるなら、LCT と同じように、期末試験も記述式ではなくコンピュータによって解答の正誤が判定されるような仕組みが必要になると思われる。

本論文では、期末試験を、クラス毎の記述式問題をクラス教員が評価した場合と、すべてをコンピュータによって人の判断が入る余地がないようにした場合の比較を行なった。その結果、興味あることがわかった。学科や学部の履修者が多い場合で、複数の教員が分担して教えて評価する場合に、本論文が、公平性をどう保てばよいのかというヒントになると思われる。

ここでは、学生の習熟度を測定するテストに、１）入学直後のプレースメントテスト、２）LCT、３）期末試験の３つを採用し、それらの評価結果としては、項目反応理論を用いて得られた ability 値を使った。プレースメントテストでは、一部、スコアそのものも用いている。

2　記述式の期末試験をクラス担当教員が評価する場合

似たような性格を持つ学科の学生を一旦一同に集め、それを複数の教員が分担して授業を実施する。その際、学生の習熟度が大きく異なる場合には、クラスを習熟度別に分けて授業を行なった方が好ましいという考え方は自然であり、広く行なわれている。例えば、150人程度の学生の習熟度を何らかの方法を用いてあらかじめ測っておき、その結果から、50人ずつの上位、中位、下位クラスのように３クラスに分けることが行なわれている。

広島工大でも基本的には習熟度によってクラス分割を行なっているが、数学の授業では下位クラスを作らないこととしている。その理由は、下位クラスで編成されるクラスの雰囲気があまりよくないという経験を持っているからである。つまり、中位クラスと下位クラスを一緒にすると、習熟度がある程度高い学生が低い学生を指導してともに学び合う姿があちこちで見られ、これが功を奏しているように見受けられることがあるからである。

図１に示す例（2017年度の線形代数 A）では、180人程度の集団を、入学直後に実施されたプレースメントテスト結果を用いておよそ40-50人程度の４クラスに分けている。プレースメントテストの成績が上位の学生を２クラスの間で均質になるように、残りの２クラスでは下位の学生が均質になるようにしている。

図１の右上の box-plot を見るとわかるように、プレースメントテスト結果におけるあるしきい値を境に、右２つが上位クラス、左２つが下位クラスに配置されている。

２段目には13回の LCT の結果をすべて用いた ability 値を示した。下位クラスは全体の中で下位傾向に、上位クラスは上位傾向に位置しているものの、クラスの LCT による習熟度は均質には見えない。LCT は同じテストを同じ条件で行なっているはずであるが、どういう訳か均質性が崩れてしまっている。LCT の実施状況や実施環境、あるいは教え方による効果などが異なったのかもしれない。しかし、わずかではあるが、プレースメントテストの結果との相関はある。

３段目は期末試験の５段階成績であり、90-100、80-89、70-79、60-69、0-59のスコアを、S、A、B、C、D の棒グラフで示している（右側に位置する程成績が良い）。プレースメントテストの結果と比較してみると、ここではほとんど相関がない。極端なケースと思われるが、下位クラスの右側（左から二番目）よりも上位クラスの右側（一番右側）の方が期末試験の成績は悪くなっている。このことは、LCT の成績と比較して考え合わせると、クラスの教員による評価のバイアスが発生していると考えざるを

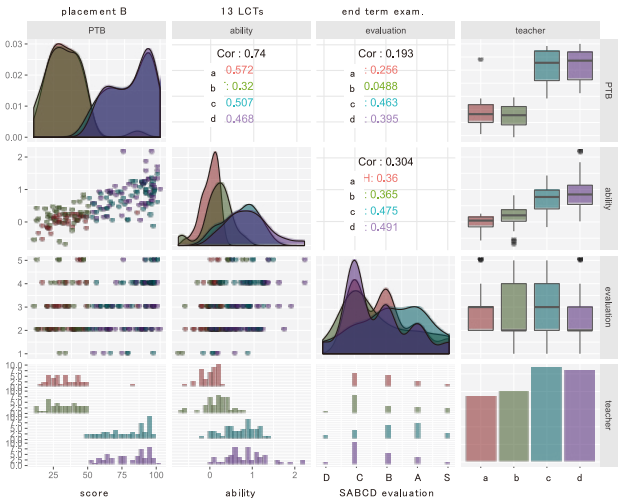


図１　プレースメントテスト結果から上位２クラス下位２クラスに分割したクラスでの、プレースメントテスト、LCT、期末試験の結果の関係
（記述式期末試験評価をクラス担当教員が行なう場合）

得ない。つまり、上位クラスを受け持った d 教員はかなり厳しい規準で評価しており、下位クラスを受け持った b 教員は全体にわたって成績が分布するような評価を行っている。

このように、プレースメントテストや LCT の結果が期末試験の結果と相関をあまり持たないようになってくると、LCT トレンドの similarity がいくら良い予測精度精度を持っていても、これを使って期末試験の正否の予測を行うことは困難になってくることが予想される。これが、クラス間で評価にバイアスが発生する可能性があると考えられる実際のケースである。

3　CBT による期末試験を用いた場合

そこで、プレースメントテストや LCT の評価と同じように、教員によるバイアスが発生しないように、期末試験の評価をコンピュータが行なう CBT 環境を作り、三者の関係を比較してみた。210人程度のクラスを、プレースメントテストの成績が均質な習熟度を持つように 3 つのクラスに分割した。CBT では、問題はすべて多肢選択型とし、回答はマークシートを使った。オンラインによるテスティングも可能な環境が作られてはいるが、WiFi の不安定な要素を考えると、公平性の観点から踏み切るにはまだ危険性を帯びているからである。ここでは、中間試験、期末試験の結果を同時に両方使って項目反応理論による評価結果（ability 値）を成績としている。

図 2 の右上の box-plot や左下のヒストグラムに示すように、プレースメントテストの成績は 3 クラスでまったく同じ分布をしている。12回すべての LCT を用いて求めた ability の結果にはクラス間で少しの違いが見られるものの、全体的には 3 クラス間でほとんど同じような分布をしていることが観測できる。また、プレースメントテストの成績と LCT の ability 値との相関はかなり高い。

更に、マークシートを用いた多肢選択型の CBT による期末試験の結果も 3 クラス間でほぼ同様で均質な分布になっている。期末試験の成績は、プレースメントテストの成績、および LCT の結果両方で高い相関を示していることがわかる。したがって、こういう場合、プレースメントテストの成績や LCT の結果を説明変数に用いれば期末試験の正否の予測は高い精度で得られることが期待される。

参考までに、プレースメントテスト A と B のタイプ、中間試験単独、期末試験単独、両者を合わせた結果、LCT ポイントなどの詳細な情報が記載された相関図を付図 1 に示す。

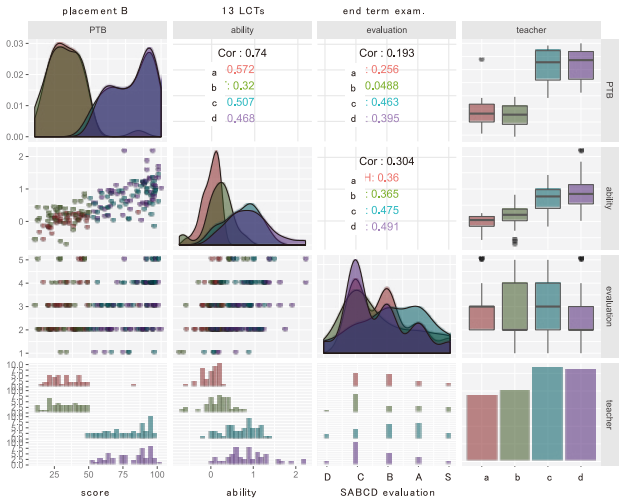


図 2　プレースメントテスト結果を均質な 3 クラスに分割したときの、プレースメントテスト、LCT、期末試験の結果の関係（期末試験評価を CBT で行なう場合）

4　LCT トレンドを用いた CBT 期末試験の合否判定予測

前節で、期末試験を教員の評価バイアスが入る余地のない多肢選択型の CBT で行なった場合、LCT の結果から期末試験の合否予測を行なったときの予測精度が向上することが期待されることを示唆した。そこで、ここでは、受験者数は200人程度と多くはないが、この傾向が見られるかどうかについて調べてみたい。データは図 2 に示したものを用いる。

期末試験の合否判定には ability のしきい値を決めて、それ以下は不合格、それ以上は合格としている。ここでは、そのしきい値を変動させながら不合格者数を、210人中20人の場合、30人の場合、40人の場合、50人の場合、と 4 つのケースに分けたときの合否判定予測結果について議論してみる。

表 1－8 に、2019年度のある学部での解析基礎 A で、LCT トレンドの similarity を用いた期末試験の合否予測を行なった confusion matrix、誤分類率、不合格の的中率について、LCT#1-#7を予測データに用いた場合、LCT#1-#11を予測データに用いた場合において、不合格率 $p \geqq 0.3$ の場合、$p \geqq 0.4$、$p \geqq 0.5$についての予測結果を示す。また、特に、予測の的中の度合いを表す図を、図 3－6 に示す。

表 1　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中20人）
LCT#1-#7利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	165	24	189
観測結果	失敗	4	16	20
	合計	169	40	209

誤分類率：0.134、不合格の的中率：40.0%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	173	16	189
観測結果	失敗	11	9	20
	合計	184	25	209

誤分類率：0.129、不合格の的中率：36.0%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	182	7	189
観測結果	失敗	17	3	20
	合計	199	10	209

誤分類率：0.115、不合格の的中率：30.0%

表 2　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中20人）
LCT#1-#11利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	170	19	189
観測結果	失敗	11	9	20
	合計	181	28	209

誤分類率：0.144、不合格の的中率：32.1%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	179	10	189
観測結果	失敗	12	8	20
	合計	191	18	209

誤分類率：0.105、不合格の的中率：44.4%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	180	9	189
観測結果	失敗	15	5	20
	合計	195	14	209

誤分類率：0.115、不合格の的中率：35.7%

表 3　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中30人）
LCT#1-#7利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	143	36	179
観測結果	失敗	6	24	30
	合計	149	60	209

誤分類率：0.201、不合格の的中率：40.0%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	151	28	179
観測結果	失敗	11	19	30
	合計	152	47	209

誤分類率：0.187、不合格の的中率：40.4%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	165	14	179
観測結果	失敗	15	15	30
	合計	180	29	209

誤分類率：0.139、不合格の的中率：51.7%

表 4　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中30人）
LCT#1-#11利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	144	35	179
観測結果	失敗	13	17	30
	合計	157	62	209

誤分類率：0.230、不合格の的中率：32.7%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	160	19	179
観測結果	失敗	15	15	30
	合計	175	34	209

誤分類率：0.163、不合格の的中率：44.4%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	169	10	179
観測結果	失敗	17	13	30
	合計	186	23	209

誤分類率：0.129、不合格の的中率：56.5%

表 5　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中40人）
LCT#1-#7利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	132	37	169
観測結果	失敗	7	33	40
	合計	139	70	209

誤分類率：0.211、不合格の的中率：41.7%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	137	32	169
観測結果	失敗	13	27	40
	合計	150	59	209

誤分類率：0.215、不合格の的中率：45.8%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	151	18	169
観測結果	失敗	16	24	40
	合計	167	42	209

誤分類率：0.163、不合格の的中率：57.1%

表 6　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中40人）
LCT#1-#11利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	131	38	169
観測結果	失敗	6	34	40
	合計	137	72	209

誤分類率：0.211、不合格の的中率：47.2%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	140	29	169
観測結果	失敗	11	29	40
	合計	151	58	209

誤分類率：0.191、不合格の的中率：50.0%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	151	18	169
観測結果	失敗	17	23	40
	合計	168	41	209

誤分類率：0.167、不合格の的中率：56.1%

表 7　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中50人）
LCT#1-#7利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	126	33	159
観測結果	失敗	5	45	50
	合計	131	78	209

誤分類率：0.182、不合格の的中率：57.7%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	132	27	159
観測結果	失敗	8	42	50
	合計	140	69	209

誤分類率：0.167、不合格の的中率：60.9%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	136	23	159
観測結果	失敗	10	40	50
	合計	146	63	209

誤分類率：0.158、不合格の的中率：63.5%

表 8　2019年度 E 学部解析基礎 A の confusion matrix
（不合格者数を、210人中50人）
LCT#1-#11利用の場合

$p \geqq 0.3$			予測結果	
		成功	失敗	合計
	成功	125	34	159
観測結果	失敗	5	45	50
	合計	130	79	209

誤分類率：0.187、不合格の的中率：57.0%

$p \geqq 0.4$			予測結果	
		成功	失敗	合計
	成功	134	25	159
観測結果	失敗	6	44	50
	合計	140	69	209

誤分類率：0.148、不合格の的中率：63.8%

$p \geqq 0.5$			予測結果	
		成功	失敗	合計
	成功	140	19	159
観測結果	失敗	10	40	50
	合計	150	59	209

誤分類率：0.139、不合格の的中率：67.8%

表1-8を見ると、誤分類率はいずれも10-20％程度で低くなっている。「不合格」と予測した学生が実際に不合格になっている的中率については、30％から70％程度にまでばらついており、不合格者数の割合が増えるに従って的中率が上がっていることがわかる。特に、表7（不合格学生50人の場合）では､学期始めから学期の中くらいまで（ちょうど中間試験の頃）のLCTの実績を使えば、$p \geqq 0.5$のとき、期末試験に不合格する学生63人のうち40人が不合格と判定され､そのときの不合格的中率は63.5％となっており、先に報告した期末試験が記述式の場合の的中率をかなり上回っている。また、表7を表8と比較してみてもそれほど差がなく、学期の終わりまでの情報を使う場合と学期の真ん中付近までの情報を使う場合の結果とがほとんど変わらないこと示されている。更に、pのしきい値の取り方についても、先に報告したように、$p \geqq 0.4$のときでも十分に予測機能を果たしていることがわかる。

表1－8のうち、特に、不合格的中の程度を視覚化するために、図3－6を作ってみた。図6を見ると、実際に50人の不合格者に対して45人以上を不合格と予測しており、先に報告した結果[8, 9]と比べて格段に予測精度が向上していることがわかる。

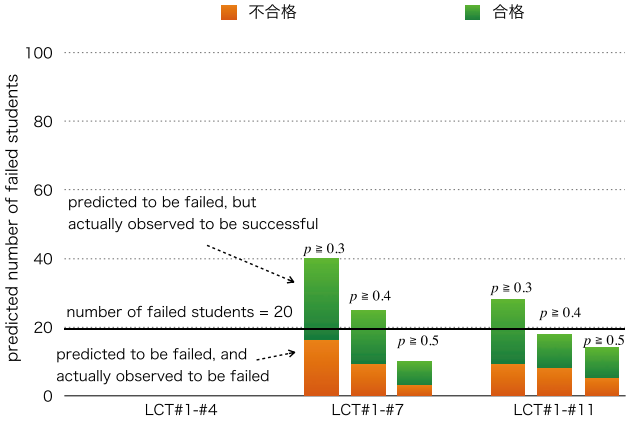


図3　期末試験の合格者予測数と実際の合格者と不合格
（不合格者数 =20人の場合）

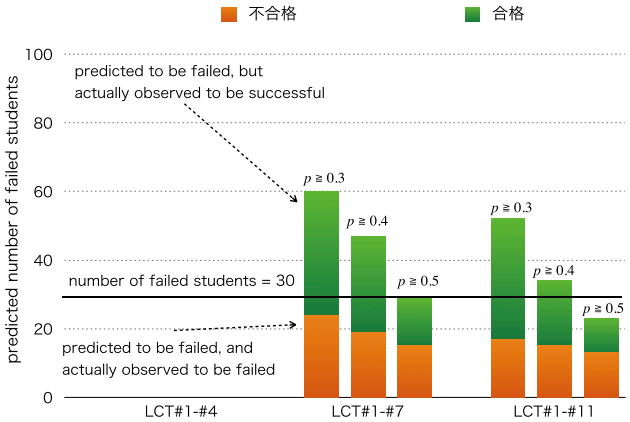


図4　期末試験の合格者予測数と実際の合格者と不合格
（不合格者数 =30人の場合）

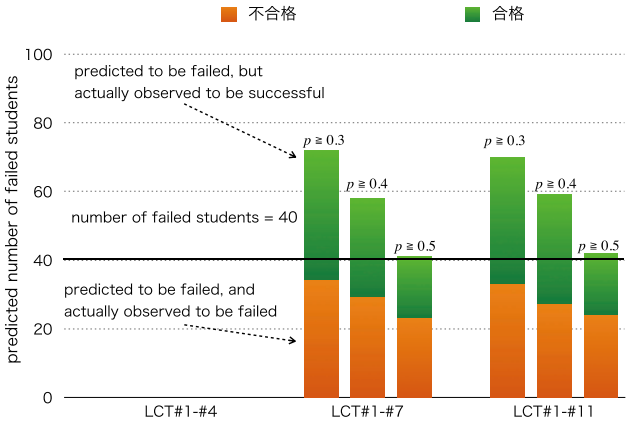


図5　期末試験の合格者予測数と実際の合格者と不合格
（不合格者数 =40人の場合）

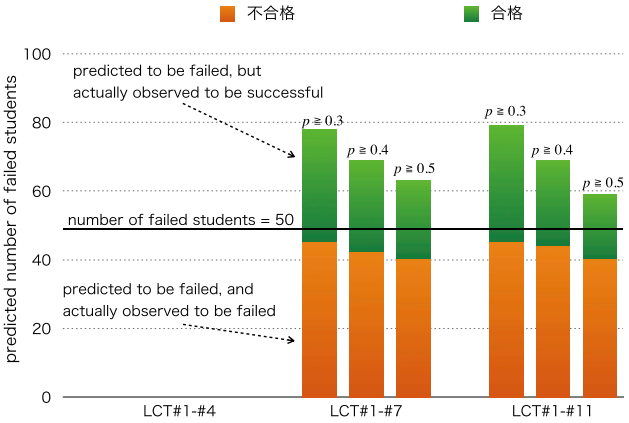


図6　期末試験の合格者予測数と実際の合格者と不合格
（不合格者数 =50人の場合）

7　まとめ

LCTを使ってできるだけ早めに学生のドロップアウトのリスクを少なくしようと、確率付きで期末試験の成否のアラートを流すことができるまで予測手法が整ってきたことをこれまで報告してきた。その際、ある一定のところまでは予測が可能であるが、教員間には評価バイアスがあるため､どうしても超えられない壁があることも感じていた。ここではその実際の評価バイアスの発生例を紹介している。

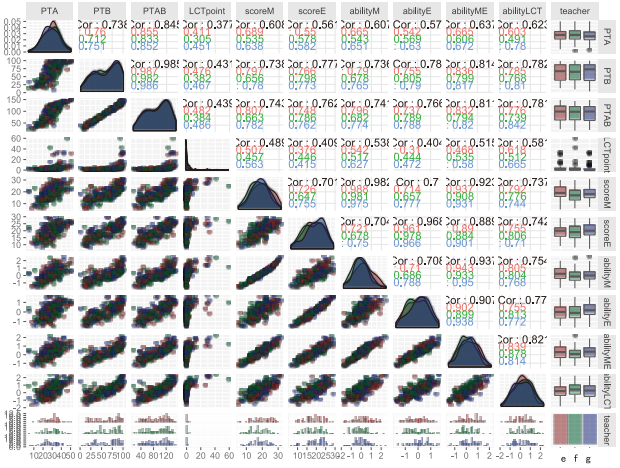
そこで、期末試験にも多肢選択式のCBTあるいはマークシート方式のテストを採用すれば、LCTを使った期末試験の合否予測の精度が上がるのではないかと期待される。実際、今回一学部だけではあるが、そのような環境を作り、このことを確認する機会を得ることができた。

LCTを使って記述式の期末試験の評価を各教員が行なう場合の期末試験の合否予測精度と、CBTを使った期末試験の評価をコンピュータだけで行なう場合の期末試験の合否予測精度を比較すると､後者の予測精度が良くなった。合否の2値分類での誤分類率は20％程度から10-20％程度まで下がり、また、特筆すべきは、前者では不合格予測者の中からの実際の不合格者の的中率は40-50％程度しか得

られていなかったものが、後者では50-70％程度にまで向上することがわかった。

付　　録

付図１に、プレースメントテストＡとＢのタイプ、中間試験単独、期末試験単独、両者を合わせた結果、LCT ポイントなどの詳細な情報が記載された相関図を示す。



付図１　均質な３クラスでの、プレースメントテスト、LCT、期末試験の結果の関係詳細
（期末試験評価を CBT で行なう場合）

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、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、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.

６）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.

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

８）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.

９）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.

10）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.

11）Martin Liz-Dominguez, Manuel Caeiro-Rodriguez, Martin Llamas-Nistal, Fernando Mikic-Fonte, Predictors and early warning systems in higher education-A systematic literature review, Learning Analytics Summer Institute Spain 2019: Learning Analytics in Higher Education. 84-99, Vigo, Spain, June 27-28, 2019.