教学 IR における気付きと事例紹介
― データを通して英語業者テストや学生授業評価を眺める ―

岡　田　耕　一

要旨

本年度取り組んできた教学 IR 活動を通じて得た気付きについて共有すると共に，事例紹介を通じて教学 IR でどのようなことが出来るかについて紹介を行う。

事例としては TOEIC，VELC の英語業者テストに関する得点分布の特徴や，得点換算の妥当性に関する検討，TOEIC 試験の複数回受験による成績の変化の様子，学生授業評価アンケート結果の回答割合の変化に関する可視化等を紹介する。

最後に，教学 IR に関する懸念事項や課題についてまとめる。

キーワード

教学 IR，TOEIC，VELC，学生授業評価

1 はじめに

筆者が所属する教学 IR 部では本年度，TOEIC の年 2 回受験の効果や，2017 年度に行われた英語カリキュラム改革に伴う学生の動向の変化に関して，データを元にした比較検討を行うべく様々な方法を試みてきた。これらの取り組みを行う中でいくつかの気付きがあった。本稿ではそれらについて情報を共有すると共に，教学 IR でどのようなことが分かるのかについて知ってもらうため，幾つかの事例について例示を行う。

2 データについて

現在，教学 IR 部で，比較検討に利用可能なデータとしては，TOEIC 及びVELCの英語業者テストのスコア，共通教育科目の成績データ，学生授業評価アンケート結果の 3 種類にアクセスが可能な状態にある。このうち，英語業者テストのスコアと共通教育の成績データは個人に紐付いているが，学生授業評価のデータは無記名回答であるため個人へは紐付いていない。

TOEIC とVELCのデータはそれぞれ，49 項目及び 46 項目で構成されており，学生番号や氏名，性別，所属，受験日等の 39 項目は概ね共通している。ただし，TOEIC では項目として記録されている Listening と Reading の個別得点がVELCではなぜか省略されている他，VELCでは入学直後のクラス分けに用いる前提で 1 回しか受験することを想定していないためか，TOEIC には存在する受験時の年次の項目が省かれている等，細かな違いがあった。他の項目はともかくとして，比較を行う上で Listening と Reading の個別得点が得られないのは惜しいところである。

既にあるデータを提供してもらうという状況では致し方ない所はあるし，現状の学生授業評価のように無記名で収集することが前提のデータについても仕方のない部分があるが，

データの容量的には微々たるものであるから，元データに存在する項目は可能な限り漏らさず記録して頂けるよう提言しておく。

3 個人に紐づいたデータによる比較

3.1 データから見るTOEICと VELC

個人に紐付いているデータに関しては，相互に相関係数等を用いた比較が可能である。ここでは， TOEIC とVELCのスコアを題材にして，どのような比較が可能か紹介する。例えばGNU R の pairs.panels() 関数を用いて TOEIC とVELCのスコアを比較すると図 1 のようなプロットが得られる。この図は，TOEIC の Listening， Reading，合計点及びVELCの合計点について，上三角成分にはPearson の積率相関係数（相関係数隣の *** は$p < 0.001$ ， **は$0.001 \leqq p < 0.01$ ， *は$0.01 \leqq p < 0.05$ を意味する），対角成分には得点分布のヒストグラム (50 点刻み)，下三角成分には散布図及び平均値（赤点）と線形近似による回帰直線（赤線）が図示してある。

このようにして比較を行うと TOEIC とV ELC のスコア間に$p < 0.001$ で有意な正の相関が存在していること， TOEIC の Listening と Reading の得点はそれぞれ TOEIC の合計点に対して相関係数にして 0.9 以上の強い相関があること， Listening と Reading の間には 0.7 前後の中程度の相関があること ,TOEIC とVELCの間にも 0.7 程度の相関があること等を確認することができる。 TOEIC とVELCの間の相関が， TOEIC の Listening と Reading の間と同程度の相関を持っていることは，両者が測ろうとしている能力とスコアにある程度の共通性が存在していること客観的に示していると言えるだろう。

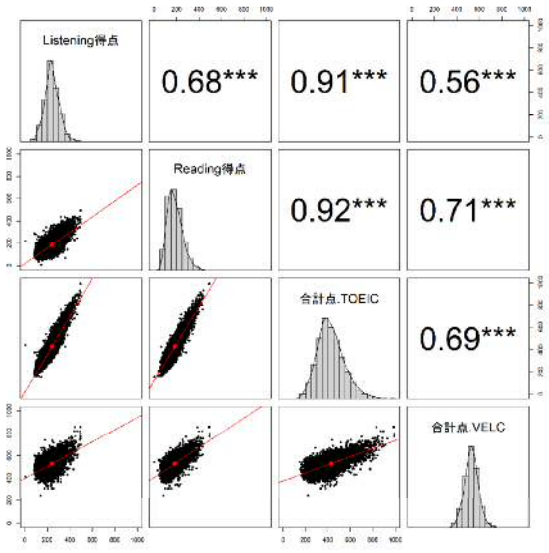

図 1 TOEIC, VELC のスコア相関

表 1 TOEIC， VELC の要約統計量

	Min	Q1st	Med	Q3rd	Max	Mean	Sd	N
TOEIC	150	345	415	500	990	428.6	118.5	8221
VELC	246	491	528	568	857	529.4	61.78	8221

一点気になったのは，ヒストグラムの広がりを見た際に TOEIC に比べてVELCの分散が小さい点である。要約統計量は表 1 の通りであった。標準偏差に約 2 倍程度の開きが生じており，差にすると 60 点の程度の開きが生じていることが分かる。平均点はVELCの方が 100 点程度高いが， TOEIC と比べると成績下位のグループは高い得点，成績上位のグループは低い得点が出る傾向が生じており，このことは散布図や回帰直線からも確認できる。つまり単純にVELCの方か難易度が低いということではなさそうな傾向が伺われる。

分散にこのような差が生じる要因の一つとしては， TOEIC の Listening45 分 100 問，Reading75分 100 問に対してVELCは Listening25 分 60 問，Reading45分 60 問のように，時間と設問数が共におおよそ 6 割程度であることの影響が大きそうに思われるかもしれない。しかし，大数の法則からも明らかなように，設問数が少ない場合は母集団の分散よりも大きな分散が生じ，設問数が増えるに従って母集団の分散に漸近するはずである。

念のためテストにおけるスコアを個人の能力値 $_0$，試験で獲得可能な最高点 $_{max}$ として，各設問に $_0$ $_{max}$ の確率で正解するというモデルを立ててシミュレーションを行ってみた。結果は予想通り設問数の増加に従って分散が母分散に漸近していくだけであった。つまり，本来であれば設問数の少ないVELCの方が分散は大きくなければおかしいはずである。ところが，今回比較した結果ではそれとは逆の傾向が表れている。

TOEIC の公式サイトにあるテストの結果について[1]によれば，「TOEIC のスコアは素点 (Raw Score) ではなく，スコアの同一化 (Equating) と呼ばれる統計処理によって算出された換算点 (Scaled Score)」であると説明されているため，このことが影響している可能性はありそうである。しかし具体的な換算式が示されていないため，実際にどのような影響が生じているのかについてはこれ以上の確認が難しかった。

3.2 TOEIC,VELC の得点換算について

異なる業者テストを利用する場合，テスト相互の得点換算式は関心の対象であろう。

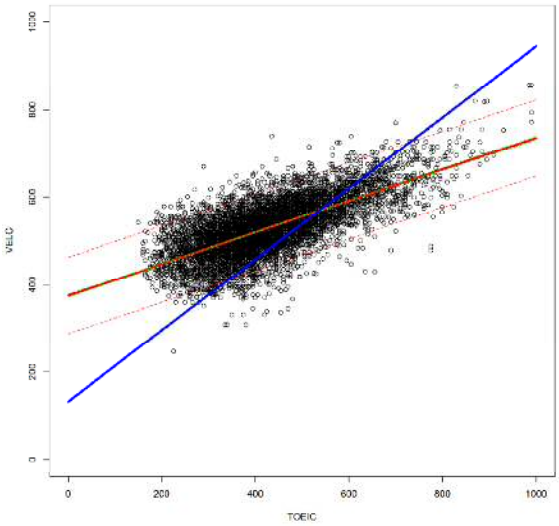

図 2 TOEIC-VELC 得点換算
回帰直線，予測区間，信頼区間，
VELC-TOEIC 換算表の直線

VELC Test 公式サイトの VELC-TOEIC 換算表[2]によれば，VELC の 300，400，500，600，700，800 点が TOEIC の 205，330，450，575，700，820 点に相当すると説明されている。前者を $_V$，後者を $_T$ と置くと，若干の誤差はあるもののおおよそ $_V$ $_T$ + の関係が成り立っており，最小二乗推定すると $a = 0.812$，$b = 133$ が得られた。この係数により $_T$ から換算した VELC のスコアを S'_V と置くと，両者の差 $_V - S'_V$ は 0.4，-1.1，1.4，-0.1，-1.6 であり誤差の範囲で直線に近似していることが分かるだろう。これに対して手元のデータから近似した値は $= 0.361$，$b = 375$ であった。本学では VELC は入学直後に受験するのに対して TOEIC は年度末に受験している。受験時期に開きがあるため，多少の誤差が生じるのは仕方がない。しかし，プロットした結果 (図 2) を見ても散布図に対して明確な乖離が生じていることが確認できる。傾きにして 0.5 近い乖離は誤差とするには大き過ぎるだろう。サンプル数がたかだか 8,000 強しかないとは言え，95%予測区間と信頼区間の範囲を考慮しても VELC-TOEIC 換算表の得点対応とはかなりの乖離が生じている。この換算表が妥当かどうかについては疑問が残る結果と言えよう。

3.3 TOEIC の受験回数と得点の推移

TOEIC の受験回数を重ねることでどの程度成績が伸びるかについても関心を持たれるであろう。図 3 では本学で確認可能な記録の範囲で各個人について受験回数を 1 回目から順にカウントして各回の点数の相関を調べた結果である。紙面の都合で 9 回までを抜粋した。散布図上で の対角線 (青線で示した) よりも上にあれば受験回数を重ねたことで成績が上昇したことを意味している。受験回数 9 回までの範囲では $p < 0.001$ で有意な正の相関が強く認められる。また，若干の例外はあるものの，赤線で示した回帰直線は

概ね青線よりも上にあることも確認できる。

例外的なのは例えば１，２回目を比較した際に高得点の領域で回帰直線が対角線と大きくクロスしている箇所である。スコアの平均値を確認すると１回目よりも２回目の受験の方がわずかながら平均点を下げている傾向は見られる（図　4）。しかし散布図で見る限りにおいては成績上位のグループで大きくスコアを下げている傾向は見られない。

これはどうも成績下位のグループの人数が成績上位のグループに比べて相対的に多いため，重心を中心として回帰直線が右回りに回転した形となり，回帰直線の高得点側が成績の下落を意味する領域にかかってしまったのではないかと考えられる。

3.4 講義科目への応用

個人に紐付いているデータは，同様にして講義科目でも様々な比較検討の材料として有用である。各科目の成績間で相関を調べたり，場合によっては業者テスト等外部の指標と相関を調べたりすることが可能である。ただし，業者テスト間に見られたような中程度以上の相関が現れるのは稀なようである。多くの場合，有意な正の相関は見られるものの 0.2 ～ 0.3 程度とかなり弱い相関に留まっていた。

各講義では目指す内容が大きく異なるため，成績間には大きな相関は発生し辛い状況はあると思われる。しかし弱くはあるものの正の相関が有意に表れている点は，学生の能力を反映して適切に成績評価が行われていること証左と言えるのではないだろうか。

4.　個人に紐付いてないデータによる比較

2017 年度に行われた英語カリキュラム改革の検証に当たっては，科目構成が大幅に変更されたこともあり講義科目の成績で比較を行うのは簡単ではなかった。一方，継続的な指標としては，学生授業評価のアンケート結果が存在した。ただし，これは無記名回答であるため個人への紐付けが存在しない。このため，相関関係を散布図や相関係数で確認するということが出来なかった。

そこで，主には t 検定を用いることで，有意に差のある分布か否かを調べ，有意な差が見いだせた場合に回答の平均値が上がったか下がったかを確認するというアプローチを取った。

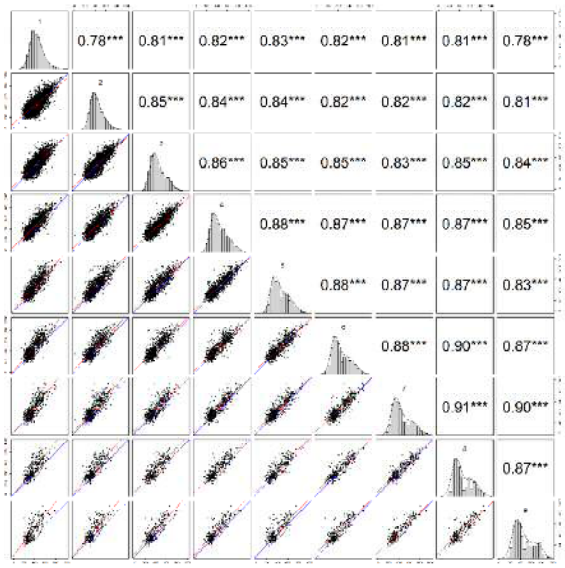

図　3 TOEIC 受験回数とスコアの相関

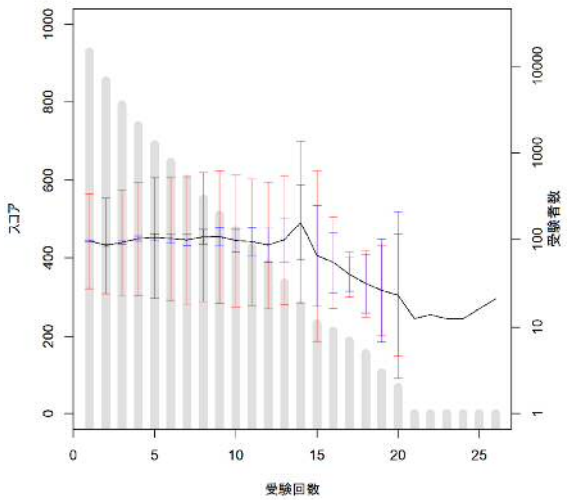

図　4 TOEIC 受験回数とスコアの推移
―：平均値，I: 標準偏差，I: 95%信頼区間

これには表　2 のような表を作成した。対角成分に回答の平均値を示すと共に，上三角及び下三角成分に平均値の差と p 値を示し，p 値により有意な差が見いだせる場合は不等号と彩色により大小関係を示すことでスコアの有意な変動箇所を可視化している。またこれに対応する図　5 ，図　6 のようなグラフを作成することで回答内容の詳細な分布や，スコアの推移について補足を行った。

この例では，2018Q1 と Q2 の間には有意な差はなかったが，2019Q1 は，2018Q1 ，Q2 に対して有意にスコアが上昇していることが t 検定により示唆されている。これを受けて図　5 ，図　6 を見ると，有意に変化した箇所で生じている回答割合の傾向や，スコア平均が変化している箇所を効率的に確認することが可能であろう。

ただし，変化を概観するツールとして順序尺度をスコア化して平均を取った値を示しているが，この値には本来意味はないのでその点には注意を要する。

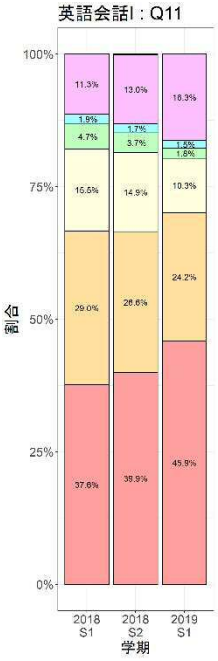

図　5 回答割合

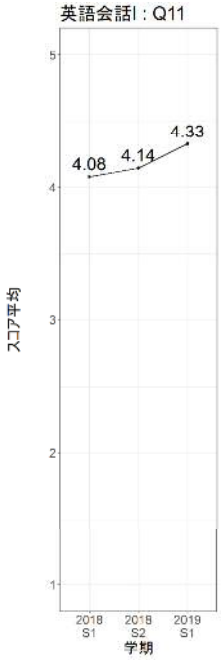

図　6 スコア平均

表　2 英語会話 I : Q11 興味関心

	2018S1	2018S2	2019S1
2018S1	4.08	-: -0.07 (p=0.277)	<: -0.25 (p=0.000)
2018S2	-: 0.07 (p=0.277)	4.14	<: -0.18 (p=0.002)
2019S1	>: 0.25 (p=0.000)	>: 0.18 (p=0.002)	4.33

5 まとめ

以上，教学 IR でどのような事が分かるかについて幾つかの事例を挙げて紹介すると共に，これらの取り組みを通じて気付いた事項等について紹介を行った。

教学 IR にはデータマイニング的手法を用いてデータの中から何らかの因果関係を見出すことを期待されているように思われる。しかしアクセス可能なデータの蓄積や紐付けが十分とは言えない場合，そこから何らかの意味のあるデータを掘り出すことは簡単ではないという現実がある。

また，カリキュラム改革の検証のような文脈では，データから改革が順調であること，または問題が生じていることを明らかにすることが使命となる。これも問題が発見できる場合は良いが，問題が発見できない場合は難しい問題が残る。

一般に，統計的仮説検定では，帰無仮説と対立仮説を設定した上で，検定によりどちらを採択するかの議論を行う。教学 IR でも仮説を設定するところが出発点と言える。

検証によって順調であることを示したいのか，逆に問題を発見したいのかによって仮説の立て方は大きく異なってくるし，どちらの立場で分析に当たるにしても，ないことの確認はいわゆる悪魔の証明である。考慮すべき事項が漏れている可能性は永遠に残り続けるため，どこまでやれば良いのかという判断は簡単ではない。データが乏しい場合はもちろんであるが，逆にデータが豊富にある場合は

組み合わせ爆発的な問題も生じる。

最近では p 値ハックが各所で問題視されているように，探し方次第ではないはずの問題を無理やり掘り起こすようなことにならないかという点も懸念材料の一つである。そのような意味で，教学 IR には一種の危うさが存在している。

以上のことから，当事者が積極的に関与し，着地点を明確にしたうえで協力しながら分析を進めるという姿勢がなれば，教学 IR の活用は難しいのではないかと考える。

本年度取り組んできた検証に関して言えば，検証内容の検討や報告の際には，表　2 ，図　5 ，図　6 に示した図表が授業科目ごとに並び，これは相当な分量となった。教学 IR においてデータの分析は試行錯誤の連続であるが，此方で用意した見せ方と相手の期待した見せ方に齟齬があり何度もグラフを作り直すようなこともしばしば発生した。あらかじめ希望するデータやその見せ方が定まっていればこれらはかなり減らせるはずだが，これも試行錯誤の中で定まってくるものなので容易ではないだろう。

データの重要性が重視される昨今において，教学 IR の可能性については疑うべくもない。しかし，今後教学 IR を進めていくうえで，利用可能なデータの拡充，確認したい事項と着地点の明確化，当事者との密接な連携については更なる整備が必要であろう。

（大学教育センター　講師）

【注】

1)　TOEIC / テスト結果について
https://www.iibc-global.org/toeic/test/lr/guide04.html

2)　VELC Test / VELC-TOEIC 換算表
https://www.velctest.org/outline/#outline5