埼玉大学紀要　教育学部，65(1)：99−106（2016）

評価方法による順位の入れ替わり　1

―シミュレーションに基づいて―

萩生田　伸　子　　埼玉大学教育学部心理・教育実践学講座

キーワード：成績評価、シミュレーション、多段階評価、バイアス

1．はじめに

多くの大学生にとって学業成績は関心事の１つであると考えられる。成績によって一学期間に履修可能な科目数が変わってくるといった条件がついていれば、卒業までの年数が延びる可能性や、特定の資格取得に必要な科目が履修できなくなる可能性が存在する。あるいは、奨学金申請で推薦順位決定に成績が利用されるのであれば、獲得できるかどうかに大きな影響を与えることは明らかであろう。成績の評価をおこなうのは教員であるが、各種の処遇に影響を与える可能性がある以上、公平・公正な評価が求められるのは言うまでもない。

ここで注意をすべきことがある。ある程度以上の規模の学部等では必修科目を配置する都合上、学生を幾つかのグループ（クラス）に分割し、同一授業科目を複数の教員による同時限開講としていることが普通である。たとえば埼玉大学教育学部心理・教育実践学講座の場合は、教育心理学概説、学校心理学（生徒指導・進路指導)、学校カウンセリング（教育相談）がそれに該当する。このような複数クラス間の成績評価にいわゆる「甘い辛い」が存在した場合、「甘い」クラスに所属する学生は若干有利ということが起こりうる。そのような問題の発生が想定される場合、公平を期するためにとりうる方法はいくつかある。たとえば同時限開講科目の担当教員が全クラスを少しずつ分担して担当した上で評価をおこなう、共通尺度上で比較するための何らかの工夫をする等である。しかし種々の理由により実現困難なケースも珍しくはない。厳しい評価がおこなわれたクラスに加点修正をおこなうというのも単純に実施する事は難しい。クラス間で評価の厳しさに差異があるように思われたとしても、実際にある特定のクラスに成績優秀な者が多く存在した可能性や、あるいは学習活動を真剣におこなおうという雰囲気がクラスのメンバーにある程度共有されるといったことの影響は否定できないと考えられるからである。

これらの点について検証を加えることは困難であるにせよ、成績を何段階評価にしたかの影響や、部分的な「甘い辛い」バイアスの混入が総合成績（順位）等にどのような影響を与えるかについての検討はシミュレーションによって可能である。そこで本稿では多段階の成績評価と類似した状況を模した人工データを生成し、その生成条件の差異や生成データの一部分に操作を加える事によって、成績の合計点に基づいた順位がどのような影響を受けるかについて検討をおこなうための素案および結果の一部を紹介する。

2．方法

ここではシミュレーションの設定の素案について述べる。状況は幾分入り組んでいるが、手順を簡潔にまとめると以下のとおりである。人工データの生成と確認および解析にはR3.1.2とSPSS Ver.19を用いた。

(1) ある特定の平均、分散共分散行列に従う多変量正規乱数を発生
(2) 条件設定に従ったデータの変換（評定を５段階化、特定の群の平均点を高くする等）
(3) 合計点（総合点）の計算
(4) 条件ごとの合計点数間の相関係数および順位の入れ替わりの比較検討

以下、各段階について説明をおこなう。冗長な処理も含んでいるがこれは成績評価の状況との対比を分かりやすくすることを意図したためである。

(1)データ行列のサイズや平均、分散共分散等の設定は任意であるが、たとえば大きさ150×５、150×10、150×15という３種類の多変量正規乱数を発生させる。これは成績評価の場面に置き換えると一学年の人数が150名、総合成績（合計点あるいは総合順位）の算出に使用される授業科目数（項目数）は5，10，15の３種類という設定である。乱数の発生は、たとえば平均ベクトルは各要素が2.7、分散共分散行列については標準偏差σが0.30および0.50、項目間の相関係数ρが.10、.30および.50の条件を想定する。この数値設定は、次の多段階データ化を見越したものだが、たとえば０から100点までの素点表記にしたいのであれば適宜変換とそれに付随する処理をおこなえば良い。

なお、項目間相関の平均が.30の場合のα係数は項目数が５の時に.68、10の時に.81、15の時に.87程度となる。項目間相関の平均が.50であれば項目数が５の時に.83、10の時に.91、15の時に.94程度の値となる。項目間相関の平均が.10であればそれぞれ.36、.53、.63程度である。

(2) 先に発生させた乱数の各値を0，1，2，3，4の５段階、もしくは０から0.5刻みとした９段階に変換する（つまり順序つきのカテゴリカルデータ化する）。換言すると前者は５段階評価を模したものであり、後者はさらに中間の評点も許容した状況である。カテゴリ化の閾値は５段階の場合

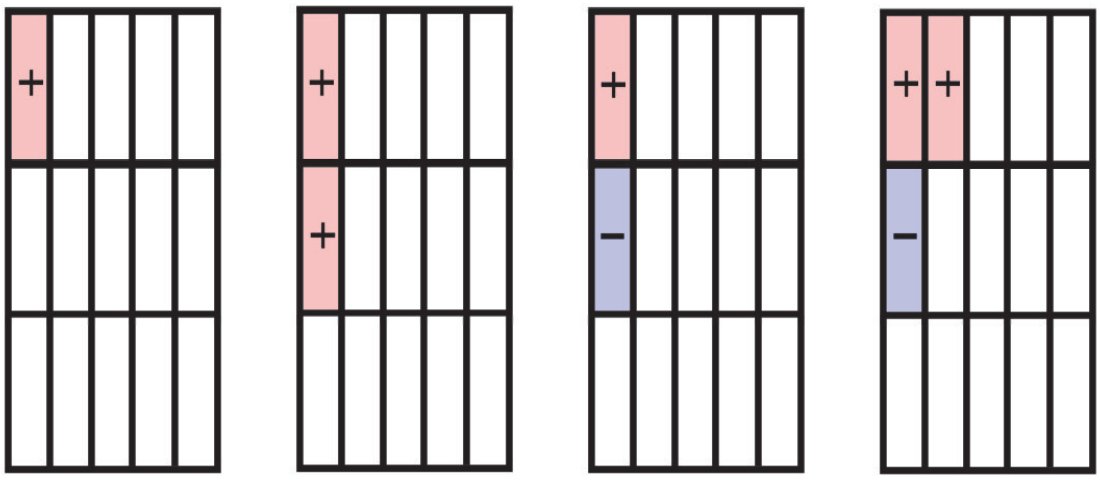

図１　データのイメージの例（＋が加点部分、－が減点部分を示す）

0.5, 1.5, 2.5, 3.5とし、９段階の場合は0.25, 0.75, 1.25, 1.75, 2.25, 2.75, 3.25, 3.75, とする（下端は含まず、上端は含む）。

仮に５段階評価をおこなった教員と９段階評価をおこなった教員が混在する状況を模すのであれば、多段階化したデータセットのうち一部が５段階評価、残りが９段階評価となっているような混合データを生成すれば良い。

上記に加えてデータセットの一部分に1.0点の加点もしくは減点の処理をおこなう。具体的には観測数を３つのブロックに分割し、項目単位でいずれかのブロック全体に一律の加点もしくは減点をおこなう。図１にその例を示す。図中の＋という表記は当該ブロックの全観測値に対して加点をおこなうこと、－という表記は当該ブロックの全観測値に対して減点をおこなうことをそれぞれ表している。

これは、それぞれの授業科目は50人ずつの３クラスに別れている（ただし授業科目ごとにクラスメンバーの組み替えは一切ない）という状況において、１つ以上のクラス全体で評価が甘め、もしくは辛めという状況を模したものである。たとえば図１の右端のケースでは一番上のブロック（クラス）の学生が受講する科目のうち２つの評価が緩め（ほとんど全員が「優」）であり、２番目のブロック（クラス）の学生が受講するクラスのうち１つは評価が厳しいという状況に相当する。

(3) 上述の設定に基づいて生成した各データセットについて行和を求める。これは学生ごとに成績の合計点が算出された状況を模したものである。

(4) 合計点同士の相関係数、もしくは順位の差を計算する。この相関係数が高い、あるいは順位の差の絶対値が全体として小さいという結果となった場合は、前述の状況設定の影響は少ないことを示していると考えられる。

以上を各条件ごとに少なくとも100回程度反復し、平均的な相関係数を算出したり、順位差の分布状況をまとめることによって、どのような条件下でどのような形での影響が生じるかについてある程度の知見が得られるものと考えられる。次項では幾つかの条件設定でのシミュレーション結果を例示する。

３．結果と考察

はじめに成績評価の甘い辛いがどのブロックにも含まれない状況設定下において、５段階評定および９段階評定を用いた際の順位差がどのような分布をしたかについてのシミュレーション結果の例を図２ａ、図２ｂに示す（項目数５、項目間相関は.50、各項目の標準偏差は0.3）。このうち図２ａ左は連続値（多段階化をおこなう前の）データの合計得点から求めた順位と５段階評定化したデータの合計得点から求めた順位の差の絶対値を横軸とした分布状況を示したヒストグラムであり、図２ａ右は連続値データの順位と９段階評定化したデータの順位の差の絶対値の分布状況を示したヒストグラムである。同様に、図２ｂ左は５段階評定化したデータの順位と９段階評定化したデータの順位の差の絶対値、図２ｂ右は５段階評定化したデータの順位と５段階・９段階の評定が混在したデータの順位の差の絶対値の分布状況の一例である。言い換えると、図２ａのヒストグラムは元のデータが同一であっても連続値データをカテゴリ化した段階でどの程度の順位の入れ替わりが生じうるのかを表したもの、すなわち多段階化の影響を示したものとなっている。５段階評定化した場合は順位の入れ替わりが20番以内の範囲に100ケース前後が収まっている一方で、50番以上の順位の入れ替わりも数ケース発生している。９段階評定化した場合の順位

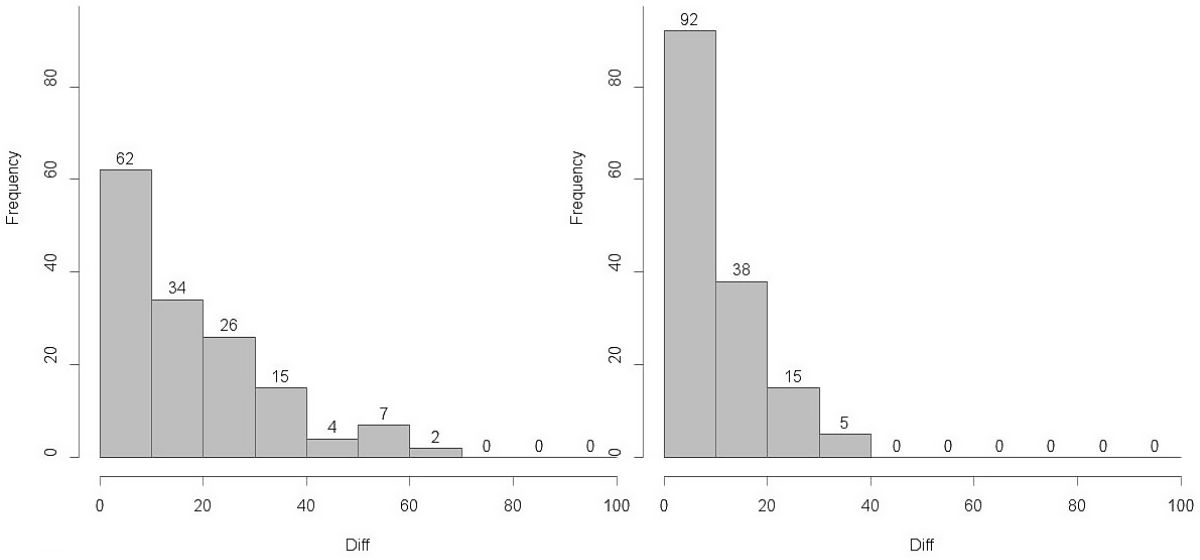

図２ａ　評定の段階数と順位の差（ρ =.50、項目数=5、σ =0.30）、連続データとの比較の例
（左図は５段階化、右図は９段階化したものとの比較）

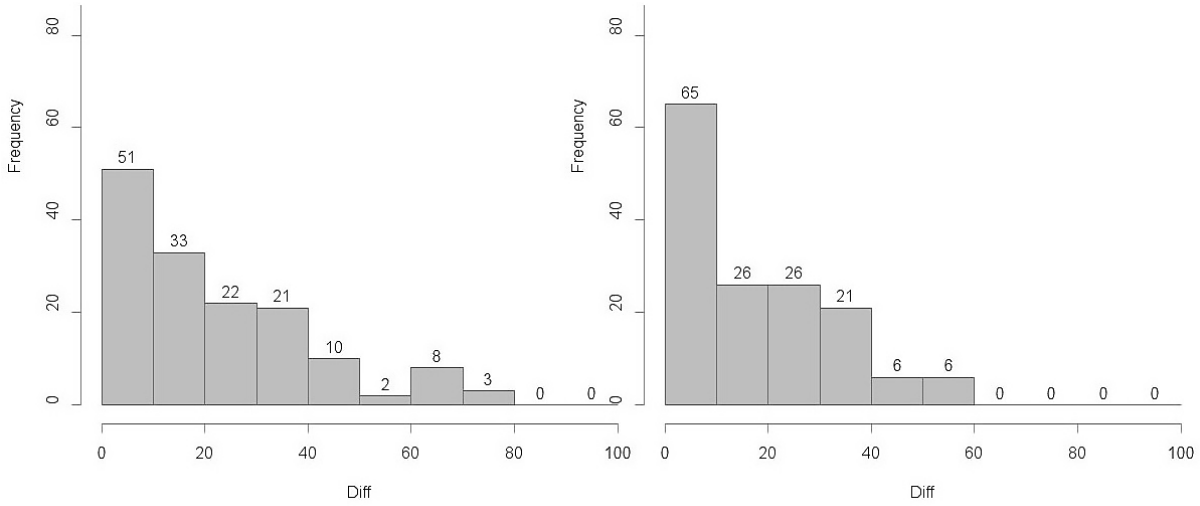

図２ｂ　評定の段階数と順位の差（ρ =.50、項目数=5、σ =0.30）、5段階評価との比較の例
（左図は９段階化したデータとの比較、右図は５段階・９段階混合データとの比較）

の入れ替わりは20番以内に130ケース前後が収まっており、50番以上の入れ替わりはほとんど生じなかった。もちろん、連続値データのような細かな刻みによる評価をおこなうことは事実上困難であることを考えれば、上述の条件下での入れ替わりについてはほぼ心配する必要はないかもしれない。

図２ｂ右左のヒストグラムは、同一データを５段階評定と９段階評定に変換した際の順位の入れ替わり、および５段階評定と９段階評定が混合した場合の順位の入れ替わりを示している。つまり５段階評定との比較例となっている。前者については特に順位の入れ替わりが大きくなっており、データセットによっては90番以上の順位の入れ替わりもみられた。しかし、先に述べた多段階化による影響に加え、評定の段階数が異なる変数間の相関係数の最大値は（双方の変数におい

て各段階への反応が１つ以上ある状況下では）1.0にはならないことを考えればこれはある意味で当然の結果であろう。後者は９段階評価３項目と５段階評価２項目を合計した際の順位と、５段階評価での５項目を合計した際の順位の差である。当然、前者よりは順位の入れ替わりが少ないという結果となった。

なお、ここに示した例は各科目の成績間の相関が全て.50という実際には存在しないような条件設定の元での結果である（ただし項目数は５であり条件としては厳しい)。多変量正規乱数発生時の相関が.30のより現実に近い条件下ではさらに大幅な順位の入れ替わりが発生する場合がある。たとえば図３は図２ｂと同じ設定で各科目間の相関係数のみ.30に変更した条件下における一例であるが、全体として分布が右側にずれている。すなわち相対的に大きな順位変動が生じていることが分かる。

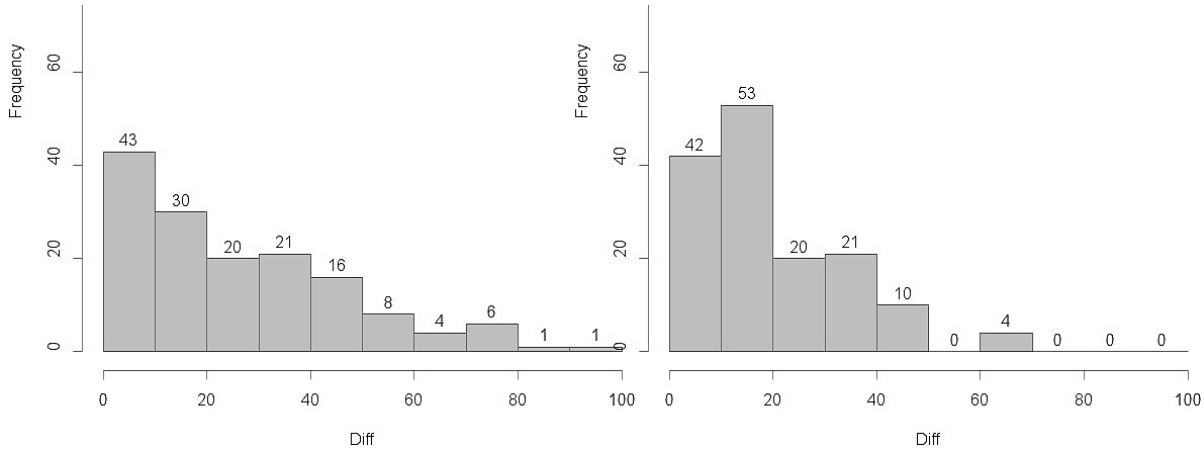

図３　評定の段階数と順位の差（ρ =.30、項目数=5、σ =0.30)、５段階評価との比較の例

次に、合計得点を算出するのに使用する項目（科目）数の影響について述べる。図４および図５は図２ｂと同じ条件下で、科目数のみそれぞれ10、15に変更した際の順位の入れ替わりを示したものである。合算する科目数が５つであるときと比較して、10科目、15科目条件では分布が全体的に左側に寄り、50番以上の順位の入れ替わりも減少している。今回設定した条件の中では順位の入れ替わりが起こりやすい５段階評定と９段階評定のペアでも、110ケース前後が20番以内の入れ替わりに収まっている。すなわち、総合順位を定めるのに使用する科目数がある程度多ければ、何段階の評定を用いるかの影響は軽減される。

なお、ここまでの議論は、評価の甘い辛いといったバイアスを取り込んだシミュレーションについてではないために本質的ではないかもしれない。しかし、元が同じデータであっても多段階の順序つきカテゴリカルデータに変換することによって、場合によっては大幅に順位が入れ替わりうるということを改めて明らかにしたともいえる。

最後に、全体の1/3にあたる特定の一クラスのみ、２つの授業科目の成績評価が甘かったという状況を模したシミュレーション結果の例を図６、図７に示す。これらは加点なしの(成績評価が「甘い」クラスが存在しない）場合の順位と、前述のとおり全体の1/3に加点がおこなわれた（成績評価が「甘い」クラスが存在する）場合の順位の差の絶対値の分布状況をヒストグラムで表したものである。図６は科目数が５のケース（図１の右端のうちマイナス部分がない状況）で、図７は科目数が15のケースである。また、各図の左側は５段階評価同士、右側は９段階同士の順位の入れ替わり状況である。

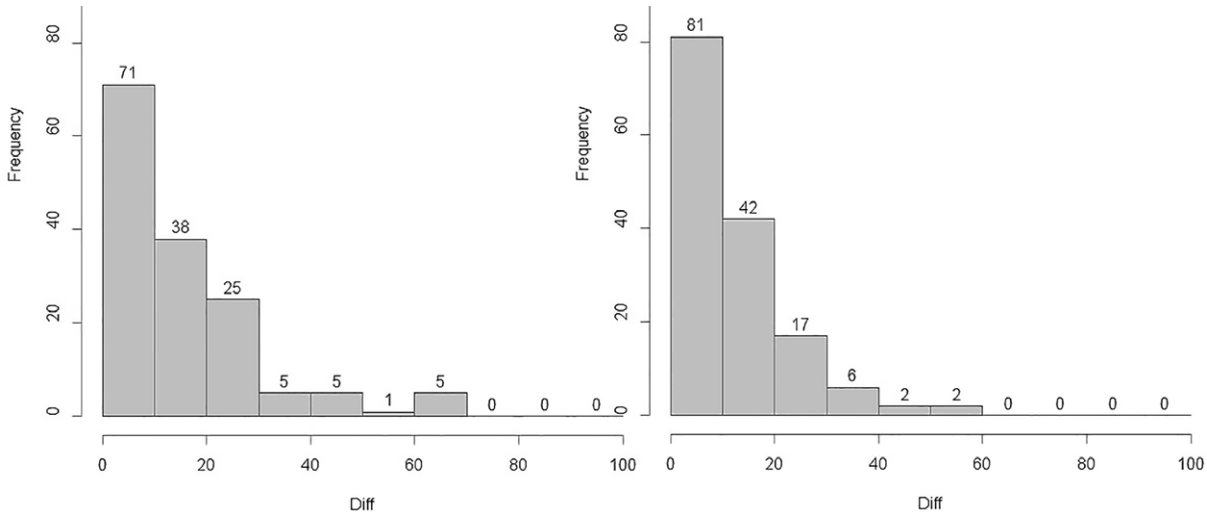

図４　評定の段階数と順位の差（ρ =.50、項目数=10、σ =0.30）、５段階評価との比較の例

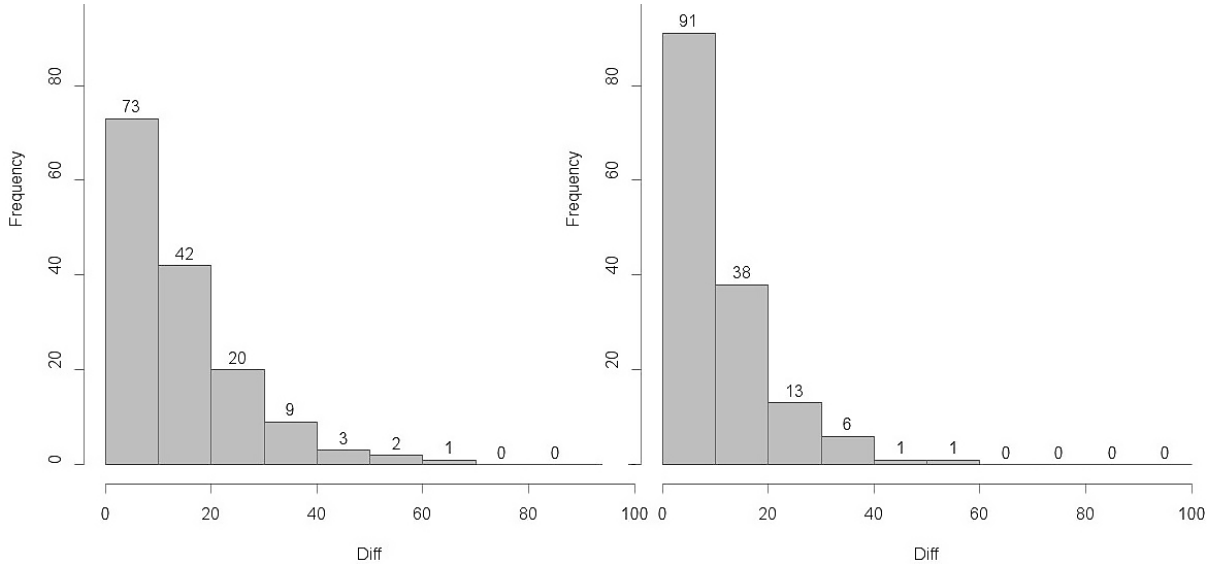

図５　評定の段階数と順位の差（ρ =.50、項目数=15、σ =0.30）、５段階評価との比較の例

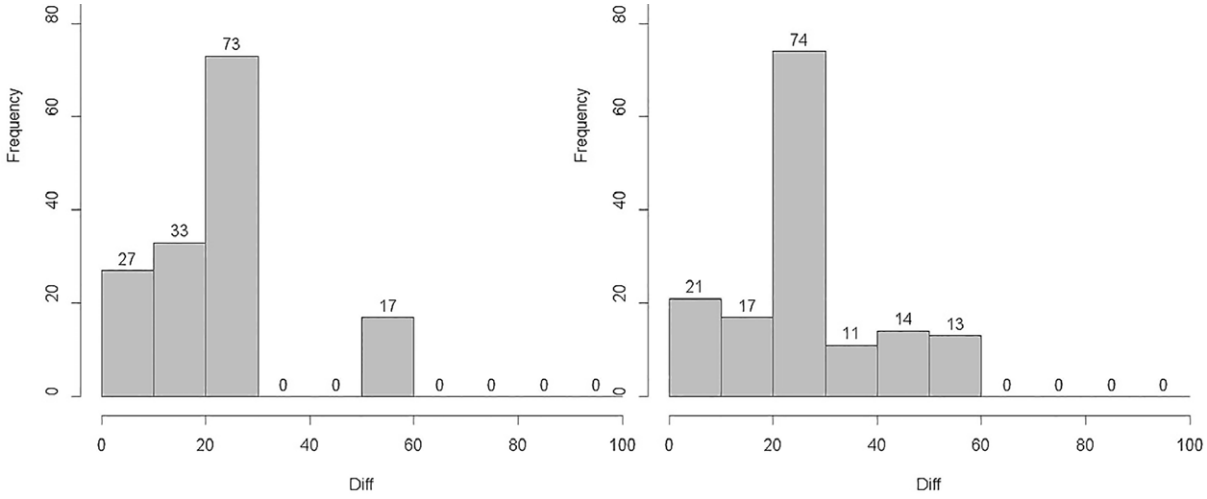

図６　評定の段階数と順位の差（ρ =.50、項目数=5、σ =0.30）
１クラス２科目に加点条件、左は５段階評価同士、右は９段階評価同士

５科目中２科目で評価が甘いクラスが存在した場合には、５段階評価、９段階評価を問わず、中程度の順位の入れ替わりが多く生じているようである。そもそも加点によってあるクラスのほぼ全員が最高評点を得ている状況では同順位も大量発生していることも想像に難くない。

他方、15科目中２科目で評価が甘いクラスが存在したという状況下では、影響はさほど顕著ではないように思われる。実際、この例では順位の変動が10以内であるケース数は５段階評定でも半数程度、９段階評価では90ほどである。もっともこれは、合計得点の算出に使用する科目数が多ければ相対的に少量のバイアスが入ったとしても結果は安定するというある意味で当たり前の結果ともいえる。

４．おわりに

本稿では幾つかの条件設定の元で、複数科目を合計した際、順位の入れ替わりがどの程度生じるかを中心に検討をおこなった。その結果、項目間相関が高く合計得点を計算するのに使用する項目数が多ければ順位の入れ替わりはそれほど起こらないという当たり前の結果となった。しかし、観測された順位の入れ替わりはあくまで今回用いた条件設定で人工データを生成した場合にはどうなるかということであり、他の条件下では結果がまったく異なる可能性がある。たとえば、実際の成績分布はほぼ考慮していないので、この結果がどの程度現実を反映しているかは不明である点には留意する必要がある。今回は考慮に入れなかった成績の分布状況と相互関係、たとえば、現実場面での科目の成績間相関は全体としてほぼ０に近いものから0.5程度のが混在している可能性も考えられるため、条件設定に加えるべきかもしれない。また、人工データ生成に使用したのは多変量正規乱数のみであり、低成績層（たとえば、ほとんどの科目で不可に該当する成績である等）に該当する乱数の混合はおこなっていない。平均ベクトルの設定や、多段階評定化する際の閾値が適切であったのかについても検討が必要である。次稿ではこれらの条件設定をどこまで取り入れるべきかを吟味した上で、シミュレーションをおこなった結果を示す。

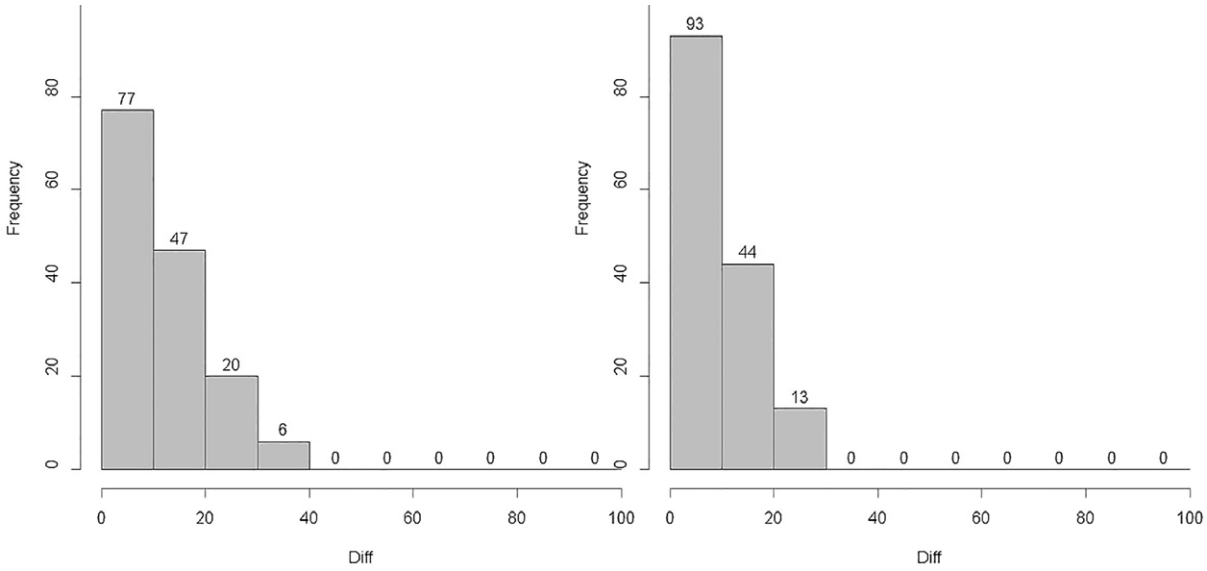

図７　評定の段階数と順位の差（ρ =.50、項目数=15、σ =0.30）
１クラス２科目に加点条件、左は５段階評価同士、右は９段階評価同士

参考文献
山田　剛史・杉澤　武俊・村井　潤一郎　2008　Rによるやさしい統計学　オーム社
石田　基広　2014　R言語逆引きハンドブック　改訂２版　シーアンドアール研究所

（2015年９月30日提出）
（2015年10月７日受理）