教員評価の方法的問題
—— 研究活動の自己評価を眺めて ——

阿久津洋巳*　菅原正和*

（2007 年 2 月 6 日受理）

Hiromi AKUTSU, Masakazu SUGAWARA

Methodological Issue of Faculty Assessments
— On Looking at the Self-assessments of Research Activities —

1. はじめに

評価という言葉は，近年の流行語になっている観がある。あたかも日本中の企業，病院，学校，公官庁等いたるところで評価が行われているような印象を受ける。自己評価もしくは第三者による評価が行われ，その結果が公表されたり昇給に影響したりする。給料や世間の評判が関わるとなれば，多くの人々は自分と自分が所属する組織の評価に無関心ではいられなくなる一方，評価者・被評価者双方とも評価に関する十分な知識を持ち合わせていないのが現状である。これは，一群のデータの扱いについて，多くの人は平均値の比較くらいしか知らないことが多いのと良く似た事情である。科学データの分析は，しかるべき知識と技術を持った専門家のみが正しく行えるということは，多くの人に理解されているかもしれない。しかし，評価もまた科学であり，正しい知識と技術を持った専門家が，これを正しく行えるということは，多くの人に理解されているとは言いがたい。評価の素人でも，大学人が 10 数人集まれば，合理的な評価を実行できる，と考える人が大学教員の中にもいるかもしれない。筆者たちは，数量的データを長年扱ってきた心理学者であり，評価の専門家ではないが，評価の科学は心理学の一分野でもあるので，ささやかながら評価の知識と技術をもっている。そのささやかな知識と技術を応用して，岩手大学教育学部の教員評価の一部門である「研究活動の評価」を評価した。

2. 報告された評価の問題点

自己評価の実態を見てみよう。現在，全国の国立大学法人○○大学では（1）教育活動，（2）研究活動，（3）社会貢献活動，（4）大学運営活動の 4 部門で，教員が自己評価を行うことが義務付けられている。2006 年に第一回目の自己評価の結果が報告され，教育学部は（2）の研究活動において，評価の基準をゆるく設定しているのではないか，という批判が岩手大学内の教育学部以外の教員の間で出た。この批判は正しいのか，もし正しければ評価方法のどこに問題があるのか，修正するにはどうすべきかを考えてみたい。

まず，どのような評価の数字が提出されたかを見る。2004 年度（平成 16 年度）からはじめよう。

全体の傾向を理解し易いように，図 1 に各評価のランク（評点）に入る教員数を全教員数で除した比率を示す。同じデータを人数で表したものが表 1 である。2004 年度の研究活動の評価は，半数以上の教員が評点 5 としたのである。平均評点

＊岩手大学教育学部

は 4.29，であり，評点の中央値は 5 である。評点が 1 から 5 まで 1 刻みで設定され，89 人の教員が報告した自己評価の中央値（評点が 89 人中 45 番目の人）が 5 であった。評価結果を額面どおりに受け取れば，見事な研究活動ぶりといえる。2005 年度（平成 18 年度）の資料は 2004 年度と似ている（図 2 と表 2 参照）。この年も，興味深いことに半数以上の教員が評点 5 としたのである。同じ基準と方法を採用するならば，次年度以降も同様な結果が得られると予想できる。

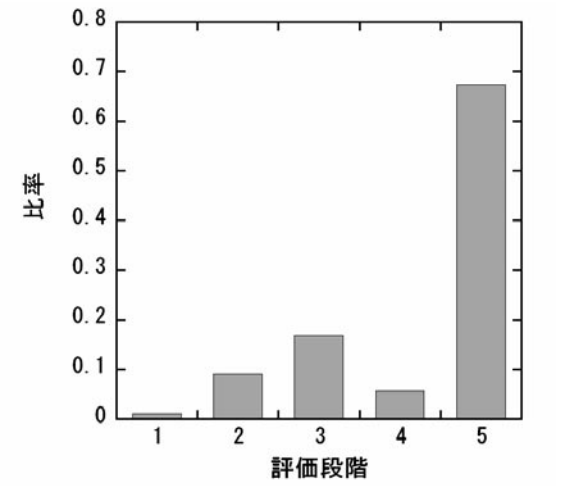

図 1．2004 年度教育学部教員の研究活動評価.
評価は 5 段階で行われ，評点 5 がもっとも優秀という評価である。各評点を持つ教員数を全教員数で除して，比率として表した。評点 5 が著しく高い比率であることに気がつく。

表 1　評点に対応する人数（2004 年度）

評点	1	2	3	4	5
人数	1	8	15	5	60

評点は研究活動を評価する指標であるが，図 1 と 2 に見るように，実際に得られた評点は，値の間隔（刻み値）が不適切なため十分役に立たない。（例えば，もしすべての人が同じ評点を得たのでは，その評点は役に立たない。）最高ランクの評点 5 に全体の約 67％（2004 年度）もしくは，約 65％（2005 年度）が含まれている。ランクの弁別的機能から見ると，半数以上の人が 1 つのランクに含まれるのは望ましいランク設定ではない。

表 1 と 2 に見られる 5 段階の評点は，自己採点（以下採点と呼ぶ）を表 3 に示す基準に照らして変換したものである。すなわち，評点＝F（採点）の関係があり，ここで F（　）は，採点から評点を求めるための関数である。ここでは採点評価関数と呼ぶ。採点が 1.4 以下ならば評点は 1 ～ 3 のどれかになり，採点が 2.0 以上ならば評点 5 が与えられる。

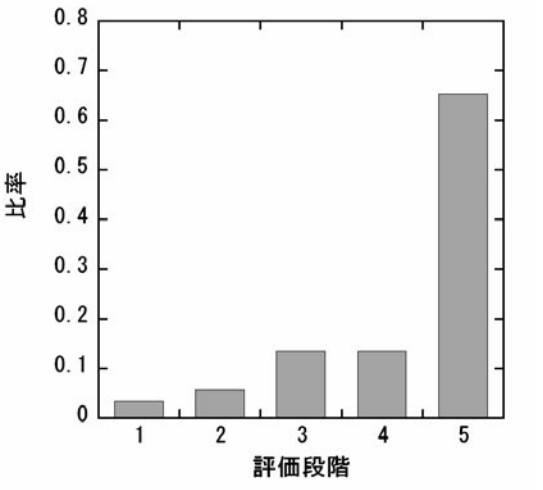

図 2．2005 年度教育学部教員の研究活動評価 .
評価は 5 段階で行われ，評点 5 がもっとも優秀という評価である。各評点を持つ教員数を全教員数で除して , 比率として表した。2004 年度と同じく評点 5 が著しく高い比率であることに気がつく。

表 2　評点に対応する人数（2005 年度）

評点	1	2	3	4	5
人数	3	5	12	12	58

表 3　評点の基準

評　点	採点
評点 5	2.0 以上
評点 4	1.5 ～ 1.9
評点 3	0.5 ～ 1.4
評点 2	0.2 ～ 0.4
評点 1	0.0 ～ 0.1

3. 採点の検討

評価過程のどこに問題があるのかを詳細に検討するため，報告された採点（0 から 25.7 まで分布していた）を調べた。以下の図に採点の度数分布を示す。

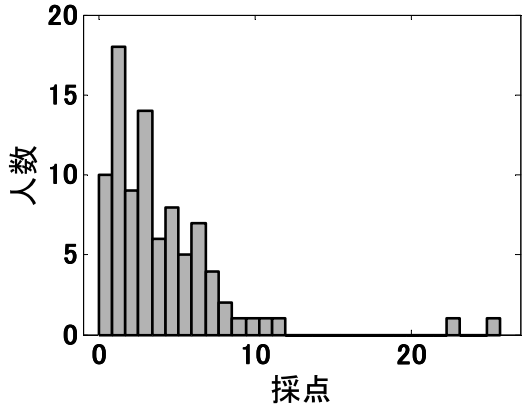

図３．2004 年度の採点の分布を示す。まとまった群から離れて２つ大きな採点値が見られる。10 以下の採点は，おおまかに見て点数が高くなると人数が減少している。

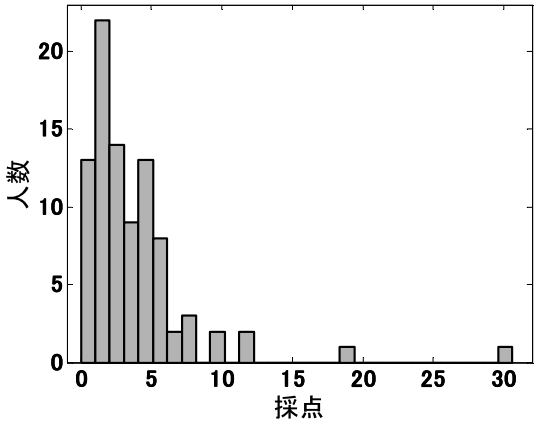

図４　2005 年の採点の分布。外れ値が４つありそうに見える。計算で４つの外れ値が確認された。分布の傾向は 2004 年度と同じで，採点が大きくなると，人数が少なくなる。

採点は，０から 25.7 まで分布した。図３で採点が２の点（評点４と５の境）を探すと，それは０に近いところにあることが見て取れる。採点がこの基準値の 10 倍の 20 を超える教員が２名いる。（この２人の採点は統計上，外れ値（outlier）とよばれ，通常の分析から除外されることが多い。）実際に外れ値を計算すると，もう一人 11.8 の採点を持つものがいる（注*）。採点が５を超える教員は，外れ値を除外しても 22 名いる（中央のランクである評点３の下位得点限度が 0.5 であるから，その 10 倍の高得点者が 22 名いる）。教育学部は非常に優れた研究活動を行う教員から構成されていると推測される。このような優秀な人たちに当てはめるには，評点基準がやや低すぎたようである。

図３の度数分布は正規分布とは異なる分布を示しているが，横軸が採点の素点であるから，これ自体は不都合でもないし不自然でもない。ついでに，2005 年度の採点の分布も調べた（図４参照）。図３と同じような分布の傾向が確かめられた。加えて，この年も特別優秀な人が数人いる。

4. 採点評価関数の検討

優秀な人たちに当てはめるには，評点基準がやや低すぎた，と先に述べたが，表３に示した評価の基準はどのような関数形をしているのであろうか。表３の数値を使って，関数形を図５に表した。

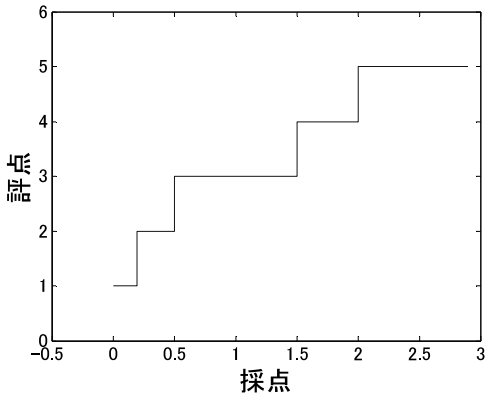

図５．2004 年と 2005 年の評価に使用された採点評価関数。横軸の採点を縦軸の評点に変換する階段型の関数である。横軸は 3.0 が最大値であるが，実際は 2.2 以上はすべて評点５として図の右に延長される。

この図５を眺めると，0.5 から 1.5 の間が 2.0 以上を除いた他の区間より広いことに気がつく。製作者の意図は分からないが，あたかも，採点の上限を 2.2 前後に予想し，採点が一様分布（どの採点区間でも頻度が同じくなる）もしくは採点の 0.5 から 1.5 の区間で最大頻度が得られる正規分布が得られれば，評点の分布も正規分布に近くなることを考慮して，作成されたかのようである。しかし，

注*　outlier は、適切な範囲から逸脱したデータポイントとして定義される。適切な範囲は、75 パーセンタイル値と 25 パーセンタイル値の差を 1.5 倍した値を 75 パーセンタイル値の上と 25 パーセンタイル値の下に設定した範囲である。この範囲から外れるデータが outlier である（Devore & Peck, 1986）。筆者たちがデータの分析に使用した MATLAB Statistics Toolbox（Math work）も同じ定義を採用する。

厳密な対称性は考慮されていない。これは評点２と４に対応する採点の区間の広さが異なること，および，評点１と５に対応する区間の広さが異なることからわかる。もっとも，対称性は適切な採点評価関数の必要条件ではない。それはともかく，この採点評価関数から明確な評価の原理（基準値決定の理由）を推測することは難しい。（この問題は，付録で再度考察する。）

実際には，図３と４に簡単化して示された細かく刻まれた（本来は連続的）段階を持つ広い範囲の採点が図５の階段型関数を介して５段階の評点に換算されるのであるから，階段型関数の階段が上昇する境界値を慎重に決めなければならない。階段の境界値が適切に設定されないと，（おそらく）製作者の意図とは異なり，今回公表された評点５が半数以上を占めるような結果になる。採点２以下に区間が狭い階段がいくつもありながら，採点２以上の領域にさらに上に登る階段がないためである。

境界値が不適当に設定された関数が役に立たないことは理解できたが，この関数を修正するにはどうしたらよいか。図５に似たような階段型関数を合理的に分布の全体にわたるように作ることが１つの方法である。しかし，2004 年度と 2005 年度の評価のようにあらかじめ境界値を決定しても，得られた採点が予想からずれていたならばどうするのか。データの分布にあわせて，自動的に境界値を変更する適応的な採点評価関数を作ることが１つの解決策であろう。（しかし，より重要なことは，階段の境界値を特定の値に決める，評価上の理由を明確にすることである。）

5. 合理的な評価の方法

もう少しうまい方法は，採点の度数分布に戻って，採点の分布が正規分布するように採点の値を変換し，その後に明確な評価の原理に基づいて決定された採点評価関数を適用することである。最初の正規分布するような変換は，心理学では標準化のための変換として知られよく使われている。具体的には，まず，採点の度数分布（ x で表す）から累積度数分布を求め，そのパーセンタイル得点（これは全体を 100 とした場合の面積にあたる）に対応する正規分布における z 値を求める。x' =F（x），ここで，F（　）は累積度数関数である。次に y=G（x）の関数を標準得点に当てはめる。ここで，G（　）は正規分布の逆累積分布関数である。このようにして変換された採点を標準得点とよぶ。2004 年と 2005 年の採点（図３と４）を標準得点に変換したデータの度数分布を図６に示す。標準得点は，扱い易いように平均 50，標準偏差 10 に z 値を線形変換する場合が多い。ここでも，平均 50，標準偏差 10 の標準得点を使う。

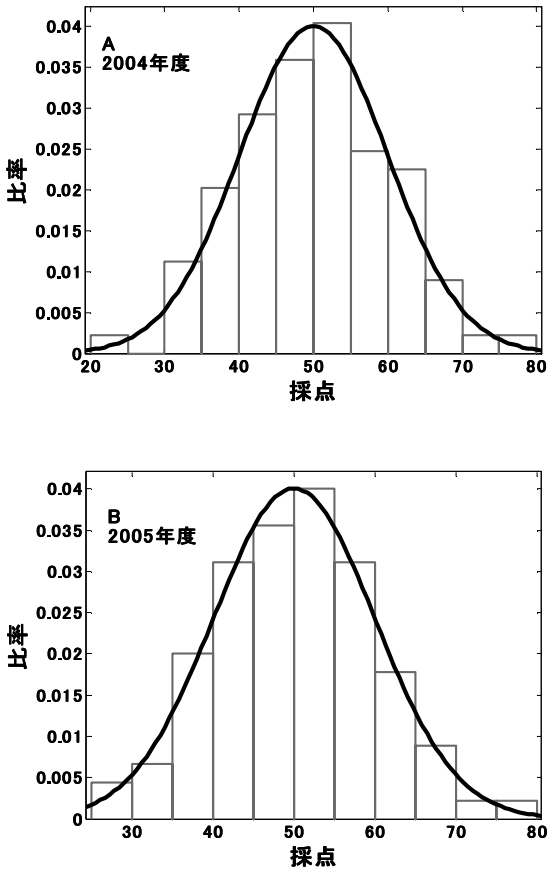

図６．標準得点の度数分布。　A が 2004 年度，B が 2005 年度である。採点値（図２と４）の度数が正規分布するように変換されている。横軸は平均 50，標準偏差 10 の標準得点である。この分布図は標準得点の相対度数（縦軸）を表し，重ねて描いた曲線はこの相対度数に適合する正規分布曲線である。正規分布が良く適合することが分かる。

評価の作業は，測定と評価の２段階からなる。測定とは，対象に合理的に数値を割り当てる作業を指す。本研究の文脈で言えば，各教員が自己採

点して提出したデータを適切に変換して望ましいデータ特性を持つよう加工する過程までが測定である。図６に示す標準得点は，心理学でいうところの「等間隔尺度」の特性を持つ。すなわち，60は30の２倍であり，40から50までの間隔は，50から60までの間隔に等しい尺度によって採点が表現されている。このような尺度値には種々の統計的分析を適用することができるので，好ましい測定値である。次の段階は，これらの測定値をある目的に合わせて解釈する過程（評価値の割り振り）である。

教員の研究活動評価の目的は，研究活動を適切にランク付けることであると仮定しよう。そのような場合，評価の第２段階は操作上，図６に示される標準得点に適当な採点評価関数を当てはめて評点（評定値）を作成することである。評価の原則を決定し，それに従って適当な採点評価関数を作成する。研究活動が中程度の人は，たくさんいて甲乙つけがたいが，特別優れた人と，特別劣る人は比較的少人数で，他から区別できる，という原則を採用したとしよう。この場合は，例えば，平均から上下に0.4標準偏差内にある採点に評点３を与え，その外側に平均から± 1.2標準偏差隔たった採点に下の方は評点２，上のほうは評点４を与え，さらにこれらの外側にある採点値にそれぞれ評点１と５を与える。この基準を採用すると，評点３を得る教員は全体の28％，評点４か５を得る教員は，全体の33.7％となる。表４に２つの異なる境界値を設定した採点評価関数を2004年度の標準得点に適用した結果を示す。表４中の人数abともに比較的適切な分布を示している。表４と最初の表１が同じ採点データを使用していることに注意して欲しい。表１の好ましくない分布は，データが悪かったのではなく（教員の自己採点が悪かったのではなく），データの処理が誤っていたためである。

ほかにも様々な合理的な境界値をもつ階段型関数を作れる。どのような関数が適切かは，評価の目的によって決まるが，この関数を決めることは本研究の範囲を超えている。いずれにせよ，測定が正しくなされているならば，評価が甘いとか，辛いとかの批判は，採点評価関数の誤った選択を指すことになる。そして，どのような採点評価関数を適用するかに教育学部の評価の方針が現れるのである。

評価過程を素点から，評価の出力までの流れを簡略に示すと次のようになる。

採点（素点）　=>　採点の標準化　=>
採点評価関数　=>　評点

表４　2004年度の採点に異なる２つの採点評価関数を当てはめた結果

評点	1	2	3	4	5
人数ａ	11	23	25	18	12
人数ｂ	6	9	22	36	16

注）人数ａは評点の境を平均から［－1.2　－0.4　0.4　1.2］倍の標準偏差隔たったところに置いた場合である。人数ｂは評点の境を平均から［－1.6　－1.0　－0.2　1.0］倍の標準偏差隔たったところに置いた場合である。

6. 他学部の評価と比較

教育学部の評価と他の学部の評価は，どのように異なるのであろうか。公表された教員評価結果に現れた数字にもとづいて，教育学部を除くほかの学部の評価と教育学部（2004年度）の評価をグラフにして図７に示す。教育学部と教育学部を除くほかの学部の間には，評価３，４，５の比率に

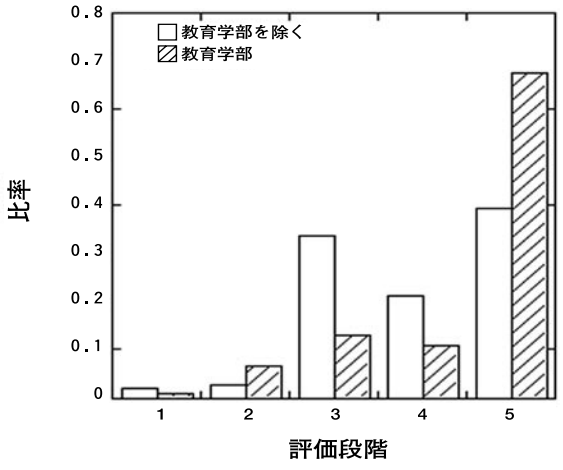

図７．教育学部を除く全学の評価と教育学部の評価。比較を容易にするために，縦軸は比率を用いる。

違いが見える。確かに，教育学部では評価５の比率が他の学部の評価（ここでは，教育学部を除く学部の評価の合計から定義する）に比べて大きい。

もし，他学部の評価に近いほうが好ましいのであれば，そのような修正は容易にできる。教育学部を除く全学の評価の分布に一致するように，採点評価関数を設定すればよいのである。このようにして計算した評価ランクに含まれる教育学部の教員数を，もとの報告の数値とともに表５に示す。

表５．2004 年度の報告された評価と計算しなおした評価

評点	1	2	3	4	5
人数O	1	8	15	5	60
人数M	1	5	30	18	35

人数Ｏは報告された評価。人数Ｍは教育学部を除く全学の評価の分布に一致するように，採点評価関数を設定して，計算した評価の人数を表す。上の図７と比較を容易にするために，表５の人数を比率に変えて図８に示す。図８の計算し直したグラフ（教育学部（新））が図７の教育学部を除くグラフとほぼ同じであることが見て取れる。評価１と２における両者間の違いは，小さな数字を扱う場合の誤差である（0.05 以下であることに注意）。評点１～５の各ランクを分ける境界値は［29.86　33.55　47.16　52.61］である。例えば，評点５を与えられるのは，標準得点が 52.61 を越える人である。（平均が 50，標準偏差が 10 であるから，これも高い基準とはいえない。教育学部以外の学部も平均するとその程度の基準なのである[注2]。）注意すべきは，表５の人数Ｍの背後にある評価の原理は，他学部と似たような評点の分布を作ることであり，それ以外の原理はない。この分布表を作成した目的は，教育学部の教員の自己採点をもとに他学部と似たような評点の分布の作成が可能であることを示すことであり，別の具体的な評価の原理を適用することではない。

[注2] 他の学部の評価の方法はわからないが、もし、評点を決める前に何らかの採点があるとしたら、その採点を標準点化すれば、ここに示したものと同じ採点評価点関数が適用できる。このような明白な操作を行わなかったにしても、平均以下に３つのランクを設け、平均をまたぐ範囲に４つ目のランクを設け、平均より少し上以上をすべて評価５としている事実に変わりはない。教育学部ほどではないが、ゆるい基準といえる。

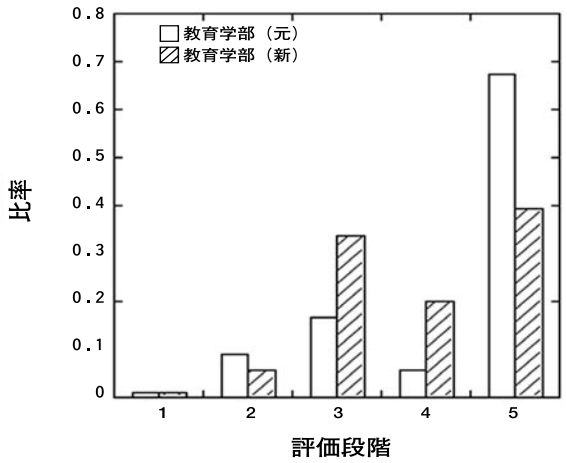

図８．2004 年度教育学部の報告された評価と新たに計算した評価。教育学部（元）が報告された評価であり，教育学部（新）が新たに計算した評価である。教育学部（新）の分布が，図７に示す教育学を除く評価の分布とほとんど同じことに注意してほしい。

7. 残された問題

なお，現行の研究活動に関する教員評価でもっとも重大な欠陥は，研究成果が長年の努力の蓄積によって得られる研究領域（例えば，文学，数学，歴史学，哲学，栽培育種学等々）が適切に評価されないことである。この問題は，研究分野が異なる教員の研究活動を比較するための適切な方法を発見するという課題に発展する。この課題を解決して初めて，研究分野が異なる教員の 研究活動（創作活動も含む）を１つの尺度の上に正当に位置づけることができる。論文の数に偏らずに学問を評価できる評価システムを創出する必要がある。教員評価の前途には解決しなければならない多くの課題がある。

8. まとめ

2004 年度と 2005 年度の教育学部教員の研究活動の評価は，用いられた評価のランク（１～５）が半数以上の教員の研究活動を弁別できなか

った。評点のランク５が教育学部全教員の 50％以上を含んでいたからである。評価の基準がゆるいという批判は妥当であろう。この問題は，評価の過程において，（１）教員の自己採点の素点をそのまま測定値とした，（２）不適切な採点評価関数を適用した，という２重の欠陥に由来する。そのために，評点ランク（１～５）が分布上著しい不均衡を生じた。2004 年の採点データを使って，教育学部の除いた全学の評価の分布と類似した評価の分布ができることを示した。2004 年度と 2005 年度の採点と評価の分析から，評価過程に何らかの改善処置をとらない限り，今後も同じ不適切な評価が繰り返されると予想できる。

引用文献

Devore, J. and Peck, R. 1986. *Statistics*. West publishing company, St. Paul, MN, USA.

池田央．1992．テストの科学．日本文化科学社．

辰野千尋．2003．絶対評価と相対評価．2003 年度改訂版教育評価法概説 . 原著者　橋本重治．改訂編集　応用教育研究所 . 国書文化

Thorndike, E. L., 1918. The Nature, Purpose and General Methods of Measurement of Educational Products. Seventeenth Yearbook of the National Society for the Study of Education, Part Ⅱ .

MATLAB 2004　The MathWorks, Inc. Natick, MA, USA.

付録

相対評価と絶対評価について

本研究が提案する評点決定法は，あらかじめ評点ランクに含まれる教員数を決めているので相対評価ではないか。このような相対評価は，教員の研究活動を評価する方法としては，不適切ではないか，という疑問を持つ人がいるであろう。そこで，本研究に関わる範囲内で相対評価と絶対評価を考察する。

（１）相対評価とは

相対評価は，次のことを意味するように使われている。評点（例えば１～５）の各クラスに入る割合（相対度数）をあらかじめ決めた評価法。評価を行う集団に照らして測定結果を解釈する。集団基準準拠解釈とも呼ばれる（辰野，2003）。

（２）絶対評価とは

相対評価に比べと絶対評価の意味は多義的である。文脈により次の３つの事柄をさす。①評点（例えば１～５）の各クラスに入る割合（相対度数）をあらかじめ決めない評価法。②ある目標がどの程度達成されたかを示す評価。日本ではしばしば達成度評価と呼ばれ，アメリカでは目標基準準拠解釈と呼ばれる（辰野，2003）。③物理測定における CGS 単位のような度量衡に基づく評価（池田, 1992）。ソーンダイク（1918）以来，多くの心理学者にとって絶対評価とは③を意味してきた。これら３つのうち，日常の会話と議論でもっとも良く使われるのは①であるが，教育評価の教科書を見ると絶対評価に①の意味は含まれない。

相対評価と絶対評価については，（１）相対評価と絶対評価のどちらが望ましいか，（２）絶対評価は可能か，という難しい問題が付きまとうが，ここではそれには触れない。

さて，本研究が提案する評点決定法は相対評価なのか絶対評価なのか。絶対評価を③の意味に解釈すれば，われわれの方法は絶対評価ではない。CGS 単位に匹敵する測定に基づく③の意味の絶対評価はまだ実現されていない。われわれが提案した標準得点の決定法は，集団における採点の分布に基礎を置き，CGS 単位とは原理的に異なる。それでは②の意味の絶対評価か。これも違う。②の意味の絶対評価を行うためには，研究活動の目標が明確にされ，その達成の度合いが操作的に定義されねばならない。例えば，Nature 誌に論文が掲載されたら，目標を 100％達成した，というぐあいに「もし x の条件を満たせば，y ％達成されたと決める」が明示されなければならない。教員の研究活動の評価方法にこれらに該当する記述は含まれていない。

それでは，①の絶対評価はどうか。①をもう一度みると，評点（例えば１～５）の各クラスに入る割合（相対度数）をあらかじめ決めない評価法とある。われわれが提案する方法は，最初に集団の得点分布を使って，標準得点を求める。次に，標準得点を使って，評価クラスの範囲を決める。各クラスに入る割合は採点評価関数を決定したときにすでに決まっている。したがって，われわれが提案する方法は，①の擬似絶対評価ではなく，その演算操作は上述した相対的評価（集団基準準拠の評価）の定義に十分適合する。

これまでの議論から，③の絶対評価は実行不可能であり，②の絶対評価には周到な準備が必要であると理解できる。唯一実行可能な絶対評価は，①の分類区分に入る相対度数をあらかじめ決めない擬似絶対評価である。2006 年に岩手大学の教育学部

で実施された評価は，（1）採点の分布を考慮せず，（2）採点評価関数を直接採点の素点に適用する，ことにより擬似絶対評価を用いたと解釈できる。

擬似絶対評価の利点は，他の教員の採点に影響されずに，各教員を独立に評価できる点である。同じ学部内に優秀な教員が多数いても，あるは反対に活動が低い教員が多数いても，その事実は各教員の評点にはなんら影響を与えない。これは，もし正当な適用ができるならば，優れた特性である。この擬似絶対評価が正当な評価として備えるべき条件は，（1）評価の弁別力を持つこと。言い換えれば，「活動の低い人」から「活動の高い人」の間をいくつかの段階で適切に区別できること。（2）長期間の使用に耐えられる。（3）評価結果の解釈に誤解を生じにくい。以上の3点であろう。この条件は，われわれが提案した評価方法にもあてはまる。2006年に教育学部で実施された評価が，(1)の条件を満たしていないことはすでに指摘した。相対評価の1つの欠点は，他の教員の採点が最終的な評価に影響する点である。例えば，全員の採点が次の年に2倍に跳ね上がったとしよう。（極端で，現実的ではないが，長期的には1.5倍くらいに上昇する可能性はあろう。）相対評価を使うと，前年に1の人は次の年も，活動量が2倍に上昇したにも関わらず，1の評価を受ける。前年4の人は，次の年も4の評価を受ける。これは確かに正当な評価ではない。しかし，この欠点は相対評価においても修正可能であろう。われわれが提案する方法に沿った1つの解決策は，採点評価関数に評点区分の境界値を適応的に移動するパラメータを導入することである。例えば，採点素点の中央値を全体の活動水準の指標とする。この指標を，採点評価関数のパラメータとして入力し，この値に従って評点区分の境界値が評価年度間で移動する。予想できる全員の採点の変動内で，多少の数値シュミレーションを行えば，活動水準パラメータと評点境界値の関係を適切に決めることができよう。その結果，全員の活動水準を考慮した相対的評価が可能となる。しかしながら，同じ年度においては，他の教員の採点が最終的な評価に影響する点は，変更できない。理由は明白で，相対評価は全体の分布を考慮に入れて，評点を割り振る過程が必須だからである。

謝辞

本研究は、2006年度岩手大学学長裁量経費（教育研究支援経費）の補助金を受けて行われた。記して謝意を表したい。