

多数の欠測値を含む相互評価データにおける 評価者の評価特性補正方法[†]

藤原康宏^{*1,*3}・大西仁^{*2,*3}・加藤浩^{*2,*3}

岩手県立大学ソフトウェア情報学部^{*1}・メディア教育開発センター^{*2}

総合研究大学院大学文化科学研究科^{*3}

学習者間の相互評価で公平なフィードバックを行うために、評価者が実際につけた評価値を、個々の評価者の評価特性を考慮して補正する方法を提案する。相互評価では、通常、全ての学習者が他の全ての学習者と相互に評価し合うことは困難であり、非常に欠測値の多いデータを扱うことになる。提案する方法では、項目応答理論のメタファを用いた評価モデルにより各評価者の評価特性を推定し、平均的な評価特性の評価者による評価値との差を用いて、評価値の補正を行う。提案した方法を実データに適用して、補正を行うことにより妥当な評価値に近づくことを確認した。さらに、実際の授業において、5回にわたって相互評価を実施し、評価結果を評価者にフィードバックしたところ、補正值をフィードバックされた評価者は評価能力が向上したが、補正前の評価値のみをフィードバックされた評価者は評価能力が向上しなかった。これらの結果から、提案した補正方法は、公平な評価、学習者の評価能力の向上に有効であることが示された。

キーワード：相互評価，教育評価，評価特性，教育データ，補正

1. はじめに

教育評価における相互評価とは、「学級集団などにおいて、生徒同士に、互いに級友を評価させる方法」（東ほか 1988）をいう。植野（2005）は、学習者同士で評価を行うことの利点として「学習者間の評価は、最も自然な評価体系であり、評価結果を受け入れやすくし、学習者の内省を引き出し、結果として知識の内化や深化を導くことや、学習動機の上昇が期待される」と述べている。相互評価の教育的効果を高めるために

は、学習者が評価結果を受け入れやすくするためのフィードバックが重要である。

一方、評価者によって、評定の甘さや厳しさなど個人差があるといわれている（渡部ほか 1998）。全ての学習者が他の全ての学習者を評価する場合は、ほぼ公平であるといえる。しかし、評価者を選択しなければならない状況では、たまたま甘い評価者にあたったのか、厳しい評価者にあたったのかによって、学習者に不公平感を与え、評価結果を納得して受け入れることが阻害されると考えられる。したがって、公平な評価を行うためには評価者の評価特性を考慮したフィードバックを行うことが必要である。

植野ら（2006）は、e-Learning で得られた大量の多段階評価の相互評価のデータを項目応答理論の一つである Graded Item Response Model (SAMEJIMA 1969) に評価者の特性パラメータを加えるという拡張をして、レポートの評価を行う方法を提案している。それぞれのレポートに対して同一尺度上で合理的な評価を行うことができ、それぞれの評価者の特性として、評価基準の厳しさ、評価の一貫性を求めることができることが特徴としてあげられる。

評価する学習者が、評価対象となっている学習者か

2007年2月19日受理

[†] Yasuhiro FUJHARA^{*1,*3}, Hitoshi OHNISHI^{*2,*3} and Hiroshi KATO^{*2,*3}: The Development of the Method to Revise an Evaluated Score in Considering of the Evaluator's Characteristic for a Peer Evaluation with Many Missing Responses

^{*1} Iwate Prefectural University, 152-52, Takizawa-Sugo, Takizawa-mura, Iwate, 020-0193 Japan

^{*2} National Institute of Multimedia Education, 2-12, Wakaba, Mihama-ku, Chiba-shi, Chiba, 261-0014 Japan

^{*3} The Graduate University for Advanced Studies, 2-12, Wakaba, Mihama-ku, Chiba-shi, Chiba, 261-0014 Japan

らも評価される場合、評価が甘くなる傾向があり（お互い様効果、藤原・加藤 2005）、お互い様効果を避けるような組み合わせで相互評価を行うと、一人の学習者ができるだけ多くの評価をした場合でも、データの半分は欠測値となる。一般的に、一人の学習者が行うことができる評価人数は限られており、その結果、一人の学習者が受ける評価数も多くはなく、このような欠測値の多い場合でも、適用できる補正方法が必要である。そこで、項目応答理論のメタファを用いた評価モデルにより各評価者の評価特性を推定し、平均的な評価特性の評価者による評価値との差を用いて、評価値の補正を行う。提案したモデルを実データに適用しその精度を示すとともに、教員の行う評価と比較することで補正方法の妥当性を示した。ただし、本研究は植野ら（2006）などが行っているような項目応答理論を相互評価に当てはめたものではなく、項目応答理論のメタファを用いた評価者の特性を反映したロジスティック回帰による簡易法を提案するものである。植野ら（2006）などの項目応答理論に対して、計算が非常に簡易であることが特徴である。

さらに、実際の授業において、5回にわたって相互評価を実施し、評価結果を評価者にフィードバックしたところ、補正值をフィードバックされた評価者は評価能力が向上したが、補正前の評価値のみをフィードバックされた評価者は評価能力が向上しなかった。

2. 補正方法

ここでは、評価者が実際につけた評価値を、その評価者の評価特性に基づいて補正する方法を提案する。

まず、同じレポートを評価した評価者の評価との関係から、評価者の特性を推定する。次に評価者の特性と平均的な評価との差を使って、実際に行った評価の補正を行う。

2.1. 評価特性のモデル化

評価者の評価特性を、項目応答理論（芝 1991）でよく用いられるモデルの1つである2パラメータ・ロジスティックモデルを元に表現する。項目応答理論は、テスト項目への学習者の応答と、現実には観測されない想定尺度上の学習者の能力を結びつける数理モデルである。大量の正誤データから、それぞれのテスト項目の特性を表すパラメータを求めることによって、学習者の推定された能力から、各テスト項目に正答する確率を求めることができる。2パラメータ・ロジスティックモデルでは、 θ_j の能力を持った学習者jが、項

目iに対してXの反応をする確率は、式(1)で表される。

$$P(X | \theta_j) = \frac{\exp\{Da_i(\theta_j - b_i)\}}{1 + \exp\{Da_i(\theta_j - b_i)\}} \quad (1)$$

ただし、 a_i, b_i はテスト項目の特性、Dは定数。式(1)では、個々のテスト項目の特性を、項目識別力 a_i 、項目困難度 b_i の2つのパラメータで表している。項目困難度はその数値が大きいほど項目が難しいことを意味し、項目識別力は値が大きいほど、項目困難度の付近での項目応答曲線の傾きが大きいことを意味し、その付近での学習者の能力を識別する力が大きいといえる。

ここで、項目反応理論のメタファとして、以下のようなモデルを考える。ただし、項目応答理論では潜在変数モデルを仮定しているのに対し、本提案では簡易的に以下のような回帰モデルにより代替するものである。項目応答理論におけるテスト項目が持つ特性を評価者が持つ特性に置き換える。平均的な評価者が、学習者jのレポートに対して、式(2)のように評価者の特性に基づいて $R_0(\theta_j)$ ($0 \leq R_0(\theta_j) \leq 1$) の評価をつけるものと仮定する。

$$R_0(\theta_j) = \frac{\exp\{Da_0(\theta_j - b_0)\}}{1 + \exp\{Da_0(\theta_j - b_0)\}} \quad (2)$$

この時、式(2)におけるaは評価に差をつける度合いを、bは評価の甘さの度合いを表す。学習者jのこのレポートでの真のできぐあい θ_j は式(3)で表される。

$$\theta_j = \frac{1}{Da_0} \log \frac{R_0(\theta_j)}{1 - R_0(\theta_j)} + b_0 \quad (3)$$

評価者iが学習者jのレポートに対して行う評価 $R_i(\theta_j)$ は、式(2)と同様にして式(4)で表されると仮定する。

$$R_i(\theta_j) = \frac{\exp\{Da_i(\theta_j - b_i)\}}{1 + \exp\{Da_i(\theta_j - b_i)\}} \quad (4)$$

式(4)に対して、式(3)を代入すると式(5)を得る。

$$R_i(\theta_j) = \frac{1}{1 + \exp\{-Da_i(b_i - b_0)\} \left\{ \frac{1 - R_0(\theta_j)}{R_0(\theta_j)} \right\}^{\frac{a_i}{a_0}}} \quad (5)$$

評価者iがレポートjに対して、 $R_i(\theta_j)$ の評価を行ったとき、 $R_0(\theta_j)$ を同じレポートjを評価した他の評価者の評価の平均を \bar{x}_j と置くことで、平均的な評価と評価者iとの評価の関係を表すことができる。式(5)に対して、パラメータを計算が容易になるように、 $D=1.7$, a_0

$=1$, $\alpha_i = a_i$, $\beta_i = b_i - b_0$ と置くと, $R_i(\theta_j)$ は, \bar{x}_j とパラメータ α_i , β_i から式(6)のように表される. 以降, この式を評価者特性モデルという.

$$f(\bar{x}_j; \alpha_i, \beta_i) = \frac{1}{1 + \exp(-1.7\alpha_i\beta_i)\left(\frac{1-\bar{x}_j}{\bar{x}_j}\right)^{\alpha_i}} \quad (6)$$

このモデルと2つのパラメータの特性をグラフにしたものを図1, 2に示す. α は評価者*i*の相対的な差をつける度合い(値が大きいほど差をつける), β は相対的な評点の厳しき(値が大きいほど甘い評価をつける)を表していると考えられる.

なお, (6)式は項目応答理論のメタファを用いているが, 項目応答理論そのものではないため, 項目応答理論に見られる理論的特性をそのまま適用することはできない. しかし, 本研究は, 評価者の評価特性を集団との関係から推定し, その差異を用いて, 実際行った評価を補正することが目的であり, (6)式を用いることにより, 簡単な計算で評価者の特性を集団の中で相対的に表現することができる.

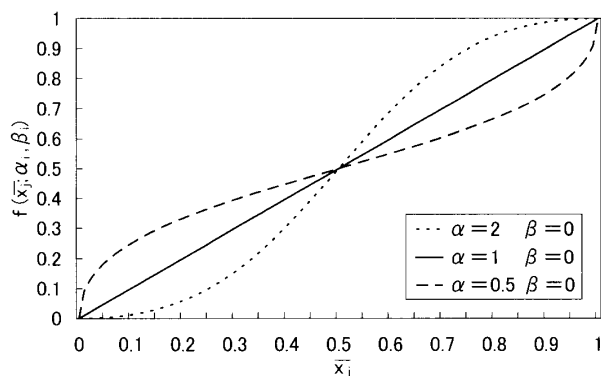


図1 モデルとパラメータ α との関係

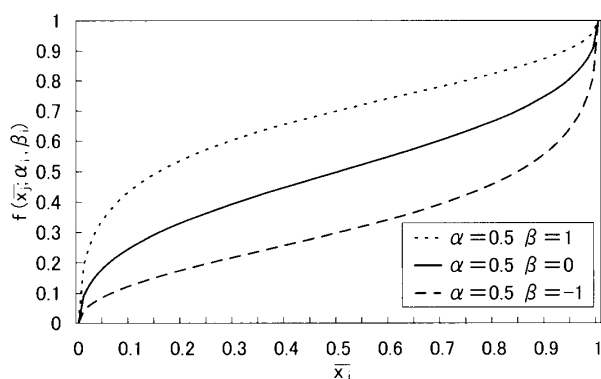


図2 モデルとパラメータ β との関係

2.2. 評価特性モデルを使った評点の補正方法

次に式(6)から求められる評価者の特性を用いて, 実

際に行った評価を補正する. 同じレポートを評価した人がつけた点数の平均 \bar{x}_j を基準にして行う. \bar{x}_j という評価を得られるレポートに対して, 式(6)を用いて i が行うと予測される評価と, 平均的な評価者が行う評価とを比較し, その差を実際に i が行った評価から減じる. 平均的な特性を持つ評価者の特性パラメータを α_0 , β_0 とすると, 評価者 i が j のレポートに対して実際に行った評価 y_{ij} に式(7)のように補正する.

$$y_{ij} - \{f(\bar{x}_j; \alpha_i, \beta_i) - f(\bar{x}_j; \alpha_0, \beta_0)\} \quad (7)$$

補正の例を図3に示す. 実線で表される評価者 i の特性曲線と点線で表される平均的な評価との差を補正量として, 実際に評価者 i の行った評価を補正する.

3. パラメータの計算方法及び精度

3.1. パラメータ推定

評価者の特性を表すパラメータの計算方法の例について述べる. パラメータは Levenberg-Marquardt 法 (MARQUARDT 1963) を用いて推定した. Levenberg-Marquardt 法は, 最急降下法と逆 Hesse 法を組み合わせた, 非線形最小二乗問題で使用される解法である. 当てはめたいモデルを $y=y(x;a)$, とすると評価関数 χ^2 は,

$$\chi^2(a) = \sum \left[\frac{y_i - y(x_i; a)}{\sigma_i} \right]^2 \quad (8)$$

となる. パラメータの初期値を a として次の手順で計算した (PRESS *et al.* 1988).

STEP1 $\chi^2(a)$ を計算する.

STEP2 λ の初期値として $\lambda = .001$ を代入する.

STEP3 連立方程式(9)(k はパラメータ)を解いて δ_k を求め $\chi^2(a + \delta_k)$ を計算する. (χ^2 の減少量が 0.1 以下の場合終了する)

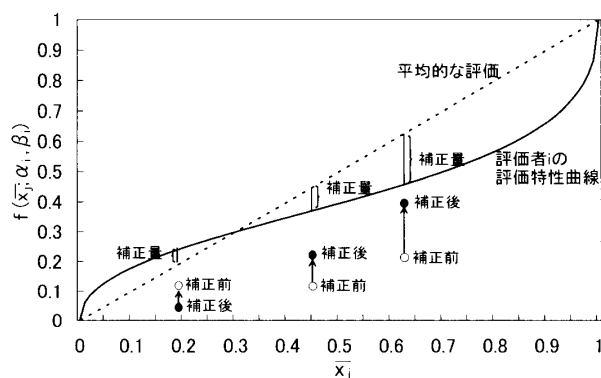


図3 補正の例

$$\alpha_{kl} = \sum \frac{1}{\sigma_i^2} \left[\frac{\partial y(x_i; a)}{\partial a_k} \frac{\partial y(x_i; a)}{\partial a_l} \right] \quad (9)$$

STEP4 $\chi^2(a + \delta_a) \geq \chi^2(a)$ ならば, λ を10倍し, STEP3に戻る.

STEP5 $\chi^2(a + \delta_a) < \chi^2(a)$ ならば, λ を1/10倍し, a を $a + \delta_a$ に更新しSTEP3に戻る.

3.2. 計算例

A大学1年生31名に対して行った『コンピュータ応用』でのレポートの相互評価を例にあげる. このレポートは, 与えられたデータを表計算ソフトを使って加工する課題で, 表計算ソフトの使い方の講義が終わった後に中間レポートとして課した. 評価項目は, 以下のとおりで, それぞれ0~10の11段階で評価させた. $R(\theta)$ の値の範囲は[0,1]であるため, 得られた評価値を x とすると, $R(\theta)=x/10$ として計算した.

- ・課題は条件を満たしているか?
- ・表の入力内容(計算式)は適切か?
- ・完成した表のレイアウトは見やすいか?
- ・適切なデータ, グラフの種類を選んでいるか?
- ・完成したグラフは見やすいか?

一人の学習者が15人の評価を行った. 相互評価は, お互い様効果(藤原・加藤 2005)を考慮して, 評価した相手からは, 評価されないような組み合わせで行った.

評価項目によって評価の特性が異なることも考えられるが, 今回は5つの評価項目が表計算ソフトの扱いに対して同じ種類の目標を扱っていると考えて, 評価項目ごとではなく, まとめてパラメータ推定を行った. 平均的な評価の特性は, 同じレポートを評価した他の14人の評価の平均を用いて計算を行った. ある評価者の行ったレポートに対する評価と, 推定されたパラメータから予測される評価特性曲線を図4に示した.

この評価者の推定されたパラメータは, $\alpha=2.16$, $\beta=-0.430$ であり, 平均より厳しい評価をしていることが分かる. 推定されたパラメータを用いて, モデルから求められた推定値と実測値の差を調べるために, 学習者ごとに平均二乗誤差(式10)を計算した.

$$e_i = \sqrt{\frac{1}{n} \sum_{j=1}^n \{y_{ij} - f(\bar{x}_j; \alpha_i, \beta_i)\}^2} \quad (10)$$

ただし, n は行った評価の回数

評価者全体の平均二乗誤差の平均は0.183であった. これは, 11段階の評価にすると平均して約2.0ずれることになる. なお, 図4に示した評価者の平均二乗誤差

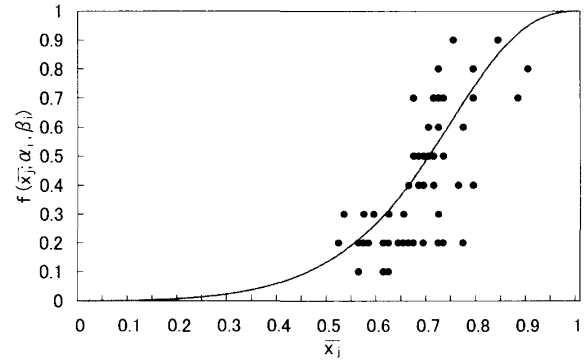


図4 推定された評価特性曲線と実測値

は0.256であった. 評価者全体の約1割にあたる4名が平均して11段階に換算すると3.0以上ずれていることから, 当てはまりが悪かった学習者に対するの検討は必要であると思われる. これについては, 「6. おわりに」で述べる.

4. 適用例

4.1. データ

B大学1年生45名に対して行った『コンピュータ入門』での相互評価のデータを用いて, 補正の効果について検証した. 課題は, プレゼンテーションソフトを使用し, アンケートデータをまとめるもので, グループワークを交えながら14回分の授業時間で作成した. 相互評価は, 15回目の授業で一度だけ行い, 一人の学習者あたり, 6人分の課題を, 次の5つの観点でそれぞれ1~5の5段階で評価させた.

- ・課題の条件を満たしているか
- ・用いられている表やグラフは適切か
- ・プレゼンテーションソフトの機能を適切に使っているか
- ・レイアウト, 色使い, 字体は見やすいか
- ・全体として内容はうまくまとめられているか

評価にあたっては, 評価した相手からは評価されず, また, グループワークを同じ班で行ったものを評価しない組み合わせとした.

4.2. 結果

学習者の推定されたパラメータの散布図を図5に示す. α の値について1.0を境に見てみると, $\alpha > 1.0$ の者は全体的に差をつけた結果, $\alpha < 1.0$ の者よりも厳しい評価をつける傾向があることが分かる. 図5の左下には離れた点が3つ存在する. これについて評価データを調べたところ, いずれもほとんどのレポートについて高い評価をつけ, 差をほとんどつけていないことが分かった.

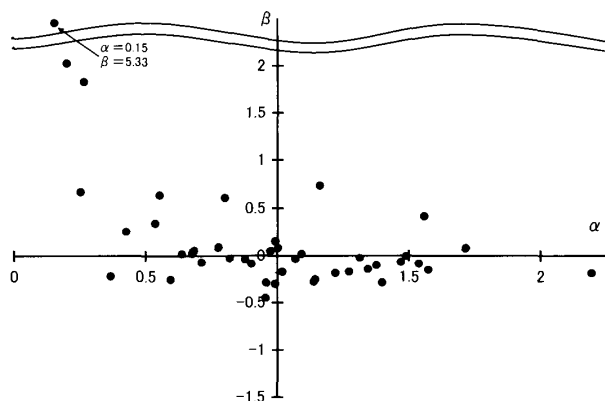


図5 推定されたパラメータの分布

次に、推定されたパラメータを用いて、モデルから求められた推定値と実測値の差を調べた。それぞれの学習者ごとの平均二乗誤差を式(10)を用いて計算し、その分布を図6に示した。平均二乗誤差が0.05の時には、5段階に変換すると1評価項目あたり0.2ずれ、同様に0.125の時は、0.5ずれる計算になる。平均誤差が0.2を超えた4名を調べると、例えば画像の使い方にこだわった評価をするなど、他の者とは異なった評価を行う傾向が見られた。全体を見ると、実用に耐える推定の精度であると思われる。このように集団とは異なった評価をするなどの理由で、推定の精度が一定値より低い学習者に対しては、補正の計算からは除外する、学習者に対して他の学習者の評価と離れていることを知らせる、教師にフィードバックを行いつどのような評価を行っているか教師がチェックし指導するなどの対処が考えられる。

次に式(7)を用いて、個人差を補正した評価値を計算した。補正前の学習者がつけた評点と教員がつけた評点との Spearman の順位相関係数（以下、相関係数）は0.303であったが、補正後の評点と教員がつけた評点との相関係数は0.427となり、適切に補正が行われていることが分かった。

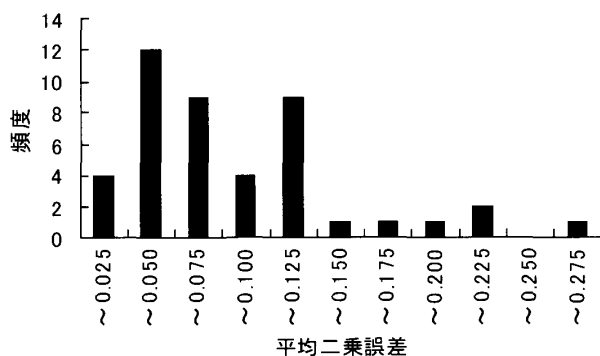


図6 推定された値と実測値との差

この結果を検証するために、別の教員に今回のレポートの評価を依頼し、補正の前後で比較した。2人の教員間の相関係数は0.566であった。追加で依頼した教員がつけた評点と補正前の学習者の評点との相関係数は0.129であったのに対して、補正後は0.235となり、最初の教員と同様の結果が確認された。

5. 相互評価支援システムを利用した実践例

5.1. 実践の概要

本論文で提案した補正を使った教育実践を行い、学習者の評価能力が向上するかどうかについて検証した。筆者らは、相互評価支援システムを開発しており（藤原ほか 2006）。このシステムのフィードバック画面（図7）に評価者特性に基づいて補正した評価値の表示を追加したシステムを開発し、実践を行った。

フィードバック画面では、学習者が受けた評価に対する補正後の評価と補正された評価の平均、他者に行った評価に対する補正後の評価と評価の特徴を表す評価特性パラメータを表示できるようにした。

A大学1年生16名に対して行った『情報処理入門』で、システムを使って相互評価を行った。8名については、フィードバック画面で今回開発した評価者の特性を補正した数値を表示し、その数値が評価者の特性によって補正されたものであることを提示した（補正表示群）。残り8名については補正した数値は表示せず、他人から受けた評価、他人に行った評価の値、平均点を表示した（補正非表示群）。Microsoft Office PowerPoint を使用して、自分の趣味や興味があることについて紹介するプレゼンテーション資料を作成させ、システムを使用して学習者間で評価させた。課題でのスライド数は5～9枚で、クリップアート、インターネット上の素材は自由に使ってよいこととした。課題を出す際に、次の5つの評価項目を提示し、それぞれ5段階で評価をさせた。評価は、それぞれの評価項目の設定理由を伝えたが、例えば、「～が含まれていれば、5をつける」など詳細な評価基準を示さなかった。評価値は、主観的に、悪いと感じれば1、良いと感じれば5を選択させ、できるだけその理由または改善方法を記述するように指示した。

- ・レイアウト、色づかい、字体は見やすいか？
- ・表現は分かりやすいか？
- ・情報の信頼性は高いか？
- ・見ていて楽しい気分になるか？
- ・全体としてうまくまとめられているか？

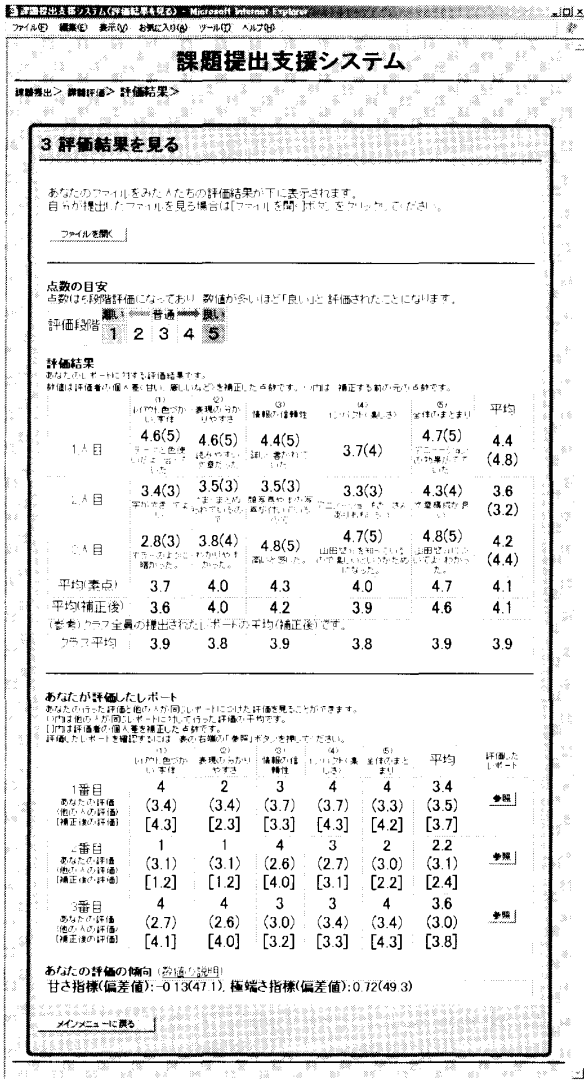


図7 フィードバック画面（補正表示群）

レポートの作成過程で相互評価を5回実施した。実践の前後で、昨年度の受講生が作成した同じテーマのレポートの評価を行わせてその変化をみた。昨年の受講生のレポートは12人分を用意し、6人ずつの2セットに分け、順序効果がでないように、それぞれの学習者に割り当て、一人の学習者は実践の前後で合わせて12人分の評価を行った。

5.2. 結果

実践の終了後に行ったアンケートの結果を表1に示す。補正非表示群にも、補正結果を表示したフィードバック画面を提示して補正についても尋ねた。

アンケートから、学習者は、評価者によって評価が甘かったり、厳しかったりすることを認識している。また、その差を不公平であると受け取る者も一定数見られた。実践で毎回補正を見て慣れてきた補正表示群では、補正することが望ましいと回答した者が多く見

表1 実践後に行ったアンケートの結果

あてはまる:5、あてはまらない:1	補正あり	補正なし
補正前の評価は、おおむね納得できた	4.13	4.50
補正後の評価は、おおむね納得できた	4.25	-
パラメータ α の意味は理解できた	3.75	3.25
パラメータ β の意味は理解できた	3.88	3.38
自分の評価の特徴に気がついた	4.25	3.38
評価者によって評価に差がある	4.25	4.38
評価者によって差があつて不公平	3.00	2.25
評価者の差を補正するほうが望ましい	3.75	2.38
学生間での評価を導入した授業を受けたい	3.88	3.00

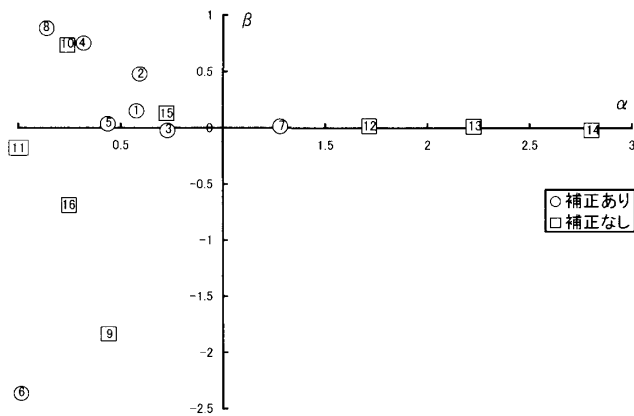
られた。補正のパラメータの意味は理解できたかという問いに、2以下と回答した者が3名見られた。自分の評価の特徴が分かったかという問いには、パラメータの意味が理解できなかった3名を除いた平均は3.92であった。補正表示群では、補正後の得点には納得できたという項目の平均は4.25であり、補正值は高く評価されていたが、補正前の得点に納得できたという項目の平均4.13よりは高かったが、補正前と補正後で有意な差はなかった。

事前、事後で昨年度の受講生のレポートについて学習者が行った評価と教員が行った評価の関係について調べた。相関係数は、事前評価では補正表示群0.260、補正非表示群0.284であったのが、事後評価では補正表示群0.431、補正非表示群0.192となった。補正結果を表示することにより、他人の行った評価をより強く意識して、適切な評価をつけられるようになったと思われる。

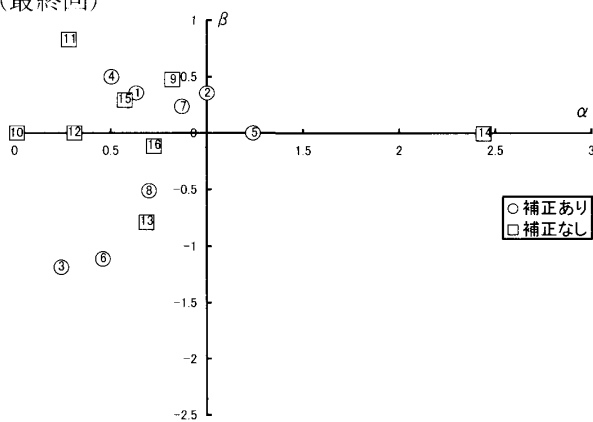
次に、評価者特性モデル(式6)を適用した結果について述べる。初回、最終回の評価のそれぞれで推定された学習者のパラメータを図8に、初回と最終回での平均二乗誤差の変化を表2に示す。

図8から、初回の評価では、評価に差をつけることができず α が低い学習者が多いことがわかる。最終回の評価では、補正表示群の学習者は α が向上していることが読み取れる。次に、フィードバック画面で補正表示を行った学習者のうち、実践前にモデルへの当てはまりがよくなかった学習者について取り上げる。学習者8は、実践前の平均二乗誤差は0.197であったのが、実践後は0.071に向上した。実践前の α は0.012であり、評価に差をつけずに、他の学習者と比べてかなり甘い評価をする傾向にあった。実践後の α は0.460に上昇し、実践前よりも評価に差をつけることができるようになった。学習者6についても、学習者8と同様に評価に差をつけることができるようになったことで、平均二乗誤差は減少した。学習者2については、実践前と実践後の平均二乗誤差からはほとんど変化が見られな

(初回)



(最終回)



図中の数字は学習者番号を表す

図8 実践前後でのパラメータの変化

表2 実践前後での誤差の変化

条件	学習者番号	実践前	実践後	差
補正あり	1	.087	.151	.064
	2	.161	.142	-.019
	3	.030	.055	.024
	4	.182	.189	.006
	5	.091	.037	-.054
	6	.149	.091	-.058
	7	.032	.101	.069
	8	.197	.071	-.126
	平均	.116	.105	-.012
補正なし	9	.236	.186	-.050
	10	.172	.190	.018
	11	.157	.226	.069
	12	.068	.122	.053
	13	.112	.134	.022
	14	.172	.165	-.007
	15	.072	.141	.069
	16	.079	.024	-.055
平均	.134	.149	.015	
全体の平均	.125	.127	.002	

かったが、学習者2は、学習者6,8と同様に評価に差をつけるようになった。学習者4については、実践前後で平均二乗誤差も変化せず、パラメータの変化もほとんど見られなかった。この学習者は、実践後のアンケートで、パラメータの理解(αとβそれぞれに対する回答の平均)については、2(あまり理解できなかった)、自分の評価の特徴に気がついたかについては3(どちらともいえない)と回答しており、この学習者については、今回の表示方法による効果が低かった。

補正表示群、補正非表示群ともに、他人の評価を見せたが、教員の評価との相関係数、パラメータ、最小二乗誤差より、補正表示群の方が補正非表示群よりも優れていた。評価がより適切になったかどうかという観点からは、補正表示の有無の要因と比較して、単に評価結果を見たことによる学習効果は、小さいと思われる。

実践後のアンケートのパラメータの理解、自分の評価の特徴への気づきと、実践前後での評価の変化について述べる。実践後のアンケートで、パラメータの理解について、4以上と回答した者と4未満と回答した者に分けて評価の変化を調べた。4以上と回答した学習者2,5,6,8はαや平均二乗誤差の変化から、評価に差をつけるようになり、評価の傾向が他の評価者に近くなった。また、アンケートで自分の評価の特徴への気づきについても同時に4以上と回答していた。4未満と回答した学習者1,3,4,7は、αや平均二乗誤差の変化から、それ以外の学習者と比較すると効果が低かった。パラメータの理解が4以上の者は、平均二乗誤差が平均0.064下がり、4未満の者は0.041上がったことから、単に補正値を表示するだけでは不十分で、自分の評価特性を理解することが必要であることが示唆される。

6. おわりに

本論文では、多数の欠測値を含む相互評価のデータにおいて、公平な評価をおこなうために、評価者が実際につけた評価値を、個々の評価者の評価特性を考慮して補正する方法を提案した。実データに適用した結果、補正方法が適切であることが確認された。また、評価特性による補正値をフィードバックする相互評価支援システムを使用した実践の結果、補正表示を行うことでより適切な評価ができるようになり、学習者も評価者の個人差を補正することを望んでいることが分かった。

今回提案した補正方法は、評価者の特性を平均的な評価者の特性との差を用いて行っているため、補正後

の評価は平均的な評価に近づく傾向がある。平均的な評価特性から離れた評価を行う評価者がいた場合、その評価が合理的であったとしても、その評価者の補正後の評価は平均的な評価に近づけられる。合理的な根拠に基づいた評価であれば、評価者の評価方法を治す必要はない。教師は、平均二乗誤差や評価者特性パラメータでこのような評価を見つけた場合、授業中に取り上げることにより、学習コミュニティの評価能力の向上につながる場合もあると考えられる。また、このような評価者に対してでも、他の評価者との評価の違いをフィードバック画面で表示し、個性的な評価をしていることを知らせることは意味があると考えている。

実データでの検証の結果、評価者特性モデルの当てはまりが悪い評価者も見られた。例えば、他の評価者と異なる基準で評価をしたり、ランダムに評価をつけたり、全員に同じ評価をつけるなど不真面目に評価を行っていたり、評価者の中での評価基準が安定していないなどの理由が考えられる。これらの評価者は、平均二乗誤差の数値などを見ることで、容易に見つけることができる。該当する評価者全て、もしくは教員が不適切な評価であると判断した評価者の行った評価については、該当する評価者を評価特性パラメータの推定から外したり、該当する評価者の評価を被評価者の画面に表示しないようにしたりするなどフィードバックに悪影響を与えない処理を行うことも考えられる。該当する評価者への対処方法としては、システムが該当する評価者に対して、他の評価者の評価と離れていることを提示することや、教師がシステムを使ってこのような評価者を検出し、該当する評価者がどのような評価を行っているかチェックした上で、個別に指導することが考えられる。他の評価者との評価の特徴の違いについてフィードバックすることで、次回以降の相互評価ではより適切な評価ができるようになると考えられる。

5章で取り上げた相互評価の実践では、分析のためにそれぞれの評価値をつけた根拠を書かせることにした。形成的評価のために行う相互評価では、根拠を書かせることは、適切なフィードバックを行うために、次の理由から特に重要である。

①レポート作成者にとって、フィードバック時に示された評価値を補完して、学習の改善につなげることができる

②評価者にとって、自身の行った評価を見直す時に、他人が行った評価の根拠が分かり、次回の評価をする際の参考にすることができる

③教員にとって、評価者特性モデルの当てはまりが悪い者や個性的な評価を行う者の評価の根拠が分かり、介入することができる。

本論文では数値データの処理について扱ったが、記述式の回答と対にしてフィードバックすることで、学習効果が期待される。

今後の課題としては、単に補正值を表示するだけでなく、学習者が自分の評価特性を理解できるような分かりやすいフィードバックを行うこと、芸術分野のように評価基準にある程度のばらつきが許容される領域など、今回取り上げなかった分野に適用し、本手法の精度を検証することがあげられる。

参 考 文 献

- 東洋, 梅本堯夫, 芝祐順, 梶田叡一 (1988) 現代教育評価事典. 金子書房, 東京
- 藤原康宏, 加藤浩 (2005) 学習コミュニティメンバーによる相互評価のための評価者選択方法. 電子情報通信学会技術研究報告, **ET2004-118** : 97-100
- 藤原康宏, 大西仁, 加藤浩 (2006) 相互評価のための課題提出評価支援システムの開発. 電子情報通信学会技術研究報告, **ET2005-94** : 5-10
- MARQUARDT, D.W. (1963) An algorithm for least-squares estimation of non-linear parameters. *Journal on Applied mathematics*, **11**(2) : 431-441
- PRESS H.W., TEUKOLSKY A.S., VETTERLING T.W. and FLANNERY P.B.(1988) Numerical recipes in C. Cambridge University Press, Cambridge. (丹慶勝市, 奥村晴彦, 佐藤俊郎, 小林誠訳(1993), 技術評論社, 東京.)
- SAMEJIMA F. (1969) Estimation of latent ability using a response pattern of graded score. *Psychometrika Monograph*, **17** : 1-100
- 芝祐順 (1991) 項目反応理論 基礎と応用. 東京大学出版会, 東京
- 植野真臣 (2005) 先端的 e-learning の理論と実践. 教育心理学年報, **44** : 126-137
- 植野真臣, 岡本敏雄, 永岡慶三 (2006) ピア・アセスメントにおける項目応答理論. 電子情報通信学会技術研究報告, **ET-2006-2** : 7-12
- 渡部洋, 平由実子, 井上俊哉 (1998) 小論文評価データの解析. 東京大学教育学部紀要, **28** : 143-164

Summary

We proposed the method of revising an evaluated score in considering of the evaluator's characteristic for an impartial feedback in a peer evaluation. Because it is difficult to evaluate all learners, peer evaluation data include many missing responses. In the proposal method, we assumed the evaluators' characteristic modeling based on the item response theory, and the evaluated score is revised by using the difference of estimated evaluator's characteristic and the evaluator of an average evaluation. In the proposal method, we assumed the evaluators' characteristic modeling based on the item response theory, and an evaluated score is revised by using the difference of estimated evaluator's characteristic and the

evaluator of an average evaluation. It was confirmed that the revised score was appropriate evaluation by real data. We practiced peer evaluation five times in our class. As a result of practice, it was found that the evaluators received feedback with revised score improved the evaluation ability, but the evaluator received only evaluation value before revision did not improved. Because the above-mentioned result, it was shown that proposed revision method was effective to an impartial evaluation and the improvement of the evaluation ability.

KEY WORDS: PEER EVALUATION, EDUCATIONAL EVALUATION, EVALUATOR'S CHARACTERISTIC IN A PEER EVALUATION, EDUCATIONAL DATA, REVISION OF AN EVALUATION RESULT

(Received February 19, 2007)