教育情報学研究　第4号（2006）
Educational Informatics Research, 2006, No. 4, 1-4

評定者内の評定のばらつきが信頼性に及ぼす影響

佐々木　典彰[1]，村木　英治[2]

[1] *東北大学大学院教育情報学教育部*
[2] *東北大学大学院教育情報学研究部*

要旨：本稿の目的は、評定者内の評定のばらつきが信頼性にどのような影響を及ぼすのかについて調べることであった。次の5つの評定者集団が設定されシミュレーションが行われた。(1) 全ての評定者における評定者内の評定のばらつきが大きい群（AH群)、(2) 評定のばらつきが大きい評定者を集めた群（GH群)、(3) ローデータ群（R群)、(4) 評定のばらつきが小さい評定者を集めた群（GL群)、及び (5) 全ての評定者における評定者内の評定のばらつきが小さい群（GL群）であった。信頼性の指標としては一般化可能性係数が使用された。その結果、評定者内の評定のばらつきは、信頼性に影響を及ぼすことが示され、AH群において最も信頼性が高く、AL群において最も信頼性が低かった。このことから、評定において信頼性を高めるためには、評定のばらつきが大きい評定者を評定者集団に多く含めることが効果的であることが示唆された。

キーワード：評定者、信頼性、一般化可能性係数

1．はじめに

近年、学校教育や人事考課等において、パフォーマンス評価（performance assessment）が導入されつつある。パフォーマンス評価の特徴の一つは、評定者が必ずしも一人ではなく、複数の評定者によって評定が行われることである。そうすると、評定者によって評定結果が異なる場合があり、信頼性に関する議論が必要となってくる。信頼性を高めるためには、一般化可能性理論（Brennan, 2001）に基づき、評定者の数を増やすことがあげられる。その一方で、ここでは評定者一人一人の特性は考慮されていない。山下・尾関（1989）及び山下・大野・尾関（1990）は、評定者の特性として寛大化傾向、中央化傾向、及び厳格化傾向を取り上げ、評定者がどの傾向であるかを示す方法を提案した。寛大化傾向とは、評定が尺度の上位に集中する傾向を指し、評定が甘いことを表し、それとは逆に厳格化傾向とは、評定が尺度の下位に集中する傾向を指し、評定が厳しいことを表す。そして中央化傾向とは、評定が尺度の中央に集中する傾向を指し、両極端の評定を避けることを表す。いずれの場合も、評定者内の評定のばらつきが小さいことを表し、一般によくみられる傾向と思われる。

そこで本論では、評定者内の評定のばらつきに注目し、それが信頼性にどのような影響を及ぼすのかについて調べることとした。具体的には、評定者内の評定のばらつきに基づいて複数の評定者集団を仮想的に作り、各々における信頼性を比較することとした。

2．方法

2.1. 評定データの生成

最初に、RESGEN（Muraki, 1992）を用い、評定データが生成された。評定者数は5名、7名、及び9名とし、被評定者数は10名、20名、及び30名とし、各々の被評定者の能力値は N (0, 1) からランダムに生成された。そして、各水準を組み合わせ、計9条件が設定された（表1）。各条件においては、乱数の初期値を変えてデータセットが20セット生成された。なお、項目数は1項目とした。

表1　9条件の設定

		評定者数		
		5	7	9
被評定者数	10	条件1	条件2	条件3
	20	条件4	条件5	条件6
	30	条件7	条件8	条件9

2.2. 分析モデル

本論では山下・尾関（1989）が示した分析モデル

$$x_{ki} = a_k + b_k\theta_i + e_{ki}$$

を用いた。x_{ki} は評定者 k が被評定者 i に対して評定を行ったときの観測値である。a_k は評定者 k における評定の甘辛を表すパラメータであり、b_k は評定者 k における評定のばらつきを表すパラメータである。そして θ_i は、被評定者 i の固有の力を表すパラメータであり、e_{ki} は誤差を表す。

2.3. パラメータの推定

ここで推定したいパラメータは a_k、b_k、そして θ_i である。山下・尾関（1989）の方法と同様に、まず、b_k を1に固定した上で、a_k と θ_i を最小二乗法によって推定し、その後、b_k を推定した。詳しくは次の2つのステップに示す。

2.3.1. ステップ1

ここでは a_k と θ_i が正規方程式

$$c = (F'F)^{-1}F'x \qquad (1)$$

によって推定される。c は $(m+n)$ 次のパラメータ・ベクトル

$$c' = (a_1, a_2, \cdots, a_m, \theta_1, \theta_2, \cdots, \theta_n)$$

であり、m は評定者数、n は被評定者数を表す。x は $(m+n)$ 次のデータ・ベクトル

$$x' = (x_{11}, x_{12}, \cdots, x_{21}, x_{22}, \cdots x_{mn})$$

である。そして、F は $(m\times n)$ 行 $(m+n)$ 列のダミー変数行列であり、$\{(k-1)n+i\}$ 行 k 列と $\{(k-1)n+i\}$ 行 $(m+i)$ 列の要素が1で、他の要素は全て0である。

$(F'F)$ の逆行列は求めることができないので、F の最終列と c の最終要素を削除し（1）式で解を求め、c の削除した要素の値は、数量化理論Ⅰ類の基準化の方法を用いて算出された。

2.3.2. ステップ2

ステップ1で得られた a_k と θ_i をもとに、b_k が正規方程式

$$b = (G'G)^{-1}G'p$$

によって推定される。b は k 次のパラメータ・ベクトル

$$b' = (b_1, b_2, \cdots, b_k)$$

である。G は $(m\times n)$ 行 m 列のデータ行列であり、$(k-1)n+i$ 行 k 列の要素が θ_i であり、その他の要素は全て0である。p は $(m\times n)$ 次のデータ・ベクトル

$$p = (x_{11}-a_1, x_{12}-a_1, \cdots, x_{ki}-a_k, \cdots, x_{mn}-a_m)$$

である。

2.4. 5つの評定者集団の設定

2.1.で生成された評定データと、評定のばらつきを表すパラメータ b_k をもとに、5つの評定者集団が設定された（表2）。(1) 全ての評定者における評定者内の評定のばらつきが大きい群（AH群）、(2) 評定のばらつきが大きい評定者を集めた群（GH群）、(3) ローデータ群（R群）、(4) 評定のばらつきが小さい評定者を集めた群（GL群）、及び (5) 全ての評定者における評定者内の評定のばらつきが小さい群（GL群）であった。言い換えれば、AH群では全ての観測値 $x_{ki} = a_k + b_k\theta_i + e_{ki}$ における b_k に $b = \max(b_k)$ が代入され、データが書き換えられたことになる。同様にAL群の全ての観測値 $x_{ki} = a_k + b_k\theta_i + e_{ki}$ における b_k に $b = \max(b_k)$ が代入され、データが書き換えられたことになる。他方GH群とGL群では、データの値はそのままで、評定者が2つのグループに分けられたことになる。この場合、評定者数が他の条件と異なることになる。そこでGH群とGL群においては、一般化可能性係数を算出するときの決定研究において、評定者数を他の条件と合わせることとした。

表2　各群の分析に用いたデータ

群	データ
AH群	$X'_{ki} = a_k + b\theta_i + e_{ki}\ (b = \max(b_k))$
GH群	$X_{ki} = a_k + b_k\theta_i + e_{ki}\ (b_k \geq \bar{b})$
R群	X_{ki}（ローデータ）
GL群	$X_{ki} = a_k + b_k\theta_i + e_{ki}\ (b_k < \bar{b})$
AL群	$X'_{ki} = a_k + b\theta_i + e_{ki}\ (b = \min(b_k))$

2.5. 信頼性

信頼性の指標として、一般化可能性係数（Brennan, 2001）が用いられた。

2.6. プログラムの作成

2.2.から2.5.については、著者がMATLABプログラムを作成し、それと2.1で生成したデータをもとにシミュレーションを行った。

3．結果

各条件において、5（評定者集団）×20（データセット）＝100の一般化可能性係数が算出された。各条件における評定者集団間の差を調べるため、クラスカル・ウォリス検定を行った。その結果、全ての条件において全体的な有意差がみられた（表3）。Scheffeの多重比較の結果を表4に示す。また、各条件における一般化可能性係数の平均値を図1～3に示す。

表3　各条件におけるクラスカル・ウォリス検定の結果（数値はχ^2値）

		評定者数		
		5	7	9
被評定者数	10	38.30**	51.27**	55.77**
	20	37.21**	43.75**	54.99**
	30	23.75**	55.34**	57.56**

** p<.01

表4　各条件において有意差のみられた群の組み合わせ

		評定者数		
		5	7	9
被評定者数	10	AH - GL** AH - AL** GH - AL**	AH - GL** AH - AL** GH - GL* GH - AL** R - AH*	AH - R** AH - GL** AH - AL** GH - GL** GH - AL**
	20	AH - GL** AH - AL** GH - AL**	AH - R* AH - GL** AH - AL** GH - GL* GH - AL**	AH - R* AH - GL** AH - AL** GH - GL* GH - AL** R - AL*
	30	AH - AL** GH - AL**	AH - R* AH - GL** AH - AL** GH - GL** GH - AL**	AH - R* AH - GL** AH - AL** GH - GL** GH - AL** R - AL*

* p<.05　** p<.01

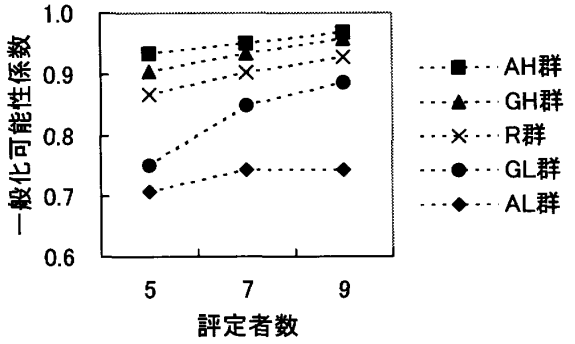

図1　各群における一般化可能性係数の値（被評定者数=10）

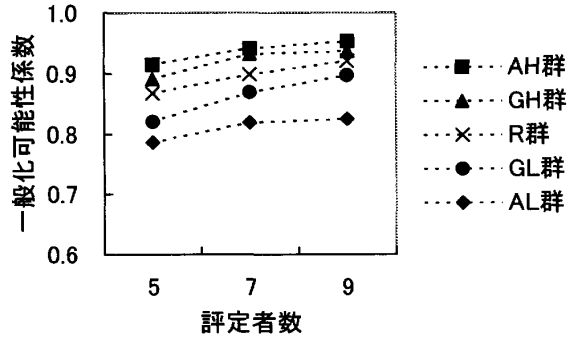

図2　各群における一般化可能性係数の値（被評定者数=20）

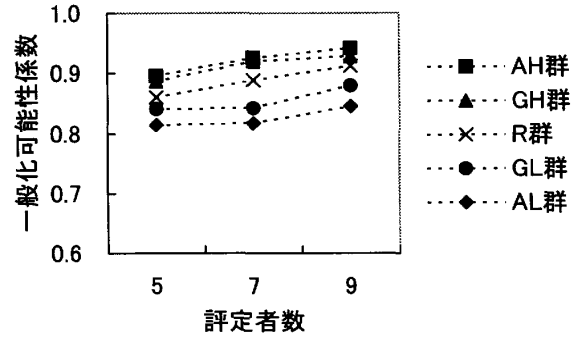

図3　各群における一般化可能性係数の値（被評定者数=30）

4．考察

評定者内の評定のばらつきは、信頼性に影響を及ぼすことが示され、AH群において最も信頼性が高く、AL群において最も信頼性が低かった。このことから、評定において信頼性を高めるためには、評定のばらつきが大きい評定者を評定者集団に含めることが効果的であることが示唆された。逆に言えば、寛大化傾向、中央化傾向、または厳格化傾向を持つ

評定者を評定者集団に含めることは、信頼性を高める観点からすれば、あまり望まれないと言える。しかしながら、評定のばらつきが大きいと言っても、おおげさに評価を行っていたり、でたらめに評定をしていたりしている場合も考えられ、注意が必要である。

また、寛大化傾向などの評定者の特性が、評定する時間や場所に依存しない継続性の高いものであるとみなすならば、各評定者のパラメータb_kをデータベース化し、評定者を決定するときの一つの資料として活用することも考えられるだろう。

文献

Brennan, R. L. 2001 *Generalizability theory.* New York: Springer-Verlag.

Muraki, E. 1992 *RESGEN: Item response generator.* Princeton, NJ: Educational Testing Service.

山下洋史・尾関守　1989　人事考課における評定傾向分析モデル　日本経営工学会, 40, 3, 177-182.

山下洋史・大野高裕・尾関守　1990　人事考課における寛大化傾向・中央化傾向・厳格化傾向の定量的分析　日本経営工学会誌, 41, 5, 336-341.

An Effect of Dispersion of Rating in Each Rater on Reliability

Noriaki Sasaki [1], Eiji Muraki [2]

[1] *Graduate School of Educational Informatics, Education Division, Tohoku University*

[2] *Graduate School of Educational Informatics, Research Division, Tohoku University*

The purpose of this study was to investigate an effect of dispersion of rating in each rater on reliability. Five groups were set for simulation as follows: (1) Group whose dispersions of rating in all raters were constantly high (AH group), (2) Group whose dispersions of rating were relatively high (GH group), (3) Raw data group(R group), (4) Group whose dispersions of rating in each rater were relatively low (GL group), and (5) Group whose dispersions of rating in all rater were constantly low (AL group). Generalizability coefficient (Brennan, 2001) was used as an index of reliability. The results showed that the dispersion of rating in each rater affected reliability significantly, and showed that reliability was the highest in AH group and the lowest in AL group. Therefore, it was proposed that raters whose ratings were widely dispersed could be used to obtain high reliability in assessments.

Key words: rater, reliability, generalizability coefficient