【原著】

主観的評価による合否決定の事例集作成に向けた試み

椎名久美子（大学入試センター），平井洋子（首都大学東京）

本稿では，小論文や面接などの主観的評価による合否決定を行う際の注意点を示す事例集作成に向けた試みとして，実際にあり得る評価方法を単純化した 2 種類のデザインに関して，一般化可能性理論を論述式課題の採点データに適用し，受験者の順位づけの安定性（信頼性）の違いを検討した。評価方法のデザインが同じでも，用いる観点や評定者の組合せによって信頼性の値が異なり，観点と評定者の選択が信頼性にかなり影響することが示された。

1 研究の背景および目的

各大学の入学者選抜で主観的な評価が行われる例として，小論文や面接などが挙げられる。 AO 入試や推薦入試の選抜材料として提出される志望理由書なども，何らかの観点に基づいて得点化するという意味では，主観的な評価の例に入るだろう。

日本テスト学会による「テストの作成，実施，利用，管理に関わる規準」（以降「テスト規準」と略記）では，主観的な評価による採点に関して，「…基本設計にそって採点できるように，評定者のトレーニングをする。また，採点後には複数評定者の評定の整合性などを分析し，必要に応じてさらに調整を加える。（基本条項 2.9 ）」と述べられている（日本テスト学会編, 2007: 86-90）。しかし，現状では，入学者選抜において，事前のトレーニングによる評定者間での採点基準の共有や，評定結果に変動を及ぼす要因に関する検討などが，全国的にどの程度実施されているのかは把握されていない。

木村・吉村(2010)は， AO 入試の信頼性評価の試みとして，一般化可能性理論 (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001)を用いて，評定値に変動を及ぼす各要因による分散成分の大きさや，評価観点ごとの評価の一致度などの分析を行っている。また，平井(2007)は，入試を想定した論述式課題を試作し，評定基準の詳しさが，評定者間での受験者の順位づけの一致度に与える影響に関して，一般化可能性理論による考察を行っている。

入学者選抜における主観的評価では，受験者の答案（面接であれば，その受け答え）を何名の評定者がどのように分担して，どのような観点で評定するかは，各大学の実状に応じてかなり異なると思われる。テスト規準に準拠した試験という点では，評価方法のデザインによって受験者の順位づけの信頼性が変わることを意識した上で，各大学で実施される主観的評価の安定性を検討することが望ましい。

本稿では，主観的評価による合否決定を行う際の注意点を示す事例集作成に向けた第一歩として，実際にあり得る評価方法を単純化したデザインを 2 種類設定し，具体的なデータを用いて受験者の順位づけの安定性（信頼性）の違いを検討する。

2 評価方法のデザインによる信頼性の違いに関する検討

2.1 評価方法のデザインと一般化可能性理論

評定者 (rater (r))，評価観点(viewpoint (v))，受験者 (person (p)) の配置という観点

からみると，様々な評価方法のデザインがあり得るが，本稿では2つのデザインを扱う。

1つは，同じ評定者が同じ観点ですべての受験者を評定するデザインである。もう1つは，評定者はすべての受験者を評定するが，各評定者は異なる観点を用いるデザインである。Brennan(2001)の表記に従えば，前者は $p \times v \times r$，後者は $p \times (v : r)$ と表記される。$a \times b$ は a と b がクロスした配置，$a : b$ は a が b にネストされた配置を示す。

表1(a)は $p \times v \times r$ の例で，評定者2名が2つの観点を共有して，n_p 人の受験者を評定するデザインを模式的に表したものである。表1(b)は $p \times (v : r)$ の例で，評定者2名がそれぞれ異なる2つの観点で，n_p 名の受験者を評定するデザインを模式的に表したものである。Xは，何らかの量的な評定値（得点）が入ることを示す。

一般化可能性理論の適用は，一般化可能性研究（ generalizability study; G 研究）と決定研究（decision study; D研究）の2つのステップに大別される。G研究では，分散分析の枠組みを用いて，得点に誤差を及ぼす要因（変動因）ごとに分散成分の大きさが推定され，D研究では，推定された分散成分をもとに，評価方法のデザインを変更した場合の信頼性がシミュレーションされる（平井，2007 ）。

表2(a)(b)は，表1に示す2つのデザインに関して，Brennan(2001)に従って得点 Xの分散を変動因ごとに分解した成分（分散成分）を，その意味と共に示したものである。受験者を変動因とする分散成分 σ_p^2 が Xの分散に占める割合が大きいほど，受験者の能力を識別できるデザインといえる。

表2(a)(b)では，それぞれ，得点の順位を乱す変動因が網掛けで示されている。得点の相対的な順位づけを乱す誤差は相対誤差 σ_δ^2 と呼ばれ，表1(a)と (b) のデザインにおける相対誤差は，それぞれ，(1)式と (2) 式で算出される。

$$\sigma_\delta^2 = \frac{\sigma_{p\times v}^2}{n_v} + \frac{\sigma_{p\times r}^2}{n_r} + \frac{\sigma_{p\times v\times r}^2}{n_v n_r} \qquad (1)$$

$$\sigma_\delta^2 = \frac{\sigma_{p\times r}^2}{n_r} + \frac{\sigma_{p\times v:r}^2}{n_v n_r} \qquad (2)$$

(1)(2) 式において，n_v は観点数，n_r は評定者数を示す。

一般化可能性係数（G係数）は，全評定者，全観点にわたって平均した得点が，評定者や観点（得点の順位を乱し得る変動因）が変わってもどの程度一貫しているかを示す指標であり，以下の (3) 式で算出される。G係数は，古典的テスト理論における信頼性係数に相当し，G係数が大きいほど，受験者の順位づけが安定していることを意味する。

$$\text{G 係数} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \qquad (3)$$

2.2 評価方法のデザインと受験者の順位づけの安定性

2.2.1 検討に用いるデータとデザイン

検討に用いるのは，与えられた資料から情報を取捨選択して判断して考えを述べる形式の論述式課題（大学入試センター研究開発部, 1999）の設問の1つを，4つの観点から3名の評定者が採点したデータである（平井, 2007）。「評価すべき側面に関する数行程度の定義文と得点範囲を与え，あとは自分で解釈して採点する」という，評定者が判断する余地が大きな評定基準を与えられた3名（r1，r2，r3と識別する）が，すべての受験者（268名）の解答を以下の4つの観点（平井，未発表）で採点した $p \times v \times r$ デザインのデータである。以降，必要に応じて以下の4つの観点を v1 ～ v4 と略記する。

オリジナリティ（v1）　分析的視点（v2）
多角的視点(v3)　　論理的一貫性(v4)

3名の評定者が4つの観点で全受験者を採点したデータからは，受験者を変動因とする

分散成分σ_p^2 がXの分散に占める割合 0.204 およびG係数 0.668 が算出される[1)]。

本稿では，評定者 2 名が 2 つの観点を共有して採点する場合（表1(a)）と，2 名が異なる 2 つの観点で採点する場合（表1(b)）を考える。各デザインに関して，観点の選び方によって，受験者の順位づけの安定性がどのように変化するかを検討する。

表 1　評定者 (r)，評価観点 (v)，受験者 (p) のデザイン

(a)　$p\times v\times r$

	評定者 1		評定者 2	
	観点 1	観点 2	観点 1	観点 2
受験者 1	X	X	X	X
受験者 2	X	X	X	X
…	…	…	…	…
受験者 n_p	X	X	X	X

(b)　$p\times(v:r)$

	評定者 1		評定者 2	
	観点 1	観点 2	観点 3	観点 4
受験者 1	X	X	X	X
受験者 2	X	X	X	X
…	…	…	…	…
受験者 n_p	X	X	X	X

表 2　得点Xの分散を変動因ごとに分解した成分（分散成分）

(a)　$p\times v\times r$

変動因	分散成分	意味
受験者 (p)	σ_p^2	受験者の能力の個人差
観点 (v)	σ_v^2	観点ごとの難易差
評定者 (r)	σ_r^2	評定者ごとの一貫した甘さ／辛さ
受験者×観点 ($p\times v$)	$\sigma_{p\times v}^2$	観点に対する得手・不得手の，受験者間での違い
受験者×評定者 ($p\times r$)	$\sigma_{p\times r}^2$	受験者につけた得点の大小の，評定者間での違い
観点×評定者 ($v\times r$)	$\sigma_{v\times r}^2$	観点ごとの平均値の大小の，評定者間での違い
残差	σ_{res}^2	デザインに含まれない要因，すべての要因の交互作用，及び，ランダムな誤差

(b)　$p\times(v:r)$

変動因	分散成分	意味
受験者 (p)	σ_p^2	受験者の能力の個人差
評定者 (r)	σ_v^2	評定者ごとの一貫した甘さ／辛さ
観点：評定者 ($v:r$)	$\sigma_{v:r}^2$	各評定者内での全観点を平均した評定値が，評定者によってどれだけ変動するか
受験者×評定者 ($p\times r$)	$\sigma_{p\times r}^2$	受験者につけた得点の大小の，評定者間での違い
残差	$\sigma_{p\times v:r}^2$	デザインに含まれない要因，受験者と観点の交互作用が評定者によってどれだけ変動するか，及び，ランダムな誤差

2.2.2 $p \times v \times r$ デザインに関する検討

$p \times v \times r$ デザインについては，v1 から v4 の観点から 2 つを選ぶ組合せが 6 通りあり，3 名の評定者から 2 名を選ぶ組合せが 3 通りある。観点と評定者の選び方の組合せ 18 通りに関して，各変動因による分散成分が X の分散に占める割合，G係数の推定値を算出する。図 1 に，各組合せに関して，表 2(a)に示す各変動因の分散成分が全体に占める割合を図示する。図 1 において，例えば，v1 & v2 はオリジナリティと分析的視点という観点を共有して採点したことを示し，r1 & r2 は評定者 2 名の組合せを示す。表 3 には，18 通りの組合せに関するG係数の推定値を示す。

受験者を変動因とする分散成分 σ_p^2 の割合及びG係数が，どの評定者 2 名の組合せでも安定して高い値を示すのは，v2 & v4 の組合せである。この組合せでは，σ_p^2 の割合及びG係数は，$p \times v \times r$ デザインで 3 名の評定者が 4 つの観点で採点した場合の値よりも高くなっている。

表 3　$p \times v \times r$（観点数 2，評定者数 2）の 18 通りにおけるG係数

観点の組合せ	評定者の組合せ	G係数
v1 & v2	r1 & r2	0.480
	r1 & r3	0.571
	r2 & r3	0.523
v1 & v3	r1 & r2	0.480
	r1 & r3	0.297
	r2 & r3	0.294
v1 & v4	r1 & r2	0.388
	r1 & r3	0.291
	r2 & r3	0.300
v2 & v3	r1 & r2	0.620
	r1 & r3	0.473
	r2 & r3	0.551
v2 & v4	r1 & r2	0.704
	r1 & r3	0.713
	r2 & r3	0.674
v3 & v4	r1 & r2	0.526
	r1 & r3	0.384
	r2 & r3	0.409

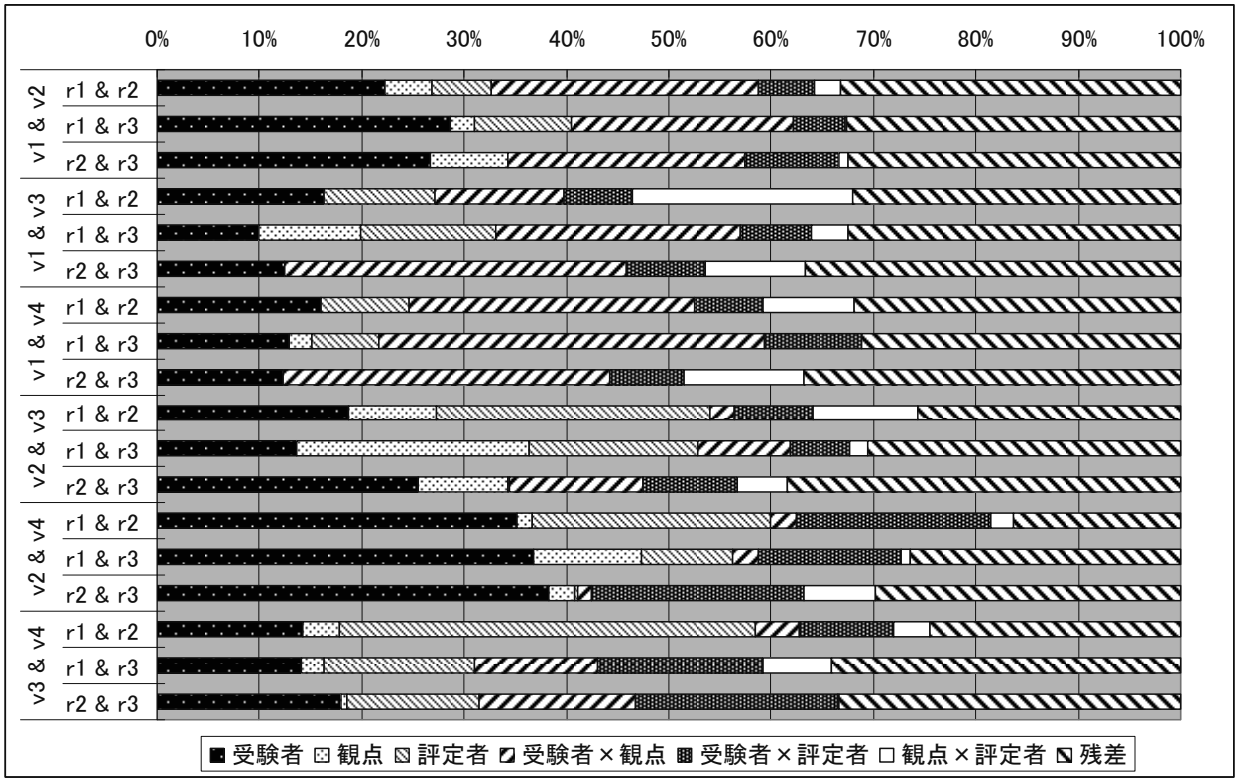


図 1　$p \times v \times r$（観点数 2，評定者数 2）の 18 通りにおける各変動因の分散成分の割合

v2 と v4 はいずれも，観点内での評定者間の相関係数が比較的高く（ 0.3 台後半～ 0.5 台前半）， v2 と v4 の間での評定者間の相関係数もかなり高い（ 0.3 台前半～ 0.5 台後半）。これは，評定者間での一致度が高い観点を用いることで，多くの評定者が多くの観点数で採点するよりも信頼性が高まる事例に相当する。

v1 & v4 の組合せでは，σ_p^2 の割合及びG係数は，どの評定者 2 名の組合せでも非常に低い。 v1 と v4 は，観点内での評定者間の相関係数が比較的高いが（ 0.3 台後半～ 0.6 台前半）， v1 と v4 の間での評定者間の相関係数が非常に低い（ 0.1 台前半～ 0.3 台前半）。 v1 & v4 の組合せは，単一の観点では評定者間の一致度が高くても，観点間での一致度が低いために，信頼性が低くなる事例に相当する。

表 3 をみると，評定者 2 名が 2 つの観点を共有して採点する場合でも， 2 つの観点の組合せと評定者 2 名の組合せによっては，G係数が 0.3 未満になり，信頼性が極端に低くなる危険があることが示唆される。

2.2.3 $p \times (v : r)$ デザインに関する検討

$p \times (v : r)$ デザインについては， 4 つの観点から異なる 2 つを 2 名の評定者に割り当てる組み合わせが 3 通りある。それぞれの観点の組合せについて， 3 名から 2 名の評定者を選んで，それぞれ異なる 2 つの観点を割り当てる組合せ（ 6 通り）に関して，各変動因の分散成分が Xの分散に占める割合，G係数の推定値を算出する。

図 2(a) に，一方の評定者がオリジナリティと分析的視点（ v1 & v2 ）という観点，他方の評定者が多角的視点と論理的一貫性（ v3 & v4 ）という観点で採点した場合について， 2 名の評定者の割り振り方 6 通りに関して，表 2(b) に示す各変動因の分散成分が全体に占める割合を図示する。例えば（ v1 & v2 ：r1／ v3 & v4 ：r2）は，評定者r1が観点 v1 と v2 ，評定者r2が観点 v3 と v4 で採点したことを示す。同様に，図 2(b) には，一方の評定者が v1 と v3 ，他方の評定者が v2 と v4 という観点で採点した場合を，図 2(c) には，一方の評定者が v1 と v4 ，他方の評定者が v2 と v3 という観点で採点した場合を示す。表 4(a)～ (c) には，図 2(a)～ (c) に対応するG係数の推定値を示す。

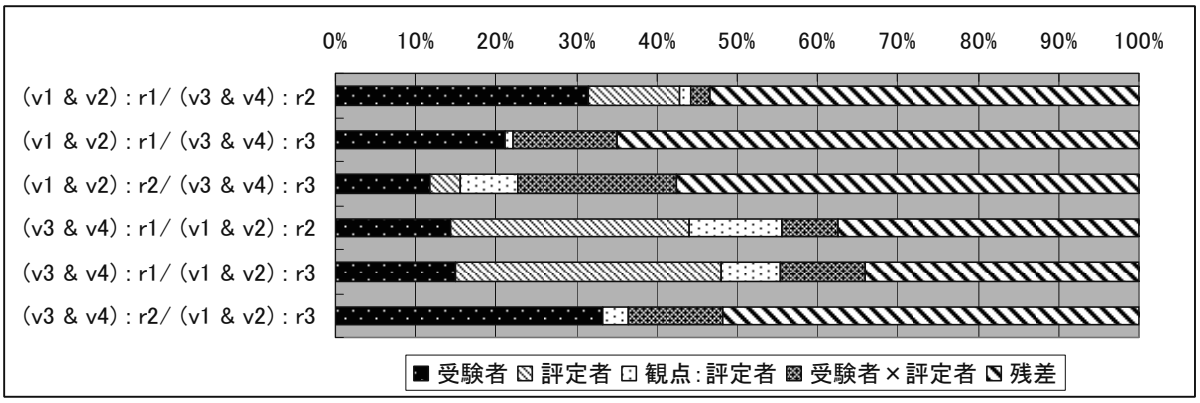
v1 & v4 ， v2 & v3 の観点の組合せで採点した場合（図 2(c) ）は，どの評定者 2 名の組合せでも，受験者を変動因とする分散成分 σ_p^2 の割合は 0.2 を超えており， $p \times v \times r$ デザインで 3 名の評定者が 4 つの観点で採点した場合とほぼ同じ割合になっている。G係数は， $p \times v \times r$ デザインで 3 名の評定者が 4 つの観点で採点した場合よりは低いものの，評定者 2 名の組合せで極端にG係数が低くなるものはない。 3 名の評定者が 4 つの観点で採点するほどの信頼性は得られないものの， v1 & v4 ， v2 & v3 の観点の組合せをどの評定者 2 名に割り当てても信頼性に大きな違いは生じない事例である。

v1 & v2 ， v3 & v4 の観点の組合せで採点した場合（図 2(a) ，表 4(a) ）は，評定者 2 名の組合せによっては， $p \times v \times r$ デザインで 3 名の評定者が 4 つの観点で採点した場合より σ_p^2 の割合が高くなる場合と低くなる場合がある。G係数は 0.325 から 0.684 まで幅広い値をとる。すなわち， v1 & v2 ， v3 & v4 の 2 つずつの観点の組合せに，評定者 2 名をどう割り当てるかによって，信頼性の値が大きく異なるものになる。

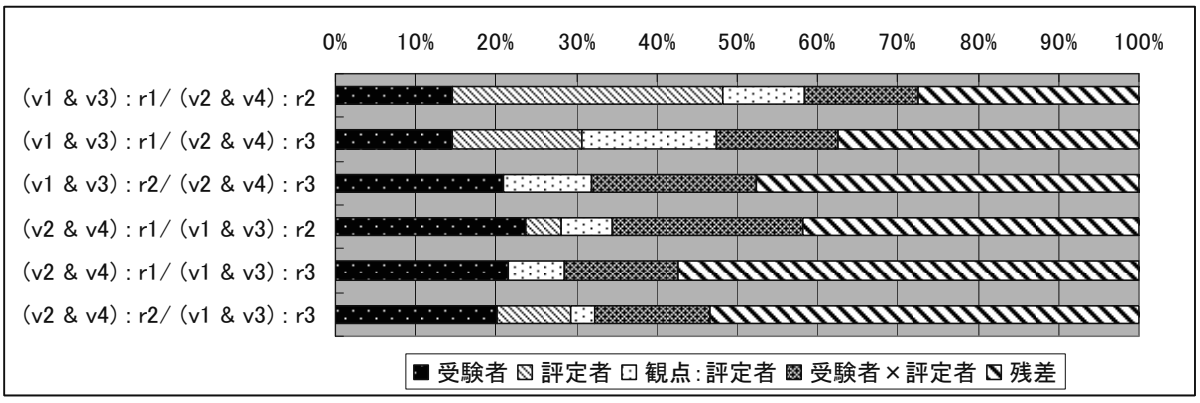
評定者r1が v1 と v2 ，評定者r2が v3 と v4 の観点で採点した場合のG係数は 0.684 で， 3 名の評定者が 4 つの観点で採点するよりも信頼性が高まる事例に相当する。一方，評定者r2が v1 と v2 ，評定者r3が v3 と v4 の観点で採点した場合のG係数は 0.325 で， 3 名の評定者が 4 つの観点で採点するよりも信頼性が低くなる事例に相当する。

表 4(a)～ (c) の 18 通りの組合せのうち，

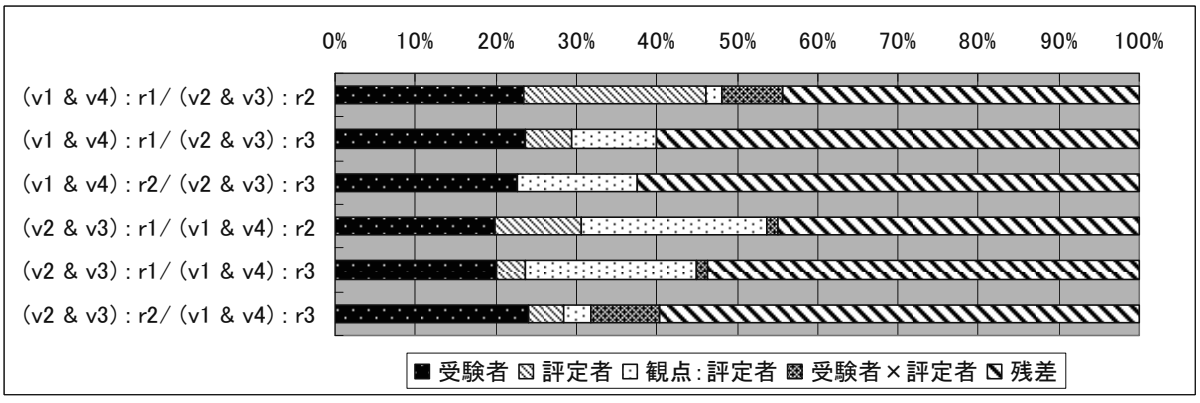
G係数が最小の値 (0.325) をとるケースを除けば，G係数は 0.4 台半ばから 0.6 台後半の値をとる。 2.2.2 で検討した $p \times v \times r$ デザインにおいても，評定者 2 名が共有する 2 つの観点の組合せによってはG係数が 0.3 未満になるケースが 3 通りあることを考えると，必ずしも $p \times (v : r)$ デザインが劣るわけではなさそうである。



(a) オリジナリティと分析的視点(v1& v2)，多角的視点と論理的一貫性(v3&v4)で採点した場合



(b) オリジナリティと多角的視点(v1& v3)，分析的視点と論理的一貫性(v2&v4)で採点した場合



(c) オリジナリティと論理的一貫性(v1& v4)，分析的視点と多角的視点(v2&v3)で採点した場合

図 2　$p \times (v : r)$（観点数 2，評定者数 2）における各変動因の分散成分の割合

表 4　$p \times (v : r)$（観点数 2，評定者数 2）におけるG係数

(a) オリジナリティと分析的視点(v1& v2)，多角的視点と論理的一貫性(v3&v4)

評定者と観点の組合せ	G係数
(v1 & v2) : r1/ (v3 & v4) : r2	0.684
(v1 & v2) : r1/ (v3 & v4) : r3	0.482
(v1 & v2) : r2/ (v3 & v4) : r3	0.325
(v3 & v4) : r1/ (v1 & v2) : r2	0.528
(v3 & v4) : r1/ (v1 & v2) : r3	0.521
(v3 & v4) : r2/ (v1 & v2) : r3	0.638

(b) オリジナリティと多角的視点(v1& v3)，分析的視点と論理的一貫性(v2&v4)

評定者と観点の組合せ	G係数
(v1 & v3) : r1/ (v2 & v4) : r2	0.510
(v1 & v3) : r1/ (v2 & v4) : r3	0.463
(v1 & v3) : r2/ (v2 & v4) : r3	0.487
(v2 & v4) : r1/ (v1 & v3) : r2	0.515
(v2 & v4) : r1/ (v1 & v3) : r3	0.501
(v2 & v4) : r2/ (v1 & v3) : r3	0.495

(c) オリジナリティと論理的一貫性(v1& v4)，分析的視点と多角的視点(v2&v3)

評定者と観点の組合せ	G係数
(v1 & v4) : r1/ (v2 & v3) : r2	0.612
(v1 & v4) : r1/ (v2 & v3) : r3	0.611
(v1 & v4) : r2/ (v2 & v3) : r3	0.592
(v2 & v3) : r1/ (v1 & v4) : r2	0.627
(v2 & v3) : r1/ (v1 & v4) : r3	0.587
(v2 & v3) : r2/ (v1 & v4) : r3	0.555

3 今後に向けて

本稿では，入試を想定したデータを用いたG研究により，評価方法のデザインや評定者・観点の組合せによって得点の信頼性がかなり変化することを示したが，入試における意思決定の改善に繋げるためには更にD研究を行う必要がある．実際の入試では，評定者の人数や配置，評定にかけられる時間などの実施上の制約条件のもとで，得点の信頼性を確保しつつ合否決定の妥当性を高めることが求められる．評定に用いる観点の数や組合せ，評定者の配置に関して，具体的な答えを得るよう努めるべきであろう．その際，信頼性の確保と，多様な観点からの評定のバランスについても考慮する必要があるのは言うまでもない．木村他(2010)では，AO 入試の評定データを一般化可能性理論によって検討することで，十分な信頼性を確保できる採点者の数や配置の検討，質問項目の改善を試みている．このような試みが主観的評価による入試を行う大学に広まることが期待される．

主観的評価による合否決定の事例集としては，評定者による順位づけの信頼性の高低が，合否の入れ替わりに及ぼす影響をわかりやすく示す必要があるだろう。今後，選抜の倍率などの要素も取り入れて，各大学で評価方法を検討する際に参考にできるような事例を蓄積していきたい。

注

1) 本稿では，受験者を変動因とする分散成分が Xの分散に占める割合やG係数の算出に，Mushquash and O'Connor (2006)によるSASプログラムを用いた。

参考文献

Brennan, R. L. (2001). *Generalizability Theory*, Springer-Verlag New York, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles.*, New York: Wiley.

大学入試センター研究開発部 (1999). 「総合試験モニター調査テスト問題集」『大学の各専門分野への適性の評価を目的とす

る総合試験のあり方に関する共同研究 最終報告書別冊』 .

平井洋子 (2007). 「主観的評定における評定基準，評定者数，課題数の効果について ― 一般化可能性理論による定量的研究 ―」『首都大学東京人文学報』**380**, 25-64.

平井洋子 (未発表). 「モニター調査(2001.2) 『自転車』採点作業記録」 .

木村拓也・吉村宰 (2010). 「 AO 入試における信頼性評価の研究― 一般化可能性理論を用いた検討―」『大学入試研究ジャーナル』 **20**, 81-89.

Mushquash, C. and O'Connor, B., P. (2006). "SPSS and SAS Program for Generalizability Theory Analyses," *Behavior Research Methods*, **38**(3), 542-547.

日本テスト学会編 (2007). 『テスト・スタンダード 日本のテストの将来に向けて』 金子書房 .