【原著】

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 ──

木村拓也，吉村宰（長崎大学）

本研究の目的は、一般化可能性理論を用いたAO入試における信頼性評価の方法を提案することである。「選抜方法の多様化、評価尺度の多元化」を基軸にする大学入学者選抜方法・制度の設計では、限られた学内の人的・時間的リソースを際限なく浪費してしまう方向にしか向かわない。本研究の成果によって、十分な信頼性を担保した上での評価観点の峻別/縮減・採点者(面接官)の削減指針が獲得可能となる。

1. 問題の所在—入試の多様化を「縮減」する論理とAO入試の信頼性評価研究の必要性

現行の大学入学者選抜実施要項「1.基本方針」に書き添えられている「選抜方法の多様化、評価尺度の多元化」という大学入学者選抜における錦の御旗は、*拡大の論理*こそあれ、どこまでの多様化・多元化であるべきかといった*縮減の論理*がそもそも内包されていない。中教審答申「学士課程教育の構築に向けて」で、AO入試や推薦入試の学力担保に憂慮が示されたりしたが（中教審2008:30)、AO入試や推薦入試などの入学者選抜方法の多様化を如何に*適正規模*で行うかを考えることが、今後、大学入学者選抜の制度設計における喫緊の課題となってくることは間違いない。

一方で、大半が記述式の問題で占められる国立大学二次試験の現場では、試験の実施・採点・合否判定と非常にタイトなスケジュールで行われるが故に、項目反応理論(IRT)が必要とする個別受験生の詳細な回答データが必ずしも得られる訳ではないものの、2次試験の教科科目については、五分位図[1]などによる大まかな識別力の検討くらいには、項目分析を行うことができる。だが、AO入試や推薦入試で行われる面接試験や小論文試験、書類審査の信頼性を担保する方法論の共有が未だ大学入試研究業界でされていないのが現状である。

そもそも、古典的テスト理論では、テスト得点をX、真値をT、誤差をEとしたとき、次式を出発的として考えられてきた。

$$X = T + E$$

このとき、古典的テスト理論では、信頼性係数$\rho(X)$を次式で定めている。

$$\rho(X) = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$$

つまり、総得点の分散に占める真値の分散の割合が信頼性係数であり、一般に、再テスト法や平行テスト法で推定してみたり、信頼性の基準としてCronbachのα係数を判断の指標として用いたりする。

一方、テスト理論の分野では、誤差Eの評価についても研究が進み、一般化可能性理論[2]と呼ばれるものが知られている。一概に誤差と言っても、試験実施時期や回数などの時間的な変動、出題される課題（質問項目）の違いといった項目間の変動、記述式のテストや面接であれば、評定者がつける評点の違いといった評定者間の変動といったものなどが混在している。一般化可能性理論とは、こうした系統誤差要因［相(fact)と呼ばれる］ごとに誤差成分を特定して、テスト得点が持つ誤差の範囲がどこまで一般化されるものかを指標化しようとして生まれたものである

(池田 2007)。既に、一般化可能性理論は、日本において、法科大学院統一適性試験(前田 2004)、医療系大学間共用試験の OSCE(宮本他 2006)のような公的試験に留まらず、大学における授業評価(豊田・中村 2004)、学生同士の相互評価(熊澤 2006)、日本語プレースメントテスト(坂野 2008)、企業の人事評価(鷲坂他 2004)、中学校における観点別評価(山森 2003)などでも応用されてきた。

また、一般化可能性理論には、分散分析の手法を利用して、各要因の大きさ(分散成分)を推定する Generalizability Study(G 研究)と、G研究で得られた分散成分から各テストデザインの信頼性を評価する Decision Study (D 研究)とがある(池田 2007)。分かりやすく言えば、前者については、面接評価や小論文採点において、いかなる誤差要因が大きく占めているのかが分かり、後者については、何人の採点者(面接官)で、いくつの観点で試験を行うと、テストとしての信頼性が担保されるのかについての解を得られる。つまり、一般化可能性理論を AO 入試のデータ分析に用いることで、冒頭にあげた、2つの問題、入学者選抜方法の多様化に際しての適正規模といった制度設計にまつわる問題と、AO 入試や推薦入試における面接試験や小論文試験、書類審査の信頼性評価の問題とを同時に解決することが予想される。よって、一般化可能性理論の大学入学者選抜における適用を提案することが本研究の目的である。

2. N大 AC の規則改正と本研究データの概要

N大学アドミッションセンター(以後、AC と略記)では、2008 年度からセンター規則が変更となり、新たに業務として「一般選抜、特別選抜、AO 入試等入学者選抜全般の方法に関する情報提供、助言及び支援に関すること」という項目が追加された。それまでは AO 入試の実施機関としての性格が強かったが、この規則改正により、能研テスト時に文部省によって導入が検討されたアドミッション・オフィスの研究機関としての本来の姿(木村 2008)に近づいた。これまでもN大学 AC では、五分位図などでの大問分析を2次科目(英数理及び総合問題)で行い、実施を担ってきた AO 1次の検討については幾つかの報告が行われてきた(吉村 2007・2008、吉村・南部 2008)。規則改正後は、AO 2次の分析も主任務の範疇となったので、分析事例として紹介することを試みる。尚、本研究で用いたデータは、A大学B学部C学科の AO 入試結果である。選抜方式は、1次で自己推薦書2様式と調査書成績が課され7人で採点している。また、募集要項に記載してある通り、2次では、「課題作文を課し、面接(数学・物理・化学・英語に関する基礎学力及び課題作文の内容等について問う)を行う」[原文のママ]。そして、課題作文を 13 人で採点し、同一の8観点で別個の質問項目で行う面接室を3カ所設けて面接官各4人で、面接(本稿では、「面接1ー3」と定義し、そのいずれもに共通の「観点1ー8」が存在する)を行い、更に、面接官3人の面接を4カ所で行う(本稿では、「面接4」と定義し、その観点を「観点1ー4」とした)など、実施要項で示された「選抜方法の多様化、評価尺度の多元化」の方針に基づいて、複雑な選抜方法が設計され、少数の定員枠の受験生に対して多数の教員の労苦を必要とする、大がかりな検査態勢を構築している。

3. 検査課題間及び評価観点間の相関と寄与

一般化可能性理論による検討に入る前に、検査課題及び検査項目の特徴を把握するために、課題(観点)間の相関係数を算出した。但し、受験者数が少数であったので、外れ値の影響を鑑みて、Spearman の順位相関係数で算出してある(表1)。ちなみに、各検査課題得点の標準偏差は、課題作文、面接1から4まで、ぞれぞれ、2.83、1.73、1.29、1.84、

表1. 検査課題間の順位相関、及び、総得点における寄与

	1次書類	小論文	面接1	面接2	面接3	面接4	総合点
1次書類	1.000	—	—	—	—	—	—
課題作文	**.220**	1.000	—	—	—	—	—
面接1	.534	.499	1.000	—	—	—	—
面接2	.498	.423	.510	1.000	—	—	—
面接3	.492	**.107**	.480	.506	1.000	—	—
面接4	**.362**	.715	.604	**.294**	**.397**	1.000	—
総合点	**.371**	.791	.642	**.346**	.421	.979	1.000
共分散比	—*	.097	.052	.032	.029	.790	1.000
配点割合との比較		>	>>	>>	>>	<<<	

左下の数字:Spearmanの順位相関係数、<:.100未満、<<:.100以上.300未満 <<<:.300以上
*:2段階選抜(但し、1次不合格なし)のため共分散比が存在しない。**太字**:.400未満

表2. 総得点における評価観点の寄与

	評価観点		共分散比	配点割合との比較
1次	書類1		.375	<
	書類2		.191	>>>
	調査書		.434	<<
2次	課題作文	観点1	.044	>>
		観点2	.033	>
		観点3	.020	>
	面接室123	観点1	.022	>
		観点2	.008	>>
		観点3	.010	>>
		観点4	.015	>>
		観点5	.008	>>
		観点6	.018	>
		観点7	.013	>
		観点8	.020	>
	面接室4	観点1	.218	<<<
		観点2	.211	<<<
		観点3	.142	<<<
		観点4	.219	<<<

<:.05未満、<<:.050以上.100未満 <<<:.100以上

20.40であり、配点割合に大差がないにも関わらず、採点の段階で、得点のばらつきに極端に差があった検査課題とまったく差がつかなかった検査課題があったことが分かる。

また、各検査課題/評価観点が総得点に与える寄与を確認するために共分散比を算出し、配点割合との比較を試みた(表1及び表2)。

表1を見ると、総合点の順位と相関が最も高いのは、面接4の順位であり、.979と高い値を示し、共分散比も.790と高い。面接4は、他の検査課題の共分散比と配点割合と比べて軒並み小さいのに対し、その差が.300以上となっており、配点以上に寄与の割合が高かったことが見て取れる。検査課題ごとの比較で言えば、他に、小論文も総得点との間に.791という高い相関を示しているが、共分散比が.097と低い。この両者の差異は、両者の標準偏差が20.40と2.83と大きく異なっていることから生じている。また、面接4の順位と総合点の順位の相関が異常に高いことも、データ数が少ないという側面は否めないが、面接1ー3での得点差があまりない中で、面接4での得点差がほぼそのまま総合得点の順位を決定付け、散布図で確認してみてもほぼ一直線であり、順位の入れ替わりが殆ど生じなかった。

更に、1次書類と課題作文、面接4および総合点との相関が低いことが分かる(注:1次得点は2次での合否に使用しない)。1次書類と総合点や面接4との相関が低いことは、1次で測っているものが2次では異なるという見解も成り立つかとは考えられるが、事前に提出し、多くは高校教員の指導が入っている1次書類と、受験生が大学の受験会場で直接その場で書く課題作文の評価の相関が低いことは、1次で選抜を行う場合には、あまり好ましい状況とは言えないだろう。

表2で、評価観点別の共分散比で細かく確認してみると、1次で言えば、書類2が、課題作文で言えば、観点1が、面接1・2・3では観点2・3・4・5が配点割合と共分散比の差が負に大きく、2次試験の総得点に対する寄与が低い。逆に、面接4の観点1・2・4では配点割合と共分散比の差がプラスに大きく、2次試験の総得点に対する寄与が高く、ほぼ面接4での結果が大きく合否に効いていることが分かる。

このように、一般化可能性理論を用いる以

前にも、こうした簡便な方法を用いることでも、各大学各学部の AO 入試（或いは、推薦入試）において、どういった選抜資料が、或いは、どういった観点が、選抜に大きく寄与したのかという、大まかな状況把握が可能である。

本稿の文脈で言うと、以上の結果も踏まえれば、相関や寄与の低いにも関わらず、多くの大学教員が採点や面接に参加している１次や課題作文の検査方法の在り方を見直したり、同じく相関や寄与の低い、２次の面接１ー３のあり方、特に、評価観点が８個もあることに問題がないのかを改めて吟味したりする必要があろう。

4. 面接官同士の評定一致度

次に、面接官の間での評定の一致度を検討した（表3)。まず、面接１ー３ごとにペアを作り、観点ごとに面接官のペア同士の順位相関を算出した。観点２の面接１及び２、観点３の面接１、観点５の面接２及び３、観点６の面接２及び３、観点８の面接１の順位相関が低く、全体のα係数も低い。面接官をひとり除外して順にα係数を再計算してみても、観点５の面接３を除いては、α係数の平均値・中央値に比して最大値がさほど向上しないことから、特定の人物の評定値が原因で順位相関を下げている訳ではなく、評価観点とそれに基づいて各面接で行われている具体的な質問項目に問題がある可能性がある。そこで、面接１ー３の評価観点の信頼性の確認、特に、観点２、観点３、観点５、観点６の信頼性を確認してみる必要がでてくる。

5. 一般化可能性理論による検討【１】

そこで、まず、一般化可能性理論による信頼性の検討を、１次書類（採点者７人、評価観点数3)、課題作文（採点者 13 人、評価観点数3)、面接１ー3(面接官各４人、評価観点数8)に対して行った。本研究における実験計画の概念図は、（$p \times h \times i$ デザイン）である（図1)。図２がG研究の結果であり、x軸の左から、受験生、採点者（面接官）、評価観点、受験生×採点者（面接官）の交互作用、受験生×評価観点の交互作用、採点者（面接官）×評価観点の交互作用のそれぞれの分散

表 3. 面接官同士の評定一致度と測定の安定性

		ペア同士の順位相関係数				α係数				
							1人除外した場合			
		平均	中央値	最大値	最小値	全体	平均	中央値	最大値	最小値
観点1	面接1	.434	.441	.731	.206	.750	.664	.647	.809	.553
	面接2	.486	.476	.799	.104	.744	.691	.699	.786	.581
	面接3	.372	.327	.661	.135	.534	.497	.424	.810	.329
観点2	面接1	.094	.154	.571	-.421	.346	.145	.346	.467	-.579
	面接2	.173	.169	.843	-.368	.687	.538	.495	.813	.348
	面接3	.481	.492	.727	.161	.577	.549	.490	.812	.405
観点3	面接1	.155	.208	.456	-.229	.340	.200	.342	.451	-.333
	面接2	.409	.418	.731	.077	.684	.573	.620	.717	.335
	面接3	.474	.453	.680	.269	.747	.676	.664	.760	.617
観点4	面接1	.555	.540	.671	.455	.700	.585	.544	.787	.463
	面接2	.322	.454	.512	.000	.535	.386	.428	.602	.087
	面接3	.289	.271	.537	.117	.502	.450	.470	.606	.252
観点5	面接1	.316	.354	.532	.048	.535	.459	.524	.624	.163
	面接2	.032	-.055	.689	-.318	.184	.074	.133	.323	-.294
	面接3	.154	.166	.562	-.314	-.025	.060	-.043	.622	-.296
観点6	面接1	.448	.469	.719	.157	.765	.695	.718	.784	.558
	面接2	.217	.176	.602	.000	.515	.410	.412	.533	.284
	面接3	.139	.104	.621	-.179	.332	.198	.155	.475	.008
観点7	面接1	.447	.418	.635	.289	.663	.548	.547	.745	.353
	面接2	.307	.297	.776	-.100	.659	.567	.590	.702	.385
	面接3	.265	.210	.501	.101	.400	.361	.355	.481	.253
観点8	面接1	-.215	-.302	.255	-.598	-.170	-.258	-.257	.278	-.795
	面接2	.388	.386	.692	-.092	.654	.581	.583	.649	.508
	面接3	.282	.279	.783	-.232	.489	.436	.420	.636	.266
総合点	面接1	.394	.374	.593	.267	.752	.667	.624	.821	.599
	面接2	.502	.464	.870	.269	.794	.736	.725	.831	.663
	面接3	.407	.513	.617	.108	.557	.531	.498	.779	.348

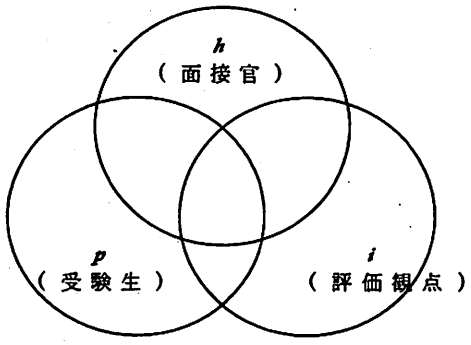

図1. 1次・課題作文・2次面接における実験計画の概念図

成分、及び、残差である。

また、Y軸の数値が分散成分の和に占める当該分散成分の割合である。その割合は、それぞれ、全体に占める、受験生の能力の散らばりの影響（受験生の主効果）、採点者（面接官）の甘さ/辛さの差の影響（採点者［面接官］の主効果）、評価観点の易しさ/難しさの影響（評価観点の主効果）、受験生ごとの採点者（面接官）との相性の影響（受験生×採点者の交互作用）、受験生ごとの評価観点への得手・不得手の影響（受験生×評価観点の交互作用）、面接官の評価観点の重視度の差の影響（採点者［面接官］×評価観点の交互作用）、及び、残差（誤差＋全ての要因の交互作用）を表している。

図2では、1次の採点者の評価基準が一貫性のなさの影響が最も割合が高くなっている[3]ものの、その後、2次の課題作文・面接での採点にあたりルーブリック作成に関する情報提供・作成支援を行い、その影響が減じたことが見て取れる。だだ、課題作文において、採点者の甘さ/辛さの影響（採点者の主効果）よりも、受験生ごとの採点者との相性の影響（受験生×採点者の交互作用）が強く見られた。受験生の評価を巡り、評価が採点者の中でまっぷたつに分かれたために引き起こされた結果である。

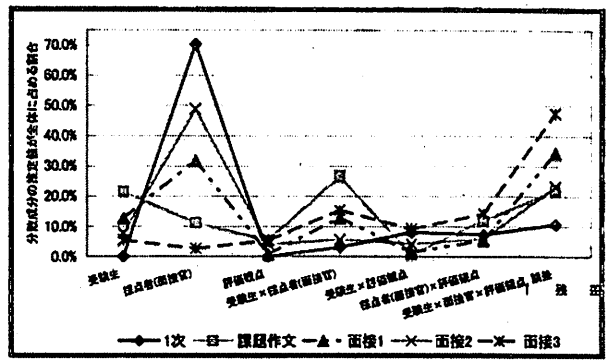

図2. 各要因の総得点に対する比率

また、図3がD研究の結果であり、課題作文での現行の評価観点数3に固定し、.7以上の信頼性係数(=一般化可能性係数）で安定した評定結果を得ると考えて、採点者4人で十分であり、例え.8以上としても採点者7人で十分となる。元の 13 人の半分 or 1/3 以下の採点者数で評定しても問題ないことが分かる。

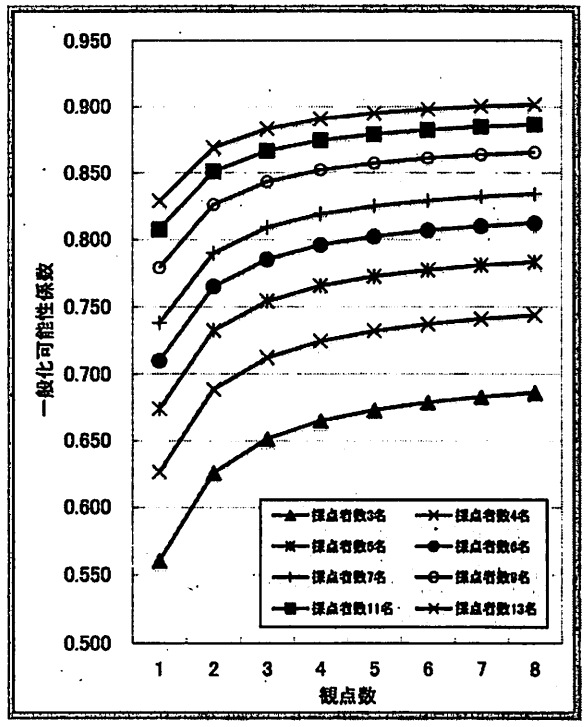

図 3. 課題作文の信頼性

6. 一般化可能性理論による検討【2】

前節では、面接1－3を個別の面接として扱ったが、一般化可能性理論を用いれば、複数の観点を同時に扱って分散・共分散成分の

推定を行うことも可能である。また、一般化可能性理論を用いることの利点1つとして、評価観点別に一般化可能性係数が算出でき、どの評価観点（及び、それに応じた質問項目）の適不適を吟味できる。そこで、一般化可能性理論を用いて、面接1ー3(評価観点数8、質問項目数［面接1ー3ごとに異なるため面接室数と一致]3、面接官4人　in 1 面接室）の信頼性評価を行った。本研究における実験計画の概念図は、$p^{\bullet} \times (h^{\circ} : i^{\bullet})$　デザイン）である（図4)。

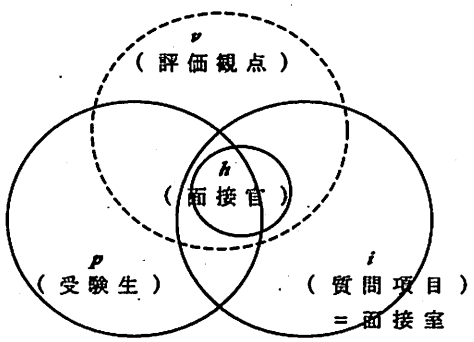

図4.　面接1ー3の実験計画の概念図
（$p^{\bullet} \times (h^{\circ} : i^{\bullet})$　デザイン）

図5で、評価観点ごとに分散・共分散成分の割合をみていくと、観点1・4・6は、受験生の能力の散らばり具合が効いており、受験生の能力をよく識別している。観点3は、受験生と質問項目（面接1ー3ごとに異なる）の相性による影響が大きく、観点3は、同じ観点で問いながら、面接1ー3ごとでの質問項目の問い方の違いによって、受験生の印象がまったく異なった質問項目であったことがわかる。観点4と7は、面接1ー3内部の面接官ごとの相性による影響が大きく、面接室内部での面接官の評価に差が見られた例でもあった。

また、図6で観点2と観点5を見てみると、一般化可能性係数の値が極端に低いことが分かる。質問項目を確認すると、問題解決能力や意欲・努力・関心といった新学力観的な観点別評価の質問項目であり、高校生にとって答えづらい質問でもあった。先の図5で観点2を見ると、受験生と質問項目の相性（受験生×質問項目の交互作用）の影響がもっとも高く、実際に面接の場面でも上手く答えられた受験生とあまり上手に答えられなかった受験生がいた印象がある。その反応から、表3で見たように、質問内容の曖昧さから面接官の評価が一貫しなかったと考えられ、次年度以降改訂が検討されるべき質問項目となった。

また、一般化可能性係数の値が高かった質問項目を確認すると、当該学部学科の専門的事項を問うたものが多く、やはり専門分野に関連するものであれば、採点者［面接官］同士の受験生に対する評価も一致し、かつ、受験生の善し悪しをよくよく識別するため、次年度以降の有力な質問項目候補として蓄積/検討されることとなった。

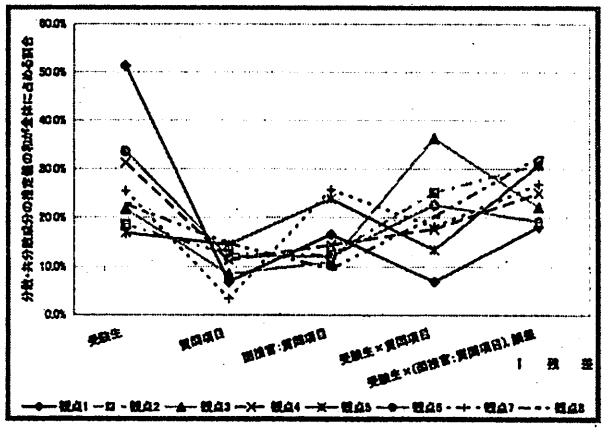

図5.各要因の総得点に対する比率

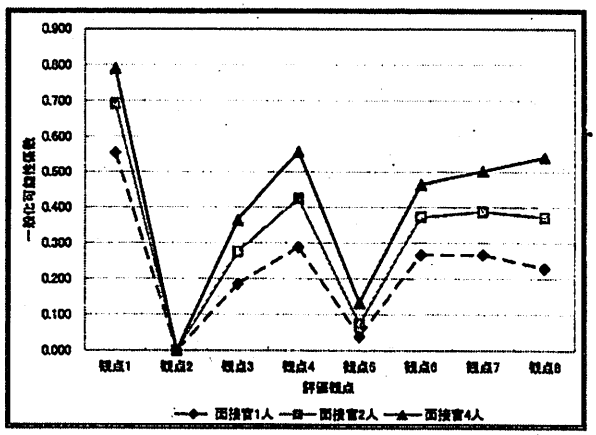

図6.評価観点別の一般化可能性係数

更に、図示はしていないが、多変量一般化可能性係数が現行の面接室1部屋あたり4人の場合、つまり、4人×3部屋 12 人で.753であるが、面接室1部屋あたり半分の2人にしても、つまり、2人×3部屋6人でも、多変量一般化可能性係数が.711となり、十分な信頼性が担保されることがわかる。

7. 結語－－AO入試の品質確保に向けて

そもそも、木村(2009)で指摘してきたように、AO入試は、選抜資料が複数であればあるほど、評価尺度が多元的であればあるほど、入学後の成績との相関が高いといった入学者選抜研究における重相関係数の誤用や選抜効果の誤読といった統計的な非常識と、じっくり手間暇かけた選抜の方がよりよい選抜だという世間の常識が不幸な婚姻関係を結んだことによって開始され、今日まで爆発的に拡大してきた。だが、全国的に蔓延した制度となり、学力を担保しない選抜方法という世間的な評価を得てしまった今の今になってさえ、その「適正規模」を考慮すべきと言った*縮減の論理*を持ち得ていないのがAO入試を巡る現況であろう。その論理を手にするためには、理念や概念の拡張では為し得ず、本研究で示してきたようにデータによる議論あってこそ成り立つ。そもそも、学力を担保しないだけに留まらず、テストとしての信頼性を担保する方向にも全くことが進んでいない。一般入試の2次科目の項目分析と同様に、どういった評価観点及びそれから導き出される質問項目が、受験生の能力をよく識別し、採点者（面接官）や質問項目に内在する、或いは、採点者（面接官）と受験性の相性などに内在する誤差を多く含むのかといった研究(=経験）を蓄積する方向性も現行の入試制度を継続するのであれば真剣に議論されるべきことであろう。

最後に、一般化可能性理論を、AO入試をはじめとした面接評価や小論文採点に適用した上での所感を述べてみたい。冒頭で述べた通り、一般化可能性理論は、米国発のテスト理論の文脈で開発されてきた。つまり、「米国発のテスト理論的な捉え方」からみれば、採点者（面接官）間の評価の一致度が高いほど、つまり、採点者（面接官）において受験生の優劣が揃うほど、誤差が小さくなる、ということで信頼性が高まる仕組みになっている。

この考え方を踏襲すれば、例えば、図2でみたように、仮に、課題作文の評価で採点者の受験生に対する選り好みが出たとなれば、その評価は「採点者（面接官）の受験生に対する選り好み」という誤差が大きいという理由で「失敗」と結論づけられる。勿論、しっかりと採点基準（ルーブリック）を定めてあることが前提として担保されていたとしても、今回の結果のように、大学入試の現場では、受験生が書いた内容についての採点者（面接官）の理解の度合いや評価がまったく異なり、評価がまっぷたつに割れることも往々にして起こりうる。そこに例えばアドミッションセンターの教員が入試現場で介入するのは時間的にも労力的にも（勿論、人間関係的にも）至難の技と言わざるを得ない。今回の場合であれば、図2を見ても、採点者（面接官）や評価観点の主効果はともに低く、採点者（面接官）の採点のバラツキや評価観点のバラツキがないことが見て取れるので、純粋に「解釈」の質とそれに対する「評価」の違いが反映されたと考えられる。

この誤差の「価値付け」というところまで考えると、一般化可能性理論を大学入試の現場に適用することが、十把ひとからげで、そのまま受け入れ、大学入試の現場からあらゆる誤差をなくす方向に進むことが必ずしも現実的・建設的であると思えない側面もある。もっと言えば、一般化可能性理論では、純粋に要因ごとに統制した際の得点データのバラツキから読み取れる誤差を評価しているに過

ぎず、採点者（面接官）の解釈から点数化する過程、つまり、各採点者（面接官）の「解釈の質」には踏み込んだものでは必ずしもないことには、十分に注意を払わなければならない。

極端な例で言えば、一般化可能性理論を突き詰め、信頼性を高めようとすれば、そういった「解釈の質」を問わない項目、例えば、「面接の最初に挨拶はできているか」とか「文法ミスはないか」とか「誤字脱字はないか」など、ものごとの有る / 無しが比較的はっきりした、誰もが評価のぶれない、非常に簡単な質問項目で揃えた方が、信頼性が高まりやすいという側面がある。こういった構造的な視点から見たとき、一般化可能性理論での結論ありきで、面接評価や小論文採点を演繹的に機能させてしまうと、中身のある良い採点表ができず、それに応じて、聞きたい / 聞くべき質問ができず、という場面に必ず陥ってしまう。このことは、一般化可能性理論を、AO入試をはじめとした面接評価や小論文採点に適用する際には、勘案しなければならない事柄となる。

勿論、高校側から首を傾げる合否結果を出さないためにも、大学教員である採点者（面接官）が満場で一致して良い生徒だと思う受験生から合格させるという側面から見れば、「採点者（面接官）間の評価の一致度が高いほど、つまり、採点者（面接官）において受験生の優劣が揃うほど、誤差が小さくなる、ということで信頼性が高まる仕組み」の「テスト理論的捉え方」はとても有効なツールとなることは間違いない。本稿で扱ったように、適正規模を保った上で、質問項目の善し悪しを判別するツールとして、一般化可能性理論から得られた蓄積——例えば、今回のように、信頼性を担保した上で、採点者の人数を定め、専門家が専門的事項を見分ける識別眼的質問項目のリストアップやノウハウの蓄積を行うこと——は、十分に大学入試の改善のためには機能するであろう。

一般化可能性理論による結論が、絶対的ではないまでも、大学入試改善の新たな方法論の一つとして、教育制度論的な考察や追跡調査等の結果も踏まえ、包括的に大学入試制度・方法を考える一助として活用される方向性が、現実的かつ建設的であると考えている。

【謝辞】本稿の完成にいたるまでには、査読者の先生がたに、数多くの貴重なアドバイスを賜りました。ここに感謝の意を表します。

【付記】本研究は、長崎大学における大学高度化推進経費「新任教員の教育研究推進支援経費」の研究成果の一部である。

注

1）五分位図とは、受験生全体を上から20%ずつの5つのグループに分け、それぞれのグループにおける得点率をグラフ化したものである。簡便な方法ではあるが、こうした手法を用い、グラフの傾きを見ることで、各設問ごとの大まかな識別力を検討することが可能となる。

2）（多変量）一般化可能性理論は、Brennan (1996, 2001a)、Feldt & Brennan(1989)、Shaverson & Webb (1991)、平井(2007)、池田(1994)、池田(2007)、角幸(2006)、中村(2003)、対馬(2007)、豊田(1994)などで紹介されている。一般化可能性理論における分散成分の推定にはSPSS 15.0 Advanced、多変量一般化可能性理論における分散・共分散成分の推定にはmGenova (Brennan 2001b) をそれぞれ利用した。

3）Brennanは、先行研究をレビューして、評定者の評価の一貫性のなさの影響が小さく、人（本研究の文脈でいうなら、受験生）と課題の交互作用の影響が大きいことを報告している (Brennan1996: 26)。だが、そもそもBrennanの報告では、用意周到に訓練された評定者が前提となっており、本研究で見てきたように、面接や作文評価においてそもそも訓練された評価者を前提とすることが難しい大学入学者選抜の文脈では当てはまり難い。この点は、平井(2007)にも指摘がある。

引用文献

坂野永理 (2008). 「一般化可能性理論による日本語口頭プレースメントテストの検討」『日本テスト学会誌』4(1)、23-32.

Brennan, R. L. (1996). "Generalizability of Performance Assessment", G. W. Phillips ed. *Technical Issues in Large-Scale Performance Assessment,* NCES96-802, 19-58.

Brennan, R. L. (2001a). *Generalizability Theory*, New York: Springer.

Brennan, R. L. (2001b). *Manual for mGenova Version 2.1.*, Iowa Testing Programs Occasional Papers No.50, 1-83.

Feldt, L.S. and R. L. Brennan (1989). "Reliability," Robert L Linn (ed) *Educational Measurement*, Third Edition, American Council on Education, (池田央監訳 (1992). 「第3章　信頼性」ロバート・L・リン編『教育測定学第3版　上巻』147-209.)

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」『人文学報』380、25-64.

池田央 (1994). 「一般化可能性理論」『現代テスト理論』朝倉書店、28-50.

池田央(2007). 「一般化可能性理論」『統計データ科学事典』朝倉書店、638-9.

角幸康太郎 (2006). 「一般化可能性理論――採用面接の信頼制度を測る」『購買心理を読み解く統計学』東京図書、215-222.

木村拓也 (2008). 「アドミッションセンターの系譜学――何故、そしてどのような入試研究が求められたのか」『日本テスト学会第6回大会発表抄録集』、88-91.

木村拓也 (2009). 『大学入学者選抜における測定評価技術の不在と誤謬』東北大学大学院教育学研究科論文博士学位請求論文.

熊澤孝昭 (2006). 「一般化可能性理論を用いた相互評価における学生評定者の信頼性の検討」『立教ランゲージセンター紀要』15、55-64.

前田忠彦 (2004). 「法科大学院統一適性試験における『表現力を試す問題』の採点と分析」『法科大学院統一適性試験テクニカルレポート 2004』商事法務、100-117.

宮本学・出口寛文・北浦泰他 (2006). 「一般化可能性理論を用いた客観的臨床能力試験における評価の信頼性の検討」『大阪医科大学雑誌』65(3)、196-201.

中村健太郎 (2003). 「一般化可能性理論」「多変量一般化可能性理論」『共分散構造分析［技術編］―― 構造方程式モデリング』71-89.

Shaverson, R. J. and Webb, N. M. (1991). *Generalizabitity Theory; A primer,* Newbury Park: Sage Publications.

対馬栄輝 (2007). 「検者間・検者内信頼性係数」『SPSSで学ぶ医療系データ解析』、東京図書、195-214.

豊田秀樹 (1994). 『違いを見抜く統計学――実験計画と分散分析入門』講談社ブルーバックス.

豊田秀樹・中村健太郎(2004)「大学における授業評価の信頼性――一般化可能性モデルと構造方程式モデリングによる4相データの解析」『心理学研究』75(2)、109-117.

山森光陽 (2003). 「中学校英語科の観点別学習状況の評価における関心・意欲・態度の評価の検討」『教育心理学研究』51、195-204.

吉村宰 (2007). 「長崎大学AO入試における書類選考データの分析」『大学入試研究ジャーナル』17、39-42.

吉村宰 (2008). 「AO入試選考書類のテキストマイニング」『平成20年度全国大学入学者選抜研究連絡協議会大会（第3回）研究発表予稿集』（取扱注意）、169-172.

吉村宰・南部広孝 (2008). 「AO入試による入学者の入学後成績と選抜方法――選抜方法改善の観点から」『大学入試研究ジャーナル』18、187-192.

鷲坂由紀子・入江崇介・内藤淳・二村英幸 (2004). 「昇格選考における論文評定の分析――多変量一般化可能性理論を用いた信頼性の検討」『経営行動科学会年次大会発表論文集』7、33-39.