2012 年度 日本テスト学会誌　Vol.9, No.1, 145 - 165

展望論文

論述式テストの運用における測定論的問題とその対処

Measurement Problems and Their Treatments in Essay-Type Tests

宇佐美 慧[1]

Satoshi Usami[1]

[1]日本学術振興会・南カリフォルニア大学

[1]Japan Society for the Promotion of Science・University of Southern California.

Correspondence concerning this article should be sent to: Satoshi Usami, Department of Psychology, University of Southern California, 3620 S McClintock Ave., SGM 826, Los Angeles, CA. 90089.　E-mail: su_686@usc.edu

論述式テストの運用における測定論的問題とその対処

宇佐美 慧[1]

[1]日本学術振興会・南カリフォルニア大学

思考力，文章力，表現力，独創性といったキーワードに代表される，受験者のもつ「高次の能力」を如何に適切に測定・評価するかについては，国内外で以前から高い注目を集めている問題である．日本でも，特に小論文試験に代表されるような，ある程度まとまった文章で回答させるテスト形式である論述式テストの普及が，入学試験や人事試験をはじめとして近年特に顕著である．その一方で，論述式テストは，測定の信頼性・妥当性・バイアスといった多くの測定論的問題を抱えるのが常であるが，この問題の概要について整理し，また論述式テストの運用の改善のための具体的な指針を得る試みは十分になされてこなかったと言える．

そこで本論文では，特に近年の国内外の論述式評価研究を俯瞰しながら，論述式テストにおける測定論的問題の内容について整理していく．そして，項目作成・採点・フィードバック等のテストの各運用段階と，関連する各々の測定論的問題を対比して整理することを通して，実践上重要である検討事項や工夫すべき点を整理・展望し，テスト開発の実務家やそれを支援する立場にある教育測定・評価の専門家にとって参考となる知見を提供していく．

キーワード：論述式テスト，小論文試験，教育測定，教育評価，パフォーマンス評価

Measurement Problems and Their Treatments in Essay-Type Tests

Satoshi Usami[1]

Japan Society for the Promotion of Science・University of Southern California.

There has been increased awareness of the importance of assessing “higher-level abilities” such as abilities of thinking, written expression, and creativity in recent years. For this purpose, the use of essay-type tests has become popular, especially in university entrance and personnel examinations. However, many researchers have pointed out that essay-type tests are subject to measurement problems such as reliability, validity, and bias, while these measurement problems and their treatments have not been fully addressed so far.

This paper overviewed these measurement problems in the literature, and summarized how these problems can be related to each phase in the construction and operation of essay-type tests. Based on this review, this paper provided test practitioners and measurement specialists with critical issues to be considered and how they can be addressed in practice.

Keywords：essay-type test, educational measurement, educational assessment, performance assessment

1 序論

資格社会という言葉によく表れているように，今ほどテスト*1のもつ社会的な影響力が大きい時代はないだろう（宇佐美, 2012）．そして，テストのもつ社会的な重要性が高まるのと並行して，テストの運用の仕組みも大きな変化を遂げてきた．例えば，テストの実施方法に関しては，大規模試験であるTOEFLに代表されるようなコンピュータ型試験の開発や，それに伴う項目反応理論の活用が挙げられる(例えば，田栗・藤越・柳井・ラオ, 2007).

他にも，テストの運用の仕組みの変化に関しては，とりわけ日本の特徴的な点として，テストの回答形式が指摘できる．それには例えば，小論文試験*2に代表されるような，ある程度まとまった文章で回答させるテスト形式である論述式テスト*3の利用が1980年代頃から急速に進んだことが挙げられる（宇佐美, 2012）．この背景には，応用的な思考能力，表現力，実技能力，独創性などといった，受験者のもつ「高次の能力」を評価する試みが，特に入学試験をはじめとして増えたことが深く関係している（宇佐美, 2010）．実際に，文部科学省(2012)によれば，現在では国公立大学の7-8割程度が小論文試験を何らかの形で利用しており，その利用頻度は長い間高い水準にある．また，論述式テストの利用および開発は入学試験に限らず，適性試験・人事試験・資格試験など様々な場面において見られる(例えば，成田・荘島・宇佐美, 2010).

しかし，論述式テストは人間のもつ複雑で多様な能力を測定する目的からすれば魅力的な手段であるが，現実には測定の信頼性や妥当性と言った，様々な測定論的問題を伴うことが教育測定・評価研究では古くから知られている（例えば，石井, 1981；平・江上, 1992; 渡部・平・井上, 1988; 宇佐美, 2010）*4．論述式テストの魅力とは裏腹に，この測定論的問題が極めて複雑であることを，テストの実務家や教育測定・評価を専門とする研究者，およびテストの運用に携わる関係者の多くは実感した経験があるだろう．また，実際に論述式テストを受けて，その評価結果に対する違和感や不公平感を覚えた経験のある人も少なくないだろう．

日本では，小論文評価或いは作文評価の測定論的問題について，ここ二，三十年の間に様々な検討がなされてきたが(阿久津・嶋野他,2005；平井, 2002, 2006, 2007；平井・椎名・柳井, 2001；池田, 1992；井上, 1996；梶井, 2001,2002；黒岩, 1991；佐渡島, 2003；平, 1995；平・江上, 1992；宇佐美,2008,2009a,2011,2012; 渡部, 1994；渡部他, 1988；渡部・平井, 1993)，現行の論述式テストにおいて十分な改善が見られたとは到底言えないのが事実であろう．このような結果，現行の個々の論述式テストにおいては，測定論的な意味としての品質に大きな差異が生じ易く，またその品質が客観的な形で十分評価されていないのが実態である．また，教育測定の専門家の不在を指摘する意見(木村, 2010)を考慮すれば，この現状が今後さらに深刻化するのは明らかである．

このような現状があるのは，研究遂行上の人的・経済的・時間的コストの問題と測定論的問題に関連する要因の複雑性・多様性の問題から，実証的な論述式評価研究の蓄積が不十分であったこと(宇佐美, 2008,2011; 渡部, 1994)も大きな理由と思われる．他にも，テスト開発の実務家やそれを支援する立場にある教育測定・評価の専門家にとって実際的な手立てが不十分であったことも大きいと考えられる．例えば，測定論的問題に関わる諸要因やそれらの影響度をまとめ，どのような解決策が講じられるべきかについて整理することは重要であるが，そのような試みはほとんどなされてこなかった．そして，既に広く利用されている小論文試験の場合であっても，その測定・評価研究に関する包括的な国内のレビューは平・江上(1992)以降，20年間なされていない．また，日本テスト学会(2007)によるテスト・スタンダードでは，2.9節(pp.86-90)で「主観的な評定による採点」の節が設けられており非常に参考になるが，紙幅の都合もあり，多様な測定論的問題を十分に概観し整理されたものではない．

さらに，より根本的な問題として，特に教育測定を専門としない実務家の間で測定論的問題が十分に認識されていないことが挙げられる．吉村他(2008)が指摘しているように，「全国大学入学者選抜方法研究協議会での各大学からの報告や，入試研究ジャーナルに掲載されている論文を見る限り，ごく一部を除きテストの質への関心はほとんどない」のが現状であろう．また，日本テスト学会(2007, pp.156-159)の記述の中にも色濃く表れているような，作文や面接を用いた評価を行う際の，テスト作成者と教育測定の専門家の間にある認識のズレは決して稀なことではない．これらは，倉元・當山・西郡(2008)や吉村他(2008)が指摘しているように，「テストの品質」という概念が実務家の間でほとんど定着していないことを意味している．

このような問題意識から，本論文では，テスト開発の実務家やそれを支援する立場にある教育測定・評価の専門家にとって参考となる知見を提供し，現行の論述式テストの運用の改善に資することを目的として，宇佐美(2009a,2011)の内容を中心に国内外の研究を整理しながら，論述式テストの測定・評価研究についてのレビューと展望を行う．まず，信頼性・妥当性・バイアス*5の三つの観点から測定論的問題の概要を整理する（第2節）．そして，それらの内容を項目作成・採点・フィードバック等のテスト運用の各段階と対比した表にまとめて整理していき（第3節），最後に総合考察を行う（第4節）．

2 測定論的問題の概観

2.1. 信頼性の問題

(1) 客観式テストと論述式テストにおける信頼性

まず，論述式テストの前に，客観式テストにおける信頼性の意味を考えてみよう．いま，表1のように，n_I 個の項目に対する，受験者 N 人分の採点データが得られているとする．このとき，j 番目の受験者の，i 番目の項目についての得点を y_{ji} とすると，j 番目の受験者の合計得点 $y_j = y_{j1} + y_{j2} + \cdots + y_{jn} = \sum_i^{n_I} y_{ji}$ がどれだけ安定しているかが信頼性の問題である*6．つまり，仮に，別の異なる n_I 個の項目を用意して再度各受験者がテストを受けることができたとして，一度目に高い（或いは低い）合計得点 y_j であった受験者が二度目も高い（或いは低い）合計得点 y_j となるような一貫性が，受験者全体においてどれだけ見られているかということである．それでは，信頼性を高めるためには何が必要であろうか．

この点を，テストによる測定を数理的な観点から扱う理論体系である，古典的テスト理論を通して考えてみよう．古典的テスト理論においては，ある測定機会 t における y_j の値を意味する測定値 y_{jt} を，真値 t_j と，真値とは無関係に存在する測定誤差 e_{jt} の和，すなわち，

$$y_{jt} = t_j + e_{jt} \qquad (1)$$

とモデル化する．そして，これらの個人差の大きさ，すなわち分散を順に $\sigma_y^2, \sigma_t^2, \sigma_e^2$ と表す．このとき，モデル上の仮定から $\sigma_y^2 = \sigma_t^2 + \sigma_e^2$ の関係が成立するが，σ_y^2 のうち σ_t^2 の占める割合，すなわち

$$\rho = \frac{\sigma_t^2}{\sigma_y^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \qquad (2)$$

を測定の信頼性として定義するのが，信頼性係数 ρ と呼ばれる指標である．

ただし，勿論，個々の測定値 y_{jt} は実際に観測される一方で，真値 t_j 或いは測定誤差 e_{jt} の値は直接知ることができない．そこで，「テストに含まれる各項目が同一の能力を一貫して測定できている程度」を意味する内的整合性の観点から，(2)式の信頼性係数 ρ を，

$$\alpha = \frac{n_I}{n_I - 1}\left(1 - \frac{\sum_{i=1}^{n_I} \sigma_{yi}^2}{\sigma_y^2}\right) \qquad (3)$$

を用いて推定するのが，クロンバックの α 係数 (Cronbach, 1951)と呼ばれる指標である．ここで，$\sigma_{y_i}^2$ は項目 i の得点の分散を意味する．(3)式より，α が大きくなるためには，n_I を大きくすること，すなわち項目数を増やすことが必要と分かる．また同様に α が大きくなるためには，σ_y^2 を大きくする必要があることも分かる．ここで，σ_y^2 が $\sigma_y^2 = \sum_{i=1}^{n_I} \sum_{i'=1}^{n_I} \sigma_{yii'}$ のように項目間共分散の和で表されることに注意すると，σ_y^2 を大きくするためには，項目間共分散，すなわち項目間相関を大きくする必要があることが分かる*7．これらより，客観式テストの文脈では，信頼性を高める際に核となるのは項目数と項目間相関であると言える．

ところが，論述式テストの信頼性を考える際には，ここまでの内容((1)-(3)式)だけでは不十分である．論述式テスト評価では測定の信頼性を高める意味や，また運用の人的・時間的コストの都合から，複数の採点者がいる場合や複数回の採点が行われる場合が多い．その結果，論述式テストの場合，表2のような受験者×項目×採点者，或いは受験者×項目×採点機会の三相のデータとなることが一般的である．つまり，表1のような受験者×項目の二相の形式である客観式テストの場合に比べてより複雑なデータの形式となる．ここまでの内容から論述式テストの信頼性を考える際の大きな問題点は，(1)式の誤差 e_{jt} の源泉には，項目だけでなく採点者等様々な要因の影響も含まれるため，e_{jt} だけではこのような誤差の多様性を表現し尽くせないという点にある．この理由から，α 係数の提案者であるCronbach自身も，α 係数から三相データの信頼性を考えることは不適切であると指摘しており，代替案として一般化可能性理論(Brennan, 2000, 2001; Cronbach, Nageswari & Gleser, 1963)を利用することを推奨している(Cronbach & Shavelson, 2004）*8.

表1: 客観式テストにおける採点データ(二相)

	項目 1	項目 2	…	項目 n_I
受験者 1	y_{11}	y_{12}	…	y_{1n_I}
受験者 2	y_{21}	y_{22}	…	y_{2n_I}
受験者 3	y_{31}	y_{32}	…	y_{3n_I}
受験者 4	y_{41}	y_{42}	…	y_{4n_I}
…	…	…	…	…
受験者 N	y_{N1}	y_{N2}	…	y_{Nn_I}

表2:論述式テストにおける採点データ（三相：2名の採点者による完全クロスデザインの場合）

	採点者 1				採点者 2			
	項目 1	項目 2	…	項目 n_I	項目 1	項目 2	…	項目 n_I
受験者 1	y_{111}	y_{121}	…	y_{1n_I1}	y_{112}	y_{122}	…	y_{1n_I2}
受験者 2	y_{211}	y_{221}	…	y_{2n_I1}	y_{212}	y_{222}	…	y_{2n_I2}
受験者 3	y_{311}	y_{321}	…	y_{3n_I1}	y_{312}	y_{322}	…	y_{3n_I2}
受験者 4	y_{411}	y_{421}	…	y_{4n_I1}	y_{412}	y_{422}	…	y_{4n_I2}
…	…	…	…	…	…	…	…	…
受験者 N	y_{N11}	y_{N21}	…	y_{Nn_I1}	y_{N12}	y_{N22}	…	y_{Nn_I2}

⑵ 一般化可能性理論

一般化可能性理論は，分散分析の方法を利用する，古典的テスト理論の一つの拡張である（Linn, 1989）．一般化可能性理論では，まず個々の採点結果の違いが，受験者・採点者・項目・採点機会といった異なる要因(相)からどの程度説明・予測できるかを推定する(G研究)．そして，これらの要因が採点結果の信頼性に与える影響を，一般化可能性係数という指標を通して見積もることで，一定の信頼性を得るのに必要な項目数・採点者数・採点回数を推定する(D研究)．日本の小論文・作文評価研究においては，池田(1973)内の問題や渡部他(1988), およびLinn(1989)を嚆矢として，平井(2007), 梶井(2001), 平(1995), 宇佐美(2009a,2011)など多くの適用例がある．

いま表2のような, 各採点者が全受験者の全項目に対する回答を採点する完全クロスデザインの状況を仮定する．この場合，j 番目の受験者の i 番目の項目への回答に対する，採点者 r による採点結果を意味する y_{jir} を，分散分析モデルでは,

$$y_{jir} = \mu + \pi_j + \alpha_i + \beta_r + \pi\alpha_{ji} + \pi\beta_{jr} + \alpha\beta_{ir} + \pi\alpha\beta_{jir} \quad (4)$$

と分解して表現する．ここで，μ は全ての採点結果についての総平均を，π_j, α_i, β_r はそれぞれ受験者 j，項目 i，採点者 r の主効果を，$\pi\alpha_{ji}, \pi\beta_{jr}, \alpha\beta_{ir}$ はそれぞれ対応する一次の交互作用効果を, そして $\pi\alpha\beta_{jir}$ は誤差（二次の交互作用効果）を意味する．いま，受験者・項目・採点者がそれぞれランダムに選ばれていると仮定すると，μ を除く各項について対応する分散を考えることができる．いま，その分散を $\sigma^2()$ で表す．例えば，$\sigma^2(j)$ は受験者の主効果の分散である．さらに，μ を除く各項の期待値は，分散分析モデル上の仮定から，全て 0 である．つまり，$E(\pi_j) = E(\alpha_i) = \ldots = E(\pi\alpha\beta_{jir}) = 0$ である．ここで $E()$ は期待値を表す.

一般化可能性理論の中の重要な概念にユニバース得点がある．ユニバース得点は，各受験者 j への採点結果の項目と採点者に関する期待値, つまり $E_i E_r(y_{jir})$ として定義される量であり，(4)式からこれは $\mu + \pi_j$ と計算される．また $\mu + \pi_j$ の分散は，π_j の分散に等しいため，ユニバース得点の分散は $\sigma^2(j)$ となる.

一方，実際の各受験者 j への採点結果 y_{jir} は，ユニバース得点の場合とは異なり，$\sigma^2(j)$ だけではなく，受験者 j に起因する全ての交互作用効果が測定誤差と

して含まれてしまう．そこで，ユニバース得点の分散に対する測定誤差分散の比を用いて定義されるのが，一般化可能性係数と呼ばれる指標である．具体的に，一般化可能性係数は，ユニバース得点の分散(分子部分)と，実際の各受験者 j への採点結果 y_{jir} の平均値についての分散の期待値(分母部分)の比として，

$$\frac{\sigma^2(j)}{\sigma^2(j)+[\frac{\sigma^2(ji)}{n_I^{'}}+\frac{\sigma^2(jr)}{n_R^{'}}+\frac{\sigma^2(jir)}{n_I^{'}n_R^{'}}]} \quad (5)$$

のように定義される．ここで $n_I^{'}, n_R^{'}$ は，後述のように分析者が別途設定する項目数と採点者数である．(5)式の[ ]内の値，すなわち受験者×項目，受験者×採点者の交互作用効果および誤差に関する項が，実際の採点結果に含まれる測定誤差分散を表す．したがって[ ]内の値が小さいほど，ユニバース得点の分散に対して測定誤差分散が小さくなるため，一般化可能性係数は大きくなる．

分散成分 $\sigma^2(j),\sigma^2(ji),\sigma^2(jr),\sigma^2(jir)$ に関しては，(4)式の分散分析モデルから推定される平均平方を用いて，それぞれ

$$\begin{aligned}
\sigma^2(j) &= \frac{MS(j)-MS(ji)-MS(jr)+MS(jir)}{n_I n_R} \\
\sigma^2(ji) &= \frac{MS(ji)-MS(jir)}{n_R} \\
\sigma^2(jr) &= \frac{MS(jr)-MS(jir)}{n_I} \\
\sigma^2(jir) &= MS(jir)
\end{aligned} \quad (6)$$

の式より推定される．ここで $MS(\ )$ は対応する要因についての平均平方の推定値であり，n_I, n_R は実際の項目数および採点者数である．これらより，一般化可能性理論による信頼性評価の手続きは，まず(6)式を利用して各要因の分散を推定し(G研究)，そして $n_I^{'}, n_R^{'}$ の値を様々に変化させて(5)式の値の推移を見ながら，一定の一般化可能性係数を得るのに必要な項目数・採点者数を推定すること(D研究)とまとめられる．

ここで重要なのは，(5)式に注目すると，論述式テストの場合，項目数 n_I だけでなく採点者数 n_R も信頼性を高める上で必要という点である．なお，表2のデータを受験者×項目×採点機会の三相のデータとみれば，同様の理由から採点回数も信頼性を高める上で必要とわかる．さらに項目間相関および採点者間の採点結果の一貫性である採点者間相関が高いということは，特定の項目および採点者による評価結果の偏りが小さいことをそれぞれ意味する．つまり，この場合(5)式の[ ]内の誤差分散 $\sigma^2(ji),\sigma^2(jr),\sigma^2(jir)$ が小さくなることを意味する．したがって，論述式テストでは項目間相関以外に採点者間相関を高めることも必要である．さらに，表2のデータを受験者×項目×採点機会の三相のデータとみれば，やはり同様の理由から，同一採点者内の採点結果の一貫性である採点者内相関も信頼性を高める上で必要とわかる．

論述式テストデータへの一般化可能性理論の適用例に 宇佐美(2009a,2011)の研究がある．あくまで単一の研究事例であるが，宇佐美(2009a,2011)は，(a)採点者数よりも項目数を増やした方が一般化可能性係数の増加が期待できること，(b)採点者数は4名を超えると一般化可能性係数への効果が頭打ちになること，(c)0.8以上の一般化可能性係数を確保する為には，採点者数が1名の場合は10以上の項目数が必要であり，また採点者数が2～3名であれば3～4程度の項目数が，さらに採点者数が4～5名程度であれば2つの項目数が必要であることを示している．

大学入試においては単一の項目に基づく論述式テストが課されることがしばしばある．しかし，平井(2007)なども指摘しているように，これは測定の信頼性の観点からすれば極めて危険である．関連して，宇佐美(2012)のシミュレーション研究でも，項目数や採点者数は一部を集中的に増やすよりも万遍なく増やす方が，信頼性を高める目的からすれば効率的であることが示唆されている．

なお(5)式より，α 係数は二相データ，すなわち $n_R^{'}=1$，$\sigma^2(jr)=\sigma^2(jir)=0$ の状況において，$n_I^{'}=n_I$ とした場合に一致することが分かる．また，採点者によって採点する項目が異なるネストデザインの場合でも，(4)-(6)式と同様の定式化により一般化可能性係数を推定する手続きが得られる．この点を含む更なる理論的な詳細は，Linn (1989)や池田(1994)が参考になる．

このように，一般化可能性理論は，二相の場合は勿論のこと，三相である論述式テストデータの場合においても採点結果の信頼性を一貫した手続きで考えることができる強力なツールと言える．さらには，Cronbach & Shavelson (2004)でも指摘されているように，各分散成分の推定値や(5)式の[ ]内に相等する採点結果の測定誤差分散を報告することで，採点結果に含まれる誤差の大きさとその要因別の割合を示すことができるのも利点である．

(3) 採点者間相関・採点者内相関

上述のとおり，論述式テストにおいて信頼性を高めるためには，採点者間相関および採点者内相関を高める努力が必要である．そのためには，例えば採点者の

訓練を行うことや，採点基準の決定に関する事前協議を綿密に行うなどの手続きが重要になる（平井, 2007; 宇佐美, 2012；この点は, 3.4節にて後述）. 宇佐美(2012)でも指摘されているように，期待される採点者間相関の大きさは，当該テストが測定する能力・適性の違いや，他にも採点者の熟練度に強く依存する．例えば平井(2008, pp.81-100)が報告しているように，綿密な事前協議と採点者訓練を行っている米国の全国学力調査（主調査）における解答構築式（論述式）項目では，複数の採点者による採点結果の（信頼性そのものとは異なる概念であるが，異なる採点者間で共通の値が仮定されている採点者間相関の大きさを反映する指標である）級内相関は0.8以上の高い値が報告されている．

その一方で，宇佐美(2009a)では，十以上の国内外の先行研究のレビューの結果から，最低限の事前協議をしているケースであっても，平均的には採点者間相関は0.3～0.4程度であることが報告されている．むしろ，場合によっては，これらよりも低い事例もある(e.g., Hayes, Hatch & Silk, 2000; 宇佐美, 2008)．採点者内相関について報告した研究は少ないが，例えば渡部他(1988)や池田(1992)は，概ね0.4～0.9程度の値を報告しており，一般に採点者間相関よりも高い値が期待される．

⑷ 採点のカテゴリ数・採点方法

他にも，実際の採点デザインや採点方法は複雑であるため，採点結果の信頼性を考える上で関連する問題がある(宇佐美, 2012)．例えば，論述式テストでは，五段階や七段階採点などの何らかのカテゴリ数を持つカテゴリ得点を通して採点が行われることが多いが，このカテゴリ数の設定は信頼性に関わる問題である．(1)式や(4)式にもあるように採点結果 y_{jir} には身長や体重のような連続的な量が暗に想定されている．一方で，カテゴリ化によって採点結果を離散化する際には情報の損失が生じるため，それが信頼性に対してネガティブな影響を与えることが一般に知られている．

実際の採点結果の信頼性がカテゴリ数の設定に応じてどのように変化するかを述べた実験的な事例報告が池田(1992)に，また100点満点で採点されたデータに対して得点を事後的に離散化することによって，後述する項目反応理論の文脈で言う情報量がどのように変化するかを検討した研究(平井・渡部，1994)がある．その結果，いずれの研究でも5カテゴリ程度で十分であるという示唆を得ている．つまり，5カテゴリ以上の細かなカテゴリ数に基づく採点は，採点者にとって負担を増やすだけでなく，信頼性の改善度も大きくないということである．興味深いことにこれは宇佐美(2012)のシミュレーションの結果とも整合的である．

さらに，採点者間で採点結果の不一致が生じた場合にどのように得点を処理するかも信頼性に関わる問題である．このような場合，単に異なる採点結果を認めてそのままにするケースもあるが，例えば他とは外れた採点結果を与えた採点者の結果を除いて得点を処理する方法や，仮に専門の採点者と非専門の採点者の双方がいれば前者による結果のみを優先する方法など様々な工夫がとられる．関連して，仮に複数の採点者による採点結果の平均値が小数点を含む場合，小数点以下を切り捨てるか（または切り上げるか），或いは四捨五入するのか，といった問題もある．これらの問題はまとめて，採点者間で採点結果の不一致が生じた際の補正方法の問題(e.g., Johnson, Penny, & Gordon, 2000; Penny & Johnson, 2011)として知られている．

これら以外にも，一つの総合的な評価観点から単一採点結果のみを与えるのか，或いは構成力・表現力・語彙力など，予め設定された複数個の評価観点に基づいた複数の採点結果を与えるのかといった，総合評価と分析的評価の選択も，信頼性に関わる問題である（e.g., Barkaoui, 2007）．評価法の選択については2.2(3)節の因子的妥当性の内容にも深く関わるため，そこで改めて触れることにする．

⑸ その他の話題

これまで見てきたように，論述式テストの採点結果の信頼性に関わる諸要因は項目間相関，項目数，採点者間相関，採点者数，採点者内相関，採点回数，採点のカテゴリ数，採点方法などに集約でき，これらは非常に複雑で多面的である．この複雑性からすると，信頼性に影響を与える各要因が相対的にどの程度の重要度を持っているのか，そしてどの要因に関わる工夫を優先すべきかについての指針を得ることも重要であろう．この観点から， 宇佐美(2012)は，採点結果の信頼性および選抜の正判別率に対する，採点者数，項目数，採点のカテゴリ数，採点者間相関・採点者内相関，採点者間で採点結果の不一致があった際の補正方法，採点結果の分布，合格率等の様々な試験実施上の諸要因の影響力について，二つのシミュレーション研究により比較検討している．その結果，特に採点者間相関・採点者内相関，および項目数と採点者数は，カテゴリ数や採点方法など他の諸要因に比べて，採点結果の信頼性や選抜の正判別率に対して遥かに高い影響力を持つことが示されている． また，合格率に関しては選抜の正判別率に強く影響することも示されている．

また，紙幅の都合から本節では十分に扱えなかった話題もある．その一つに，採点者の負担を減らし，特に信頼性やバイアスの問題を解決する目的で開発が進められてきた自動採点システムがある．自動採点システムは，まだ日本での導入例は見られないが，特に大規模試験を中心として，今後の応用が期待される．自動採点システムについての詳細は石岡(2004,2006,2012)が参考になる．

2.2. 妥当性の問題

妥当性の問題は，当該の論述式テストが測定している能力・適性といった構成概念の実体は何か，ということである．仮に測定の信頼性が十分であっても，意図していた能力・適性から著しく乖離した構成概念を実は測定していたとすれば，そのようなテストの妥当性は低いと言わざるを得ない．また，論述式テストでは，「表現力」・「構成力」・「独創性」のように構成概念そのものが極めて抽象的である場合が多いため，客観式テストの場合以上に妥当性の検証が難しいように思われる．そのため，テスト項目の作成・開発段階から，測定を意図する構成概念を予め明確化しておくことが重要であり，またそのことがテストの仕様の決定や，妥当性を評価する際の判断基準を与えると言える．

しかし，そもそも妥当性をどのように定義づけるのかに関しては現在でも議論の尽きないところであり，妥当性の概念は時代とともに変化を遂げている．例えば村山(2012)に詳しく説明されているように，妥当性は「基準連関妥当性*9」・「構成概念妥当性」・「内容的妥当性」の三本の柱（三位一体観）から説明されることが多かったが，現在では構成概念妥当性という単一の統合的概念で捉えられるものという見方が主流になってきている．

この点からも推察されるように，信頼性の場合に比べて，妥当性の議論はずっと抽象的である．また，以下に述べるように妥当性の意味合いは非常に多面的であることから，妥当性のどの側面が重要になるのかは個々のテストに応じて異なる．そのため，妥当性に関する問題や対処法について一義的に述べることは，前節の信頼性に比べてずっと難しいという点は否めないだろう．しかし，妥当性の諸側面について知ることは，テスト実施後の妥当性の事後的な検証については勿論のこと，テスト項目の作成・開発段階も含めたテストの運用を見直す上で極めて重要である．

本節では，このような考えに沿って，論述式テストに独自に含まれる問題点を加味しながら，構成概念妥当性の一部として位置づけられる各々の妥当性の諸側面について述べていく．また，以下では，説明の便宜上タイプ分けに基づいた妥当性の説明を行うが，以下で述べるような妥当性のタイプ分けは唯一のものではなく，そしてこれらだけで妥当性の検証が全て終わる訳ではない点には注意が必要である．

(1) 内容的妥当性

内容的妥当性とは，「テストに用いられる課題内容が，それを用いて結論しようとしている測定内容のいかによい見本となっているかを示す概念」(池田, 1973, p178）である．例えば客観式テストを通して小学生の計算能力を測るということであれば，そのための様々な教科目標（例えば，二・三桁程度の整数の加減乗除の計算ができること，また分数・小数の計算ができること，百分率を理解していること，四捨五入などの数的処理が適切にできること，一定の時間内でこれらの計算ができること，現実の問題に対して計算式を立てて応用できること）を立てて，それに関わる項目群を網羅的に洗い出し，実際のテストにおいてこの項目群から項目の不備や偏りがないように出題ができているかという問題になる．

論述式テストの場合，内容的妥当性を高める際に運用上二つの大きな問題点がある．一点目は，「項目の不備や偏り」を考えること自体が難しいという点が挙げられる．この理由には，上述のように，論述式テストでは測定しようとする能力・適性が客観式テストに比べて抽象的である場合が多いことが挙げられる．他にも，関連して論述式テストでは，一見同じテーマを扱う課題内容でも回答形式や回答手順の設定(詳細は3.1.節を参照)について多様な選択肢がありえるため，その設定によってはテスト作成者の知らないうちに測定する能力の不備や偏りを生んでいる可能性がある．

二点目は項目数の問題である．一般に論述式テストでは，主に時間的制約の都合から，出題可能な項目数に強い制限が課せられる場合が多い．そのため，大学の小論文試験においてしばしばみられるように，項目数が一題のみである場合も珍しくなく，これは内容的妥当性の観点からすれば大きな問題である．項目数が不十分であれば，採点結果が当該の課題内容や回答形式に非常に強く影響を受けてしまい，採点結果の信頼性や後述するバイアスの観点からしても問題になる可能性が高い．

また，テストの直感的・印象的な意味としての適切性である表面的妥当性(池田, 1973, p.182)も，内容的妥当性の観点からよく指摘される問題である．例えば，実際はより限定的な能力の測定を行っているにも関わらず，見かけ倒れの標題を付けた能力やテストが安易に提唱・発表されることは，表面的妥当性の問題として考えられる．表面的妥当性の問題は，客観式テストの文脈で池田(1973, p.182)より古くから指摘されているが，論述式テストの場合は，扱われる構成概念が複雑なことが多いためにより深刻であるように思われる．このように，内容的妥当性は，論述式テストの場合においてはより問題となり易いため，項目数を可能な限り増やすことや，項目作成段階において作成者間の事前協議を綿密に経るなどの多くの工夫と努力を払うべき問題である．

(2) 基準連関妥当性

基準連関妥当性は，「問題としている行動特性と関連のある外部変数或いは基準測度とテスト得点とを比較することによって判定される」妥当性である（池田,

1973 p.183）.

論述式テストの文脈では，小論文試験の得点とセンター試験科目別の得点を比較する例が挙げられる(例えば，荒井・大久保・石岡・宮埜, 2012; 石岡・荒井・大久保, 2012). 具体的に，石岡他(2012)の事例では，小論文試験と英語の得点の間に中程度の相関がみられた一方で，数学や理科の得点とはほぼ無相関であった. このような作業は，小論文試験で測定している能力の実体を明確化するだけなく，今後のテスト作成の際にも有用であろう.

池田(1973, p.183)も述べているように，基準連関妥当性は操作的に定式化しやすく，経験的な検証が比較的容易である. しかし，一般に論述式テストではその運用そのものに大きな時間的・経済的コストが生じやすいこともあり，データの収集は容易ではない. そのため，一回のテスト運用を通して，基準連関妥当性も含め様々な測定論的問題に関する検証が可能な限り同時に実行できるような採点デザインやデータ収集上の工夫を考えることが望ましいだろう. データ収集上の工夫としては，例えばアンケートを用いて，外的基準の情報やテストの運用の改善に活かすための情報収集を行うことは有効である.

⑶ 因子的妥当性

因子的妥当性は，各項目に対する採点データに対して因子分析法を適用することにより，テストで実際に行われている評価についての質的構造を明らかにしていく作業である. 例えば，ある客観式テストが，計算能力を測定するために開発されたとすれば，各項目の出来不出来は計算能力の個人差を反映するはずである. このような場合，正の項目間相関，および因子分析モデルにおける一因子性が期待される.

ただしこのような例とは異なり，実際の論述式テストの場合，測定を意図する能力は単一でないことも多い. 例えば，文章力を最も高次の能力と位置づけ，低次に論理構成力・表現力・語彙力を想定して，低次の能力を反映するような複数の下位項目群を設定する場合である. このとき，各下位項目群間の相関が一定程度あるとともに，各下位項目群内ではより高い相関が見られることが望ましい. つまり，全体と部分それぞれの一因子性の見方が適切と判断されることが望まれる. この点に関する検証の仕方には幾つかの考え方があるが，各下位項目群内で個別に因子分析を適用する場合や，他にも階層的因子分析モデルに基づく方法などがある（例えば，心理検査場面の事例として，宇佐美・名越他，2011がある）.

以上のように，(a)「項目間の相関関係」は因子的妥当性を評価する際の主要な側面となるが，この点は客観式テストの場合も同様である. 加えて，論述式テストにおいて因子的妥当性を検証する際には他にも考慮すべき点がある. すなわち，分析的評価を用いた場合の(b)「複数の評価観点間の相関関係」や，他にも，(c)「採点者間の評価構造の等質性」という側面が含まれる.

(b)「複数の評価観点間の相関関係」については，個々の項目内での評価構造を明らかにするために検証されるものであり，(a)「項目間の相関関係」と類似した目的を持つものである. 例えば，宇佐美(2011)は，303名の高校生に対し，英語の早期教育の是非に関して問題文のみを与えた小論文課題，および日本の親の子育ての態度に関する３つのデータをみて，要約と意見を求める小論文課題を実施した. そして，実験的な目的から，「語句」，「表現の正確さ」・「語彙力」・「課題内容の解釈」・「簡潔性」・「主張の明確性」・「構成」・「一貫性」・「説得力」・「独創性」・「要約」・「形式」の計12種類の評価観点に基づいて，専門の採点者2名と非専門の採点者2名の計4名が採点を行っている. そして各課題別の採点データに対して因子分析を適用した結果，「語彙力」や「語句」・「表現の正確さ」が言語能力と解釈される因子に，また「構成」・「説得力」・「主張の明確性」など残りの評価観点が文章能力と解釈される因子にそれぞれ高く負荷を示すような，二因子に基づく解釈の適切性が示唆されている. さらに，渡部他(1988)や平(1995)の場合においても，概ね類似した意味を持つ二因子から説明される評価構造が抽出されている.

2.1(4)節で述べたように，単一の項目に関して，一つの総合的な評価観点から単一の採点結果を与えるのか，或いは構成力・表現力・語彙力など，予め設定された複数個の評価観点に基づいて複数の採点結果を与えるのか，といった評価法の選択に関する問題は複雑である (e.g., Barkaoui, 2007). この点については，信頼性の観点から分析的評価法の有効性が示されている場合もある一方で（e.g., Breland, 1983; Huot, 1993），二つの評価法の間で系統的な差異があるとは判断しにくい事例（例えば，渡部他,1988 ; 宇佐美, 2011）も多く，むしろ総合評価に比して時間的コストが膨大にかかる割に信頼性の向上に寄与しないという根拠から分析的評価に関して否定的な意見もある(Wiseman, 1949). ただし，総合評価の場合，受験者にとって評価がどのような観点を通して行われているのかを知ることは容易ではない. その一方で，分析的評価の場合は評価観点別に採点結果をフィードバックできるため，テストのアカウンタビリティからすれば分析的評価は望ましいとする指摘もある（宇佐美, 2009a）. また，平・江上(1992)でも指摘されているように，分析的評価法により，上で示したような，因子分析に基づく評価構造の検証を行うことができるという利点は無視できないであろう.

ところが，さらに根本的な問題として，分析的評価

法による採点結果と，総合評価法による採点結果は異なる構成概念を反映している可能性があるという指摘があり（渡部他, 1988），実際にこれらの2つの採点結果の間に十分に高い相関が見られなかったことを報告している研究（梶井, 2001,2002; 渡部他, 1988; 吉川他, 2006）もある．この点に関する実証的な研究はまだ十分には蓄積されていないと思われるため，今後の測定論的問題に関わる研究課題の中でも特に重要と言えるであろう．

最後の(c)「採点者間の評価構造の等質性」の検証も実際上重要な問題となる．採点結果の信頼性を高めるためには採点者数を増やすことが望ましいが，採点者間で評価構造が十分に等質であるかについての明確な根拠が得られない限り，複数の採点者に基づく採点結果を加点・平均することの意味は希薄になる．もし評価構造が等質でないと考えられる場合，特定の採点者による評価には偏りがあると考えられるため，因子的妥当性だけでなく評価のバイアスの観点からも問題になりうる．評価構造の等質性の検証の仕方も様々であるが，例えば各採点者が行った採点データに対して因子分析法を適用し，採点者間で概ね類似した因子パタンを示していることを確認する手続きが最も簡便であると言えよう．

(4) 交差妥当性

交差妥当性の問題とは，当該の受験者集団で示された妥当性の証拠が，他の受験者集団の場合でも同様に成立するかということである．特に論述式テストを研究場面で用いる場合，運用上のコストの都合から，多くの集団に対して一度に試験を実施しデータを収集するのが困難であることが多い．その結果，例えば特定の地域・学校・年齢の集団からデータが収集される場合がほとんどである．そのため，可能な限り追加データの収集をして，交差妥当性の検証を行うための工夫をすることが望ましい．

また，交差妥当性には，当該のデータを利用して作成された基準の適切性を再評価する，という側面も含まれる．例えば，あるテスト得点を通して他のテストの合否や得点を予測する場合，または同様にテスト得点を利用して個人を幾つかの少数のクラスに分類（ランク分け：3.3節を参照）する場合が関係する．このとき，推定されたこれらの予測式や分類のための基準値が当該の受験者集団のデータに過度に依存している場合があるため，他の受験者集団においてもそれらの適切性が示されるとは限らない可能性がある．そのため，追加データを収集して，他の受験者集団でも予測式や基準値が適切に機能しているかを検証することが重要である．

(5) 結果妥当性

結果妥当性，或いは関連して本質的妥当性と呼ばれるものは，テストを実施した結果として受験者や関係する事物全般に生じた変化についての社会的・教育的な意味としての適切性を表す概念である．そのため，これまで述べてきた，当該テスト固有の統計的結果から検証される妥当性とは性質の異なるものである．

例えば，宇佐美(2009a)にもあるように，テストを受けた経験そのものが受験者の学習観・動機づけ・学習方略に影響を与えることが指摘されている．例えば村山(2003)は，記述式テストと空所補充型テストを用いて，学習方略やノート書き込み量などの行動指標に与える影響を検討し，記述式テストを課された群では深い処理の学習方略使用が促進されたことを指摘している．また鈴木(2011)は，ルーブリックの提示による評価基準や評価目的の教示が，「テストの実施目的・役割に対する学習者の認識」としてのテスト観に与える影響を示している．

このように，ある論述式テストを実施することが，そのテスト形式や評価基準，評価方法，またはフィードバックを介して，受験者の価値観や行動全般にどのような変化を齎すのかは，テストのアカウンタビリティにも繋がる重要な問いであろう．

(6) 採点と妥当性の関係

論述式テストの場合，構成概念を如何に明確に定義した上で項目を入念に作成したとしても，テストデータが個々の採点者による採点結果から得られる以上，採点作業が適切に行われない限りは無意味になってしまう．例えば，個々の採点者が半ば無意識的に持つ評価観点の種類や重みづけ，また採点方法の違いは，採点者間(内)相関などを通して信頼性に影響を与えるだけでなく，因子的妥当性を中心とした妥当性(および後述するバイアス)を損なう根本的な原因となり得る．このような採点の妥当性とも言うべき採点妥当性を高めるためには，3.1節で述べるように作成される項目内容にも依存すると考えられるが，事前の事前協議や採点者訓練を綿密に行うことがやはり重要である．具体的には，どのような評価観点を通して各項目の採点を行うのかについて採点者間で共有し，採点者訓練を経てそれが実行できるようになるまでの作業を入念に行うことが必要である．しかし，実際のテスト作成場面では，時間的コストの理由や，他にも「各々の採点者が個別の評価観を持っていれば全体の採点結果には反映されるため，評価観点は採点者間で異なっていても良い」という，少なくとも測定論的に見れば誤った認識をもつテスト作成者もいることから，この点は必ずしも実現されているとは言えないのが現状であろう．採点者訓練や事前協議については3.4節で改めて触れる

ことにする.

2.3. バイアスの問題

採点のバイアスの問題も，公正なテストの運用を阻む危険因子である．一般に種々のバイアス要因は，採点者の熟練度，課題内容，採点者訓練の内容，受験者の特性，採点者の疲労などによる複雑な相互作用の結果生じるものが多い．さらに，当該の採点時に生じる場合もあれば生じない場合もあるなどの複雑な性質を持っている.

バイアスとして広く知られている要因としては，文字の美醜効果 (e.g., Chase,1979 ; Eames & Loewenthal, 1990)や，評定する順序によって同じ文章の採点結果が変動する問題である系列効果(e.g., Hughes, Keeling, & Tuck, 1980, 1983b)，さらには時間制限，採点日数，受験者の人種・性別・性格・魅力，教師の期待，採点日，採点者の評定の厳しさなどがある.

このように，バイアスとなりうる要因は極めて多岐に渡り，各々のバイアスの影響力やその発生条件については十分に検討されていないものが多い．ただし，上述のように，各バイアスは項目や採点者・受験者に関する様々な要因についての複雑な相互作用の結果生じている場合が多いので，項目作成や採点時の工夫，および採点者訓練の徹底を図ることが重要である（採点者訓練については3.4節で後述する）．バイアスに関するより詳しいレビューは宇佐美(2008,2009a)を参照されたいが，以下ではその重要度の高さから特に考慮すべき，系列効果と文字の美醜のバイアス要因について述べる.

まず，系列効果については多くの研究（Hale, 1975; Hughes, Keeling, & Tuck, 1980, 1983b, : Hughes & Keeling, 1984）で見出されており，その影響を取り除くのは難しいとされている．さらに，前・後いずれの順番で採点される方が評価は不当に高くなるのかについては，結果は一貫していないようである．系列効果バイアスに対処する一つの工夫としては，3.2節でも後述するように，採点者間で採点する答案の順番をランダム化することや，複数回の採点を行うことが望ましい.

また，文字の美醜バイアスはJames(1927)にもあるように，とりわけ古くから検証されてきたものである．日本でも文字の美醜バイアスに焦点を当てた研究はあるが（例えば渡部・曹, 1992；宇佐美, 2008），その数は十分には多くない．例えば，Hughes, Keeling, & Tuck, (1983a)で報告されている実験においては，単一の作文の評価において文字の美しい文章はそうでない文章に比べ，採点の結果に25 点満点中平均1.32 点(100点満点であると約5 点)分の差が生じたことが示されている．一方，採点者である教師は文字の美しさに左右されないことを示した研究（Eames & Loewenthal,1990）もある．一つの対処法として，文字の美しさの影響を排するためには，パソコン打ちした答案によって採点を行うことは有効であると考えられる．しかし，それによって採点結果が全体的に甘くなるといった新たなバイアスが生じる可能性があるが，パソコン打ちの答案による採点結果についてはまだ十分に検証されていないようである(吉村, 1991, 1992, 1993).

3 論述式テストの各運用段階に関わる測定論的問題の明確化と実践上の工夫について

これまで述べてきたように，論述式テストの孕む測定論的問題は，信頼性・妥当性・バイアスのいずれの観点から見ても非常に多面的で複雑である．そのため，単に測定論的問題を概観するだけでは，実際の論述式テストの運用に資するような実践的な示唆を得ることは容易ではない．そこで，本節では，テストの各運用段階に関連する測定論的問題を明確化しながら，実践上必要な工夫について整理していく.

表3は，テストの項目作成・採点・フィードバック等の各運用段階における検討事項と，それらに対応する測定論的問題およびテストの公平性等に関わる諸側面をまとめたものである．表3は，当初肥田野(1972)を参考に作成した宇佐美(2009a)の内容を，前節までの内容も踏まえ大幅に改編したものである．表では，各検討事項に直接関わると考えられる測定論的問題に○を，また特に関連や影響度の強いと考えられる測定論的問題に◎を付している.

以下では，表3の内容を段階的に見ていきながら，テストの運用上重要な観点であるものの前節までには十分に触れられなかった話題を中心に補足していく.

3.1. 項目作成(段階 1-1)

テストの品質に最も直結すると言っても過言でないのが項目作成段階である．項目作成段階では，どのような能力・適性を，どのような目的（例えば，選抜・分類・予測…）で，またどのような課題内容や形式を通して測定していくかを，項目数や課題の難易度，他にも回答時間・制限字数を含め多面的に吟味していく．この段階から，測定を意図する能力・適性といった構成概念を予め明確化しておくことが，後にテストの仕様や，妥当性を評価する上での判断基準を与えることに繋がる.

また，課題内容については，内容的妥当性や因子的妥当性などの妥当性の諸側面はもちろんのこと，信頼性にも大きく影響しうるため重要度が高い検討事項である．宇佐美(2008,2009a)では課題内容と信頼性およ

表3: テストの各運用段階における検討事項と，対応する測定論的問題(1/3)

	測定論的問題												
	信頼性				妥当性					バイアス	経済的・時間的・人的コスト	フィードバック	公平性などその他
	項目数・採点者数	項目間相関	採点者間(内)相関	採点方法・その他	内容的妥当性	基準連関妥当性	因子的妥当性	交差妥当性	結果妥当性				
1 試験実施前の諸決定													
1-1 項目作成													
(1)論述式テストの実施目的は，選抜か，分類（診断）か，予測か，または受験者の動機づけを高めるなど他の目的も兼ねたものか.												○	○
(2)どのような能力（構成概念）の測定を意図した論述式テストにするのか.		○	○		◎		◎						
(3)上記の能力を測定するための課題内容（テーマ）は何にするか.		○	○		◎		◎						
(4)項目数の設定をどうするか.	◎				◎						○		
(5)課題形式（課題型・素材型・データ型）はいずれにするか.		○	○										
(6)回答手順（論述の仕方やその内容に関する指定）はどうするか.		○	○										
(7)回答時間の設定はどうするか.	○			○									
(8)制限字数の設定はどうするか.	○	○	○	○			○				○		
(9)各項目の難易度は予想される受験者集団の能力からして適切か.													○
(10)出題する項目は，測定を意図する能力を反映する項目領域全体から偏りなく選ばれているか.	○				◎								
1-2 採点デザイン													
(1)採点者数・採点回数の設定をどうするか．また，各採点者が採点する項目の分担等に関する採点デザインはどうするか.	◎										○		
(2)評価法は総合評価にするか分析的評価にするか．また，分析的評価の場合，どのような評価観点を設定し，そしてどのようにフィードバックに活用するか.				○			○				○	○	
(3)採点カテゴリ数(3 カテゴリ，5 カテゴリ等）はどのようにするか．また，各カテゴリ内で設けられた採点基準は適切か.				○									
(4)各採点者が複数回の採点を行い，採点結果が一致しなかった場合，採点結果の補正を行うか．さらには，それはどのような補正方法にするか.				○									
(5)採点者間の採点結果が一致しなかった場合，採点結果の補正を行うか．さらには，それはどのような補正方法にするか.				○									
(6)各項目の各カテゴリ得点に対応する，答案例（ベンチマーク）を作成するか.			○							○			
(7)系列効果が生じないよう採点の順番をランダムにする等の工夫をするか.										○			
(8)パソコン打ちして電子媒体で保存した答案をもとに採点するなど，文字の美醜を含め採点のバイアスが生じにくくなるような工夫を行うか.										○	○		

表3: テストの各運用段階における検討事項と，対応する測定論的問題(2/3)

	測定論的問題												
	信頼性				妥当性					バイアス	経済的・時間的・人的コスト	フィードバック	公平性などその他
	項目数・採点者数	項目間相関	採点者間(内)相関	採点方法・その他	内容的妥当性	基準連関妥当性	因子的妥当性	交差妥当性	結果妥当性				
1-3 フィードバックの方法													
(1)得点の開示は素点に基づくものにするか．項目反応理論に基づく尺度化は行うか．またさらに，何らかのランク分けに基づく段階評価を行うか.												○	
(2)項目内容およびフィードバックの内容・方法は，教育的・社会的な観点から見た際に適切であると考えられるか.									○			○	
1-4 事前協議・採点者訓練													
(1)各採点者は各項目および各カテゴリ内の採点基準や採点方法を十分理解しているか．そして，その理解の共有化を図るための事前協議を行うか.			◎				○			○	○		
(2)ベンチマークを用いた採点の練習など，採点者訓練を行うか.			◎				○			○	○		
1-5 その他の運用上の検討事項													
(1)想定される受験者集団（規準集団:norm)はどのように定義されるか.												○	○
(2)アンケート等を用いて，外的基準の情報やテストの運用の改善に活かすための何らかの情報の収集を行うか.						○						○	○
(3)テストの実施やフィードバックの時期，場所，方法等の諸決定はどうするか.									○			○	○
(4)障害をもつ受験者を考慮した運用等，テスト固有の性質とは別の観点から考えた際のテストの公平性について問題はないか.										○			○
2　採点段階													
(1)（一部の）答案を採点した結果として，設定した評価基準が適切であると再確認できるか.			○				○			○			
(2)採点作業の途中に，ベンチマークの答案や異なる採点者間・採点者内の採点結果を用いて，測定の安定性についての検証を行っているか.			○				○			○	○		

び妥当性の関係について論じている．各々の論述式テストは，固有の課題内容や回答形式，また回答時間や制限字数・教示文が設定されていることに伴って，解答と考えられる内容の質や幅の広さは，テスト間で変動することが予想される．この，解答として想定される内容や幅の広さを，宇佐美(2008)は解答内容の自由度と呼んでいる．また，解答内容の自由度の構成要素として，課題内容の具体性・客観性などの(a)課題の内容的要素と，回答形式・回答手順・制限字数・回答時間に関する設定を中心とした(b)回答方法の要素，の二つに分けて論じている．

(a)については，例えばグラフの結果を読み取って論述するデータ型の形式の項目への解答は，それが客観的に示されている図の内容を基にしている為に解答内容の自由度が小さくなると考えられる．その結果，測定する構成概念の領域は狭くなる一方で，他の条件が同じであれば測定の信頼性は高くなると予想される．また，しばしば歴史や経済分野の論述式テストに見られる，歴史的な事実や知識的要素を含んだ内容を扱う場合にも，その課題内容の客観性から，解答内容の自

表3: テストの各運用段階における検討事項と，対応する測定論的問題(3/3)

	測定論的問題												
	信頼性				妥当性					バイアス	経済的・時間的・人的コスト	フィードバック	公平性などその他
	項目数・採点者数	項目間相関	採点者間(内)相関	採点方法・その他	内容的妥当性	基準連関妥当性	因子的妥当性	交差妥当性	結果妥当性				
3 測定論的問題の検討													
(1)複数の採点者が採点した場合，採点者間相関はどうか．また複数回採点を行った場合，採点者内相関はどうか．			○										
(2)複数の採点者が採点した場合，採点者間で評定の平均値や，評価構造（相関）に大きな差異はあるか．							○			○			
(3)項目間相関から見た，テストの内的整合性はどうか．		○											
(4)一般化可能性理論を適用して各分散成分を推定し，また異なる項目数や採点者数等の下で一般化可能性係数を推定するとどうか．	○												○
(5)テスト得点と外的基準との相関関係はどうか．						○							
(6)項目間相関の情報から評価構造を調べるとどうか．							○						
(7)複数の評価法(総合評価・分析的評価)を用いた場合，それらの間の採点結果の相関関係はどうか．また(1)-(6)の一連の結果を比較するとどうか．				○			○					○	
(8)必要な場合，採点結果から尺度化やランク分けをするとどうなるか．												○	
(9)上の一連の検証結果は，他の受験者集団においても同様に成立しているか．								○					
(10)特定のサンプルについての解析結果の比較などを通して，採点結果のバイアスの有無を検証するとどうか．										○			

由度は狭まりやすいであろう．

(b)の回答方法の要素としては，例えば「遺伝子組み換え作物を食品として取り入れることに反対か賛成か．」という問いから，「遺伝子組み換え作物を食品として取り入れることに反対か賛成かを，自分の支持する意見に向けられる反論を考慮して論述せよ」のように回答手順を設定すれば，論述の構成や内容を部分的に指定することになる為に，解答内容の自由度は相対的に小さくなることが予想される．Christian, Timothy, Richard & Bud (2002)は回答方法の設定の仕方とテストの測定内容との関係について詳細に論じている．他にも，阿久津・菊池他(2006)は，課題内容や回答方法の設定の違いと採点者間相関との関係について述べている． このように解答内容の自由度は，信頼性や妥当性を中心に測定論的問題全般に直結すると言える．

また，解答内容の自由度を広げるほど様々な能力の測定可能性が高くなるという魅力に一見惑わされやすいだろう．しかし，反面，事前に設定された採点基準からでは判断の難しい答案が増えることや，本来の意図とは逸脱した回答を示す答案が増えるなど，間接的にテストの信頼性や妥当性に悪影響を及ぼす危険性が増すと考えられる．そのため，テストの作成時には不必要に解答内容の自由度が高い課題内容や回答方法になっていないか吟味する必要があるだろう．

さらに，測定論的に見て重要な影響力を持ちながらも，項目作成段階ではやや軽視されている感のある検討事として，上記(b)の回答方法の要素で指摘した，制限字数と回答時間が挙げられる．例えば，宇佐美(2009a, 2011)は崎濱(2005)の研究から得られた示唆をもとに，制限字数の設定の違いが採点結果の信頼性・構成概念妥当性・バイアスに与える影響を実験的に検証している．その結果，基本的な回答手順を守りある程度の論理構造を有している回答になっているか否かを評価する程度であれば，短い制限字数に基づく論述式テストでも測定論的には十分妥当な評価を達成しうることを指摘している*10.

制限字数の要因は，項目数の設定上限や採点者の負担にも深く関わる問題であるため，測定を意図した能力の差異に応じた制限字数ならびに項目数の設定可能

性について，今後さらに実証的な検討を進めていく必要があるだろう．この点は回答時間についても同様に指摘できることであり，これらの見直しが測定論的に見た際の論述式テストの品質の改善に大きく資する可能性がある．

3.2. 採点デザイン(段階1-2)

採点デザインについての諸決定も，測定論的問題に関わる重要な検討事項である．例えば，採点者数・採点回数・各採点者で分担する項目の決定，評価方法（分析的評価・総合評価）・採点カテゴリ数・採点者間・採点者内で採点結果の不一致があった場合の補正方法の選択，ベンチマークの作成，採点時間の決定などが含まれる．評価方法・採点カテゴリ数・補正方法の選択については，2.1(5)節で述べたとおり，それらは項目数や採点者数および採点者間相関・採点者内相関に比べて信頼性への影響力は小さい．しかし，評価方法の選択に関しては，2.2(3)節でみたように，とりわけ分析的評価の場合，一般に採点作業に要する時間的コストの代償を伴うが，因子的妥当性の検証がしやすいだけでなく，フィードバックに評価結果を活用しやすいという利点がある．

また，ベンチマークの作成は高い採点者間相関・採点者内相関を維持し，特定の採点者からのバイアスを抑制する上で重要である．さらに上記以外の検討事項としては，2.3節で述べたような系列効果が生じないための採点の順番の検討が挙げられる．他にも，時間的コストを伴うが，文字の美醜効果を除くためにパソコン打ちした答案を利用する等の工夫も一考すべき点であろう．

3.3. フィードバックの方法(段階 1-3)

どのような方法を通して結果のフィードバックを行うかは，2.2(5)節で見た結果妥当性にも関わる重要な問題である．特に客観式テストの文脈においては，項目プールを作成し当該受験者集団や項目に依存しない評価を実現する目的や，また測定の信頼性を能力水準別に検証する目的で，さらにはCAT(Computer Adaptive Test) の枠組でテストを運用する為に，項目反応理論（Item Response Theory；IRT）の利用が進んでいる（宇佐美, 2010）. Lawley(1943)やLord(1952), またLord & Novick (1968)がIRT の理論的基礎を築き上げてから今日に至るまで，項目反応を表現するモデルは多数提案されている. それらに関しては, Baker & Kim (2004), 村木(2011), 豊田(2005), van der Linden & Hambleton (1997)などが参考になる.

特に大規模な試験の文脈では，論述式テストに基づく採点データも，IRT の枠組みを用いた解析が有用であるケースは多い．例えば，論述式テスト項目が客観式テスト項目と共に出題される場合や，他にも分析的評価を複数の採点者が行う採点デザインでテストが実施された場合である．論述式テストデータを扱うための項目反応モデルには多相ラッシュモデルや宇佐美(2010)の多値型項目反応モデルなど様々なものが提案されている．一方，総合評価による論述式テストデータを行う場合一般に見かけ上の項目数が不十分であることや，2.2(3)節で述べたとおり総合評価と分析的評価では測定している内容が異なる可能性があるため，この点IRTを念頭においた運用を行う際には注意する必要があるだろう．また，等化などIRT特有の作業に関しては，論述式テストデータを扱った場合の実施方法の検討や経験的証拠の蓄積が今後必要であろう．

他のフィードバック方法としては，例えば採点結果について素点（合計得点）だけを受験者にフィードバックするのではなく，例えば能力判定A,B,C,D,Eなど，何らかの手続きで離散化し作成されたランクを併用する場合がある．このようなランク分けによって学習の達成水準の把握を直感的に容易にすることや，また受験者に対してフィードバックを合理的に行うことができるという利点がある．また他にも重要な点として，素点そのものの信頼性が高くない場合は，一定範囲内にある得点は同一の能力水準としてみなしてランク分けをした方が，受験者に対して不適当な評価やフィードバックを与える可能性を低くできるということが挙げられる．

ランク分けの手続きとしては，素点，或いは項目反応理論を用いて推定された潜在特性値の記述統計や分布をみて，得点範囲とランクとの対応関係を事後的にまとめる手続きがある．他にも，より統計的に洗練された手続きとして，ノンパラメトリックな有限混合モデルとして考えることのできる，荘島(2007)やShojima(2008)による潜在ランク理論(Latent Rank Theory; LRT*[11])を用いる方法も近年多く報告されている．

ランク分けは測定の信頼性を考慮する上で有効な方法である一方で，ランク分けをすること自体で生じうる様々な影響については無視できない問題である．例えば，離散化された各ランク自体は，水泳の場合のように，「息継ぎができるレベル」，「10kmの遠泳ができるレベル」といった質的で明確な内容を通常は意味しないため，各ランクにどのような意味を付与し，それを受験者にいかにフィードバックするかはテストのアカウンタビリティにも関わる問題である（宇佐美, 2009b）．この点に関しては，仮に分析的評価法を用いている場合や，或いは客観式テスト項目も含め複数個の項目が設定されている場合では有効な手続きがある．すなわち，このような場合，各ランクにおける各項目の評価点の平均やランク間のその推移をみて，各ランクにおける学力の達成水準と学習目標の対応を

can-do-chart(植野・荘島,2010, pp.106-108) にしてまとめることが有用であろう．

3.4. 事前協議および採点者訓練(段階 1-4)

一般に採点者訓練および事前協議は，テスト運用上の時間的コストのため必ずしも十分になされないことが多い．しかし，2.2(6)節で述べたように因子的妥当性を中心としたテストの妥当性に，また同時に採点者間相関や採点者内相関を介して信頼性に，さらには採点結果のバイアスにも強く影響を与える点でもある（石井, 1981; Penny, et al, 2000; 宇佐美, 2012）．事前協議では，各項目に対する評価観点および評価方法等について決定し，採点者間で共有化することが必須であろう．特に評価観点については，採点者によって質的に異なる評価基準を有している可能性が高いが，事前協議の段階で綿密に議論をして共有化することで，各採点者内の評価基準がより具体的になり精緻化されるという利点がある．

また，採点の訓練を経た者といっても，当該の論述式テストの運用のために短期的な採点者訓練を受けた者を意味する場合もあれば，長期的に論述式テストの作成・採点に携わっている専門家・熟練者を意味する場合もある．現実的には専門の採点者を十分に確保することは容易ではないので，しばしば前者程度の訓練でも十分に採点結果の信頼性が高くなることが重要である場合も多い．

採点者訓練の効果を検証した研究はあまり多くないが，特に信頼性の観点からは，Knoch, Read & Randow (2007)やLumley & McNamara (1995)は採点の練習を積み一貫した評価ができるよう訓練された採点者らによる評価は，その信頼性が優れていることが示されている．また，鷲坂・二村他(2001)の研究では，採点基準の明確化の工夫のほかに，二時間程度の採点者訓練を行うことを通して，採点経験のない非専門家による採点も，専門家による採点と概ね遜色ない信頼性が確保された事例が報告されている．

3.5. その他の運用上の検討事項(段階 1-5)

これまで見てきたもの以外にも，測定論的問題に関わる運用上の検討事項がある．例えば，想定される受験者集団の学歴・年齢層・男女比・出生地区分等の情報から受験者集団の平均像を明確化することが挙げられる．この点は，テストの規準集団像を明確にし，各テスト得点が絶対的にどのような意味を持つかについての情報にも関連してくるため，フィードバックの内容や方法の検討の際にも必要である．

また，外的基準についての情報やテストの運用の改善に活かすための情報を得る目的で，アンケート等を用いて受験者から何らかの情報収集を行うか否かについても，重要な検討事項である．これは2.2節で述べた基準連関妥当性にも関わる点である．他にも，テストの実施やフィードバックに関する時期，場所，方法等についての諸決定も，フィードバックの効果を介して結果妥当性に関わる問題である．さらに，大規模試験ではとりわけ広く認識されているように，障害を有する受験者を想定したテスト運用についても，採点結果のバイアスを始めとした，テストの公平性に直結する重要な検討事項である．

3.6. 採点段階(段階 2)

これまで見てきたように，採点に関わる諸決定は，採点デザインや採点者訓練など，テストの実施前に関わるものが多い．その一方で，採点段階でも適宜判断していくべき，測定論的問題に関わる検討事項がある．

例えば，一部の答案を採点した際に，様々な理由によって想定外の内容を含む答案が現れ，そのため暫定の評価基準では評価点を根拠立てて決定できない状況が生じうる．そしてこの点が，採点結果の信頼性・妥当性・バイアスの問題に直結する可能性が考えられる．そのため，テストの実施前に設定した評価基準が適切であると再確認できるかどうか採点者の間で検討し，必要に応じて修正していく作業は重要である．

また，採点者間で採点結果を比較する作業や，他にもベンチマークを用いて採点結果の偏りの有無を調べる中間作業も，一定の時間的コストを伴うが，特に採点者間相関および採点の妥当性を維持し，さらにはバイアスを抑制する目的からは必要不可欠である．

3.7. 測定論的問題の検討(段階 3)

テストの実施後に行われる測定論的問題の検討においては，単に考慮すべき検討事項が多く複雑であるだけでなく，それが統計的な手続きに直結しているものがほとんどであるため，それらの検討の実施は全てのテスト開発者にとって容易なことではない．この場合，教育測定・評価に関わる専門的な知識・経験を有する専門家との連携が不可欠である．しかし，この点からすると，木村(2010)で指摘されているようなテストの専門家の不足の問題がいかに深刻であるかは明らかであろう．

本論文では，紙幅の都合から，測定論的問題の検討に関わる技術的な内容についての多くを割愛せざるを得なかったが，Blok (1985)，平井・椎名・柳井(2001)，梶井(2001, 2002)，平(1995)，宇佐美(2009a,b, 2010, 2011)，渡部他(1988)などで報告されている分析事例は参考になるであろう．

4 総合考察

本論文では，現行の論述式テストの運用の改善に資することを目的として，まず論述式テストにおける測定論的問題の内容を整理した．そして，論述式テストの各運用段階における検討事項と対比させながら，実践上配慮・工夫を要する点についての整理と議論を行った．

テストは，性格・適性・能力・興味などの個体差の記述・評価を通して，選抜や診断・分類等の社会的に重要な意志決定場面において中心的な役割を担う製品と言ってよい．ところが，道具としての製品の安全性やその品質の検証が行われることは極めて当然であっても，冒頭で述べたように，測定論的観点からテストの品質を維持・評価・改良することの必要性に関する認識は一般に薄いと思われる（宇佐美,2011）． 無論，極めて周到な手続きに基づいてテストを運用しているケースも確かにある一方で，安易な計画に基づいたテスト作成と評価を行ってしまう例も少なくないと思われる．さらには，教育測定そのものに関しても，それを人間を無機質な存在として捉え，そして評価を機械的・画一的に捉えるアプローチであるかのような誤解も，残念ながらあるように思われる．

吉村他(2008)にもあるように，テストの品質とその維持・評価・改良の重要性，および教育測定・評価に関わる技術・知識の啓蒙は容易ではない．本論文で行った測定論的問題のまとめや議論した対処法は，その有効性や実行可能性からしても多くの課題を残しているが，テスト開発関係者に対して測定論的問題についての認識の強化を図ることが，真に公正なテストの運用の実現の為には何よりもまず望まれることなのかも知れない．

そのためには，例えば，吉村他(2008)が述べているように，テストの専門家を交えた小論文や面接の実施マニュアルの作成は，テストの品質の評価・維持・改善に役立つ実践である．そして，このような実践に加えて論述式評価についての実証的研究の蓄積や，得られた知見を整理するために本論文のようなレビューや解説を行うことは，論述式テストの運用に伴う測定論的問題への早期認識および解決に寄与すると考えられる．そして最終的には，宇佐美(2009a,2011)にもあるように，このような試みを通して測定論的問題だけでなく，教育行政・大学経営などの多角的な視点からも，論述式テストの運用上の問題を（単に語るだけではなく）実際に調査して研究し，論述式テストを改善する試みが今後少しでも増えることを期待したい．

謝辞

本論文の執筆にあたり，二名の査読者の先生からは大変多くの建設的なコメントをいただきました．心より感謝申し上げます．

注釈

1) 主に選抜の目的で利用される入学試験や小論文試験は勿論のこと，資格試験や人事試験・適性試験，また性格検査や臨床検査に代表される様々な心理検査についても，テストという用語の範疇に含まれる．

2) 大野木(1994)は，小論文試験を，回答形式を基準に課題小論文・素材小論文・データ小論文に区別している．課題小論文は一定の長さの課題文が項目に先立って提示される形式であり，また素材小論文は課題文を伴わずに提示されたテーマに即して論述する形式で，一行問題とも呼ばれる．データ小論文は与えられた図や表・データの内容をもとに論述する形式である．

また，小論文試験以外にも作文という用語も良く使われる．作文は物語の創作や随筆などの文脈で用い，必ずしも論述的な要素を伴うものを意味しないが，本論文での議論は作文にもあてはまるものである．

3) テストという言葉については，その用途や実施形式，実施方法の違いに応じて様々な分類がなされており（例えば，池田, 1973），それと同時に多くの用語が存在する（大野木, 1994; 宇佐美, 2009a）．客観式に対応する言葉は， 論文体テスト（東・梅本他, 1988），論文式テスト(肥田野, 1972)，他にも論述式・記述式や，解答構築式（村木・斉田, 2008）など多様である(宇佐美,2009a)．平井(2008)にもあるように，実際のテストでは作文に近い形式である叙述型の項目も含まれうることから，論述式・論文式等よりは解答構築式などの用語を用いることが適切であるように思われる．しかし，慣習に従って，本論文では論述式という表現を用いる．

また，大野木(1994)の議論にもあるように， 論述式テストの分類方法も多岐にわたる．大野木(1994)は，記述量が多い順に論文形式・小論文形式・論述と分類している．あくまで目安であるが，本論文では小論文形式および論述，すなわち100-2000字程度の記述量を伴う文章についての測定・評価の問題を扱う．

4) 無論，論述式テストの利点と欠点としては，測定論的問題に関わるものだけでなく，「経済的・時間的・人的コストを多分に要するテスト形式であること」や「文章表現独自の面白さが学習者の興味を喚起させる可能性があること」，「受験者に表面的な知識だけを覚えれば良いとする，機械的暗記主義的方略の助長を抑制させる利点があること」など様々なものが指摘されている．宇佐美(2009a)は，東・梅本他(1988)，肥田野(1972)，池田(2007)，大野木(1994)などを中心に，論述式テストの利点と欠点を，歴史的経緯と共にまとめている．

5) バイアスの問題は広義には信頼性や妥当性を包含する問題と考えられるが，論述式評価の文脈においてはバイアスの発生源は項目内容そのものよりは採点者の特性に帰せられる場合が多いことや，海外の文章評価研究ではバイアス研究が１つの主軸となっていることからも，本論文ではバイアスの話題を独立に設けて議論する．

6)採点者間の採点結果の安定性・一貫性を評価する際は，採点結果の一致度や級内相関といった他の指標を利用することもある(例えば，平井,2008)．

7)項目間相関を高めるには，内的整合性の高い項目群を用意するだけでなく，受験者集団の能力の個人差が十分に大きいことも必要である．これは，(2)式において真値の分散が大きいほど ρ が高くなることにも表れている．

8) Cronbach & Shavelson (2004)は，二相の場合であっても，α 係数を計算する上での要件である，互いにランダムで平行な項目に基づくデータが得られることは現実には稀であるという理由から，一般化可能性係数を通した評価が望ましいと指摘している．

9) 基準関連妥当性と呼ばれることもある．

10) 無論，測定の目的によっては，制限字数を短くすることで別の様々な問題が生じることが予想される．より詳細な議論については宇佐美(2009a,2011)が参考になる．

11)荘島(2007)や宇佐美(2009b)等ではニューラルテスト理論(Neural Test Theory : NTT)と呼ばれるものである．

参考文献

阿久津洋巳・島野恵美子・熊谷賢・佐々木和歌 (2005). 課題レポート評価における評定者間の一致. 岩手大学教育学部付属教育実践総合センター研究紀要, *4*, 75-80.

阿久津洋巳・菊池梢・鈴木安澄・鈴木光・渡邉愛枝 (2006). 論述式テストの研究(1) －採点者間の一致度－.岩手大学教育学部付属教育実践総合センター研究紀要, *5*, 115-122.

荒井清佳・大久保智哉・石岡恒憲・宮埜壽夫 (2012). 複数課題の小論文試験得点と他教科科目得点との関連 日本テスト学会第10回大会発表論文抄録集,88-89.

東洋・梅本堯夫・芝祐順・梶田叡一編 (1988). 現代教育評価事典　金子書房

Baker, F.B., & Kim, S.H. (2004). Item response theory：Parameter estimation techniques (2nd ed). New York: Marcel Dekker.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. Assessing writing, *12*, 86-107.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, *22*, 41-52.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. Applied Psychological Measurement, *24*, 339-353.

Brennan, R. L. (2001). Generalizability Theory. Springer-Verlag.

Chase, C. I. (1979). The impact of achievement expectations and handwriting quality on scoring essay test. Journal of Educational Measurement, *16*, 39-42.

Christian, M. R., Timothy, W.B., Richard, R. S., & Bud, W. (2002). How to prepare effective essay questions. BYU Faculty Center, Brigham Young University Testing Services.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, *16*, 297-334.

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. The British Journal of Statistical Psychology, *16*, 137-163.

Cronbach, L.J., & Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, *64*, 391–418.

Eames, K., & Loewenthal, K. (1990). Effects of handwriting and examinee expertise on assessment of essays. The Journal of Social Psychology, *130*, 831-833.

Hayes, J. R., Hatch, J. A., & Silk, C. M., (2000). Does holistic assessment predict writing performance?: Estimating the consistency of student performance on holistically scored writing assignments. Written Communication, *17*, 3-26.

肥田野直 (1972). 心理学研究法7 テストⅠ 東京大学出版会.

平井洋子 (2002). 論述的課題による高次思考能力測定の試み －採点内容の検討－ 人文学報, *326*, 17-30.

平井洋子 (2006). パフォーマンス・アセスメントによる高次思考能力の測定の研究 H14-16科学研究費成果報告書.

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.

平井洋子 (2008). 作文の評価. 荒井克弘・倉元直樹 (編著) 全国学力調査 日米比較研究 金子書房: 81-100

平井洋子・椎名久美子・柳井春夫 (2001). 文系学生向き総合論述問題による能力測定の試み. 大学入試

センター研究紀要, *30*, 1-20.
平井洋子・渡部洋 (1994). 小論文評点のカテゴリ化に関する測定論的考察　行動計量学, *21*, 21-31.
Hughes, D. C., Keeling, B. F., & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. Journal of Educational Measurement, *17*, 131-135.
Hughes, D. C., Keeling, B. F., & Tuck, B. F. (1983a). Effects of achievement expectations and handwriting quality on scoring essays. Journal of Educational Measurement, *20*, 65-70.
Hughes, D. C., Keeling, B. F., & Tuck, B. F. (1983b). The effects of instructions to scorers intended to reduce context effects in essay scoring. Educational and Psychological Measurement, *43*, 1047-1050.
Hughes, D. C., & Keeling. B. F. (1984). The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, *21*, 277-281.
Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In: Willamson, M. & Huot, B. editors, (1993). Validating holistic scoring for writing assessment: theoretical and empirical foundations, Hampton Press, Cresskill, NJ. 206-236.
池田央 (1973). 心理学研究法8 テストⅡ　東京大学出版会.
池田央 (1992). テストの科学 ―試験に関わる全ての人に― 日本文化科学社.
池田央 (1994). 現代テスト理論　行動計量学シリーズ 7 朝倉書店.
池田央 (2007). 日本テスト学会学会賞受賞記念講演資料 於東京大学
井上俊哉 (1996). 論述式テストの利用について: 客観テストと比較して　東京家政大学研究紀要, 7-16.
石井巖 (1981).「論文試験」とその評価について　行動計量学, *8*, 22-29.
石岡恒憲 (2004). 記述式テストにおける自動採点システムの最新動向　行動計量学, *31*, 67-87.
石岡恒憲 (2006). 小論文/エッセイの自動採点システム過去,現在そして未来, 大学入試研究ジャーナル, *16*, 41-47.
石岡恒憲 (2012). エッセイおよび作文テストにおけるコンピュータ利用と自動採点　日本テスト学会第10回大会発表論文抄録集, 26-29.
石岡恒憲・荒井清佳・大久保智哉 (2012).　小論文試験得点と他教科科目得点および受験者属性との関連. 日本テスト学会第10回大会発表論文抄録集, 72-75.
James, H. (1927). The effect of handwriting on grading. English Journal, *16*, 180-205.
Johnson, R., Penny, J., and Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. Applied Measurement in Education, *13*, 121-138.
梶井芳明 (2001). 児童の作文はどのように評価されるのか？ ―評価項目の妥当性・信頼性の検討と教員の評価観の解明― 教育心理学研究, *49*, 480-490.
梶井芳明 (2002). 大学生は児童の作文をどのように評価するのか？ 日本教育工学会論文誌, *26*, 33-44.
木村拓也 (2010). 日本における「テストの専門家」を巡る人材養成状況の量的把握　日本テスト学会誌, *6*, 29-49.
Knoch, U., Read, J., & Randow, J.V. (2007). Re-training writing raters online: How does it compare with face-to-face training? Assessing Writing, *12*, 26-43.
倉元直樹・當山明華・西郡大 (2008). AO 入試の実情調査(1) ―大学入試の多様化とAO 入試―. 日本テスト学会第6 回大会発表論文抄録集, 82-83.
黒岩督 (1991). 教師の作文評価と作文の数量的指標の関連 学校教育学研究, *3*, 33-45.
Lawley, D.N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, *61*, 273-287.
Linn, R.L. (Ed.) (1989). Educational measurement (3rd ed.). Macmillan. (池田 央・藤田恵璽・柳井晴夫・繁桝算男 (編訳) (1992). 教育測定学 第3版. みくに出版.)
Lord, F.M. (1952). A theory of test scores. Psychometric Monograph, No 7.
Lord, F. M., & Novick, M. R. (1968).Statistical theories of mental test scores. Reading, Mass：Addison Wesley.
Lumley, T., & NcNamara, T. (1995). Rater characteristics and rater bias. Implications for training. Language Testing, *12*, 54-71.
文部科学省 (2012). http://www.mext.go.jp/b_menu/houdou/24/09/__icsFiles/afieldfile/2012/09/07/1325256_1.pdf (2012年10月16日閲覧)
村木英治 (2011). 項目反応理論 シリーズ行動計量の科学8　朝倉書店
村木英治・斉田智里 (2008). 調査デザインの考え方と方法. 荒井克弘・倉元直樹 (編著) 全国学力調査 日米比較研究　金子書房: 66-80.
村山航 (2003). テスト形式が学習方略に与える影響 教育心理学研究, *51*, 1-12.
村山航 (2012). 妥当性：概念の歴史的変遷と心理測定学的観点からの考察　教育心理学年報, *51*,

118-130.
成田秀夫・荘島宏二郎・宇佐美慧 (2010). ニューラルテスト理論を用いた大学生のジェネリックスキルを測定する試み　日本テスト学会第8回大会発表論文抄録集
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
大野木裕明 (1994). テストの心理学　ナカニシヤ出版
Penny, J. & Johnson, R.L. (2011). The accuracy of performance task scores after resolution of rater disagreement: A Monte Carlo study. Assessing Writing, *16*, 221-236.
Penny, J., Johnson, R. L, & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rublic. Assessing Writing, *7*, 143-164.
佐渡島紗織 (2003). アメリカにおける作文評価研究 -1997年以降の動向. 国語科教育, *53*, 42-48.
鷲坂由紀子・二村英幸・山岸建太郎 (2001). 採用選考における作文評価 : 達成動機測定の試み. 経営行動科学, *14*, 153-159.
崎濱秀行 (2005). 字数制限は書き手の文章算出活動にとって有益であるか？ 教育心理学研究, *53*, 62-73.
荘島宏二郎 (2007). ニューラルテスト理論. 日本テスト学会第5回大会発表論文抄録集, 174-177.
Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. DNC Research Note, RN08-01.
鈴木雅之 (2011). ルーブリックの提示による評価基準・評価目的の教示が学習者に及ぼす影響―テスト観・動機づけ・学習方略に着目して― 教育心理学研究, *59*, 131-143.
田栗正章・藤越康祝・柳井晴夫・C.R.ラオ (2007). やさしい統計入門　講談社ブルーバックス
平直樹 (1995). 物語作成課題に基づく作文能力評価の分析. 教育心理学研究, *43*, 134-144.
平直樹・江上由実子 (1992). ESSAY TEST の方法論的諸問題に関する研究の動向について　教育心理学研究, *40*, 108-117.
豊田秀樹 (2005). 項目反応理論(理論編) 朝倉書店.
植野真臣・荘島宏二郎 (2010). 学習評価の新潮流 シリーズ行動計量の科学4　朝倉書店
宇佐美慧 (2008). 小論文試験の採点における文字の美醜効果の規定因 ―メタ分析及び実験による検討―.　日本テスト学会誌,　*4*, 73-83.
宇佐美慧 (2009a). 小論文試験による評価データの心理計量学的性質の検討 -制限字数の影響に焦点を当てて-　東京大学修士論文(未刊行論文)
宇佐美慧 (2009b). ニューラルテスト理論の応用可能性 ―方法論的課題の考察と多値型モデルの適用例―　日本テスト学会誌, *5*, 65-79.
宇佐美慧 (2010). 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル. 教育心理学研究, *58*, 163-175.
宇佐美慧　(2011). 小論文評価データの統計解析 ―制限字数を考慮した測定論的課題の検討―. 行動計量学, *38*, 33-50.
宇佐美慧 (2012). 論述式テストを通した評価と選抜の信頼性に関わる諸要因の影響力についての定量的比較検討. 日本教育工学会論文誌, *36*, 451-464.
宇佐美慧・名越斉子・肥田野直・菊池けい子・斉藤佐和子・服部由紀子・松田祥子 (2011). 社会適応スキル検査の作成の試み ―検査の信頼性・妥当性・臨床的有用性の検討―　教育心理学研究,　*59*, 278-294.
van der Linden W.J, & Hambleton, R.K. (1997). Handbook of modern item response theory. New York: Springer.
渡部洋 (1994). 小論文試験の特徴とその利用法について 学校教育研究所年報, *38*, 48-59.
渡部洋・平井洋子 (1993). 段階反応モデルによる小論文データの解析. 東京大学教育学部紀要,　*33*, 143-150.
渡部洋・曹亦薇 (1992). 小論文評価における字の美しさの影響について. 東京大学教育学部紀要,　*32*, 253-256.
渡部洋・平由実子・井上俊哉 (1988). 小論文評価データの解析. 東京大学教育学部紀要, *28*,143-164.
Wiseman, S. (1949). The marking of English composition in grammar school selection. British Journal of Educational Psychology. *19*, 200-209.
吉川愛弓・岸学 (2006). 作文の評価項目に関する検討 ―意見文の評価は何に影響を受けるのか― 東京学芸大学紀要, *57*, 93-102.
吉村英 (1991). ワープロ文字と手書き文字の違いが文章の内容の評価に与える影響(Ⅰ). 日本心理学会第55回大会発表論文集, 753.
吉村英 (1992). ワープロ文字と手書き文字の違いが文章の内容の評価に与える影響(Ⅱ). 日本心理学会第56回大会発表論文集, 539.
吉村英 (1993). ワープロ文字と手書き文字の違いが文章の内容の評価に与える影響(Ⅲ). 日本心理学会第57回大会発表論文集, 803.
吉村宰・木村拓也 (2008). テストスタンダードを満たす大学入学者選抜を目指して―N 大学における事例―日本テスト学会第6回大会発表論文抄録集, 78-81.