

論述式テストの測定論的問題再考

—主要な論点の整理とその現実的解決のために—

宇佐美 慧
(筑波大学大学院人間総合科学研究科)

はじめに

入学試験や入社試験をはじめとして、応用的な思考能力、表現力、構成力、独創性、他にも広い意味での実技能力といった、一般に客観式テストでは測定が難しいとされる、受検者のもつ「高次の能力」を測定・評価することのニーズが高まっており、それに伴い論述式テスト（小論文試験）・面接試験・パフォーマンステストによる測定・評価が広く社会的に浸透してきたことは周知の事実であろう（宇佐美, 2008, 2011, 2013）。ところが、この「高次の能力」を測定・評価する試み自体は非常に魅力的である半面、特に論述式テストの場合を中心に、それに基づく測定の信頼性・妥当性、或いはバイアスといった測定論的問題（測定論的問題全般を扱ったレビューとしては、平・江上, 1992; 宇佐美, 2013）が根強く指摘されている。とりわけ最近では、様々なテスト開発場面において項目反応理論（IRT: 論述式テストへの適用を想定した項目反応モデルとしては、例えば、宇佐美, 2010; 宇都・植野, 2014）の実用が期待されているが、この測定論的問題を解決せずに項目反応理論に則った運用を進めることは現実的ではないだろう。

測定論的問題を解決するためには、例えば、(i) 十分な数の項目や採点者を用意すること、(ii) 採点基準について採点者間で十分な協議を経て基準を共有すること、(iii) 採点の訓練を十分に行うこと、といった点が挙げられる。重要なのは、このように、論述式テストの場合、テストの仕様についての諸要因（項目数や項目間相関）のみだけでなく、採点者に関わる諸要因（採点者数、採点者間相関、採点者内相関）についても配慮が必要とされるという点である。これによって、測定論的問題を考慮した論述式テストの運用を行うとなると、客観式テストに比べて、時間的・人的・経済的コストがしばしば大きくなってしまふ。しかし、とりわけ大学入試（や適性試験）のような集団式のテストを実施する場合、受検者の能力・特性をなるべく簡便に測定したいという意図も少なからずテストの実施者側にあるのが一般的である。したがって、論述式テストを運用する際には、運用上のコ

ストの大きさと測定論的に見た際のテストの質の高さの二つを天秤にかけながら、現実的にどのような形で運用が可能なのかを考えることが必要になる。ところが、測定論的問題の概略やその解決のため現実的な手立てについてはこれまでに多くが既に語られているものの、測定（テスト）の専門家が慢性的に不足している我が国では（例えば、木村, 2010）、そもそも測定論的問題の深刻さやそれを考えることの意義を、テストに関わる様々な人の間で十分に共有できているとは言い難い。加えて、測定論的問題は、非常に多面的で複雑な問題であり、「測定論的に見た際のテストの質の高さ」を正しく吟味することは、実際の運用場面ではしばしば困難な作業となっているように思われる。

このような現況を踏まえると、教育測定の観点から論述式テストの運用の在り方を見直す際には、測定論的問題の深刻さを改めて知ると同時に、現実的な解決策のあり方を考え、議論を深め問題意識を共有していくことが実り多いであろう。本発表では、宇佐美 (2013) のレビューで挙げた論点の中でも、特に重要度の高いものや、またはテストに関わる教育測定専門家・実務家・テストの作成者の間でまだ共有が十分にされていない点を挙げて、解決の指針とともに整理する。また、関連して、今後の更なる議論や検証が必要な、測定論的問題に関わる重要度の高い論点を取り上げる。

論点の整理と解決の指針

1. 測定論的に主要な（共有すべき）論点

(a) 項目数と測定論的問題の関係

項目数は信頼性（内的整合性）に関わるのは勿論であるが、「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）としての内容的妥当性にも非常に強く関わる。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が、その結果項目数が一題のみである場合も珍しくなく（平井, 2007）、これは信頼性のみならず、内容的妥当性の観点から見ても大

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、更には採点のバイアスの観点からしても問題になる可能性が高くなるだろう。

(b) 信頼性を高めるための方法—項目数・採点者数—
とりわけ採点の訓練や採点者間での事前協議を綿密に行うことができた場合、採点者間相関は高い水準となることが期待できる。このような場合は、採点者数を多く設けたとしても、元々の採点結果の間に高い一貫性があるため、その信頼性の改善の寄与度は相対的に小さくなる。一方、現実の場面においては、項目間相関は採点者間相関よりも平均的にしばしば低くなる傾向がみられる(例えば、宇佐美, 2011)。したがって、採点者数・項目数は、いずれも信頼性を高める上での必要条件であるが、特に項目数を増やすことの重要性は高く、また上記(a)の論点も踏まえると、それは内容的妥当性を高めることにも繋がる。

宇佐美(2011)では、実際の論述式テストデータに一般化可能性理論(詳細は、Linn, 1989; 宇佐美, 2013)を適用した結果、あくまで単一の研究事例であるが、(i) 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(つまり、信頼性)の増加が期待できること、(ii) 採点者数は概ね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、(iii) 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名の場合は概ね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問程度の項目数が、さらに採点者数が4~5名程度であれば2問程度の項目数が必要であることを示している。また、(ii),(iii)にも関わる点であるが、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されている。相対的には項目数を増やすことの重要性がより高い場合が多いと考えて良いだろうが、この点は実践上踏まえておくべき観点であろう。さらに関連して、採点者間相関や項目間相関の(平均的な)予想値が分かれば、任意の採点者数や項目数のもとで信頼性の高さがどの程度になるのか予め見積もることができる(または、複雑な採点計画の下では、シミュレーションを利用すれば同様に見積もることができる)ため、実際の運用ではこのような事前の概算をしておくことが得策ではないかと思われる。

(c) 採点訓練・採点方法

冒頭で述べたように、採点基準について採点者間で十分な協議を経て基準を共有することや採点訓練を十分に行うことは勿論重要であるが、採点訓練の手続きや、採点方法と信頼性の関係については明確

に論じられていない、或いは十分な研究蓄積がない側面のように思われる。ベンチマーク(各評価得点に対応する典型的な答案例)を得点ごとに幾つか作るとは、採点訓練に活用できるだけでなく、採点基準について見直す上でも役立つ。また、米国の全国学力調査に関する平井(2008)の興味深い報告例にもあるように、採点の途中に採点者間で結果を比較したり、ベンチマークの答案を実際の答案の中に混ぜて採点結果の偏りの有無を調べるような中間的作業も、一定の時間的コストを伴うが、一定の採点者間相関・採点者内相関を担保し、また採点のバイアスを抑制する上では必要不可欠である。

また、採点方法については、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対処するかはしばしば問題になる。これに関しては、単に評価点の平均をとる場合や、より熟練した採点者の結果を優先する(あるいは重みづけする)などの場合がある。これらのうちどの方法が良いかは、個々の採点者の熟練度に依存する問題であるため一概には言えないが、少なくとも項目数や採点者数、あるいは項目間相関・採点者間相関・採点者内相関に比べれば、その方法の違いによる信頼性への影響度は非常に小さいことが分かっている(宇佐美, 2012)。さらに関連して、論述式テストでは、五段階や七段階などの何らかのカテゴリ数をもつカテゴリ得点を用いて採点が行われることが多いが、このカテゴリ数の設定についても、それが信頼性に与える影響は、項目数や採点者数などに比べればずっと小さいようである。

(d) 回答内容の自由度と測定する能力・特性の関係

各々の論述式テスト(項目)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、想定される回答の内容の質や幅の広さは、テスト(項目)間で変動することが予想される。この、回答として想定される内容の幅の広さを、宇佐美(2008)は回答内容の自由度と呼んでいる。また、回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(i) 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(ii) 回答方法の要素、の二つに分けて論じている。

(i) については、例えばグラフの結果を読み取って論述する形式の項目への回答は、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回答内容の自由度が小さくなると考えられる。その結果、測定できる能力・特性の範囲は狭くなる一方で、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、回答内容の自由度は狭まりやすいであろう。(ii)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」のような論述の手順(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、回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。回答内容の自由度は、信頼性や妥当性、つまり測定論的問題全般に直結する観点である。

回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るだろう。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、本来の測定意図とは逸脱した回答を示す答案が増えるなど、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可能性も考えられる。したがって、テストの作成時には、測定を意図した能力・適性の定義について明確化するとともに、それに対して不必要に回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ことは重要である。

2. 今後の議論・検証が必要な論点

(a) 制限字数・制限時間

回答内容の自由度の関係と信頼性・妥当性の関係は全般的に研究ベースでの検証がまだまだ必要な領域であろう。特に、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論点として、前節(d)の(ii)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(2011)は、制限字数の設定の違いが採点結果の信頼性・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(無論、測定の目的によっては、制限字数を短くすることで却って別の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, 2011を参照のこと)。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・特性の違い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。

(b) 得点の離散化

入学試験よりは資格試験や適性試験の文脈の方がより一般的であるが、採点結果について素点(合計得点)だけを受検者にフィードバックするのではなく、例えば能力判定の結果をA,B,C,D,Eなど、何らかの手続きで離散化し作成された段階(ランク)を併用する(またはそれだけを利用する)場合がある。このような得点の離散化にはどのような意義があって、どのような文脈で使い得るのだろうか。評価段階数が比較的少数の方が、受検者が自身の能力水準を直感的に把握する上では便利かもしれない。また実施者側からすれば、一定の得点範囲の受検者層には同じ内容のフィードバックをした方が運用上合理的なこともあるだろう。

他にも重要な点として、素点そのものの信頼性が低い場合は、測定誤差の大きさを考慮しながら、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、受検者に対して不適当な評価を与える可能性は低くなるであろう。つまり、任意の評価段階数のもとで、ある受検者が真に属する評価段階に正しく割り当てられる(=適切な評価を受ける)可能性は、評価段階数が少ない方が高くなるであろう。ただ、離散化をする上では、段階数をいくつに設定し、またどのような方法の下で、どこの得点範囲に対して各段階を設定するかは、純粋に測定論(統計)的な問題のみからでは解決し得ない複雑性がある。また、各段階が質的にどの程度の能力水準を意味するのかを記述することは、段階数を非常に少なく抑えた場合を除けば、しばしば難しい問題でもある(宇佐美, 2009)。

また、上記の利点はあくまで各受検者の能力とフィードバック上の整合性に関することであり、基本的には、素点の方が離散化した段階得点よりも情報量が多いことには注意する必要があるだろう。したがって、一般に入学試験のような、得点に基づく受検者の序列化(順位化)が最重要とされる文脈においては、通常は素点を利用の方が望ましいであろう。つまり、各受検者にどのようなフィードバックを与えることが適切なのかという点からすれば離散化をすることの意義はあるが、あるテスト得点が高い受検者は、低い受検者よりも能力・適性水準が高いことは、確率的に見れば正しい判断のはずである。しかし、得点の離散化の意義とその適正な使用の文脈、および離散化をする上での方法論上の問題については、まだあまり議論と検証がし尽くされていない領域であるように思われる。

おわりに

本発表では、測定論的問題の中でも重要度の高いものや今後より議論または検証が必要な論点について触れた。最後に、これらを考える以前の問題として踏まえるべき、基本的な幾つかの論点について更に挙げておきたい。まず、テストの項目作成段階から測定論的問題を十分に考慮するという点である。具体的には、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回答手順を通して測定していくかを、項目数・採点者数や課題の難易度、採点計画、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とが必要である。とりわけ、測定を意図する能力・適性といった構成概念の意味内容を予め明確化しておくことは、後のテストの仕様の決定や、テストの妥当性を評価する上での判断基準を与えることにも繋がる。

次に、本発表では、測定論的問題と言いながらも、信頼性に関わる点を重点的に議論したが、それは信頼性が妥当性を充足する上での必要条件だからである。ただし、妥当性に関しても、特にその示し方については踏まえるべき基本的な観点があるだろう。妥当性自体は、構成概念妥当性という単一の統合的な概念で捉えるべきという見方が強まっているが、妥当性が充足されていることを当該のテストで示す上では、どのような側面の妥当性（例えば、内容的妥当性、基準関連妥当性、因子的妥当性、結果的妥当性）について議論しているかを明確にすることは重要である（論述式テストにおける妥当性の諸側面については宇佐美, 2013 を参照のこと）。実際に、一側面の検証に留まりながら、妥当性全体が宛も示されたと安易に述べてしまう事例は後を絶たない。一方で、単一の研究や調査データからは全ての妥当性の諸側面を調べることは非常に難しく、妥当性の証拠は本来は研究の蓄積を経て徐々に積み上がっていく性質のものであることを踏まえるのも重要であろう。

冒頭でも述べた点であるが、最後に、テストの専門家の不足の問題も再度挙げておく必要があるだろう。測定論的問題を踏まえて運用上の現実的な解決を目指すためには、教育測定についての専門的な知識・経験が少なからず必要であり、そのための人材の育成がまず必要と考えられるが、教育測定の専門家、または一定以上の知識・経験をもつ実務家・作成者の数はまだ十分ではない。したがって、このような人材育成のための継続的な教育・啓蒙活動は、我が国ではまだ当然重要度が高く、また緊急性も高い事項と言えるだろう。

引用文献

- 阿久津洋巳・菊池梢・鈴木安澄・鈴木光・渡邊愛枝 (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.
- 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.
- 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.
- 池田央 (1973). 心理学研究法 8 テスト 東京大学出版会.
- 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.
- Linn, R.L. (Ed.) (1989). Educational measurement (3rd ed.). Macmillan. (池田 央・藤田恵璽・柳井晴夫・繁樹算男 (編訳) (1992). 教育測定学 第3版. みくに出版.)
- 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.
- 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 (2009). 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). 論述式テストを通じた評価と選抜の信頼性に関わる諸要因の影響について — 定量的比較検討. 日本教育工学会論文誌, 36, 451-464.
- 宇佐美慧 (2013). 論述式テストの運用における測定論的問題とその対処 日本テスト学会誌, 9, 145-164.
- 宇都雅輝・植野真臣 (2014). ピアアセスメントの低次評価者母数を持つ階層ベイズ項目反応理論 教育システム情報学会発表論文集, 465-466.