

# マークシート形式のテストは何を測っているのか？ —櫻井・藤井・岩坪・伊藤・松田論文へのコメント—

名古屋大学 村上 隆

はじめに この論文は、入試センター試験の問題を記述式（筆記解等方式）に書き換えて、それを元の問題とともに同じ高校生の集団に対して実施し、その結果を比較することによって、マークシート形式のテストが測っている能力の性質を明らかにしようとしたものである。

筆者はまず、大学入試センター研究開発部の先生方が、単に入試センター試験の得点の分析だけではなく、（もしかしたら）センター試験では測れない学力の存在を示すことになるかもしれないような、センター試験と個別試験の関係にまで視野を広げた泥臭い研究に取り組まれたことに心から敬意を表する次第である。しかも、そのことが、大学の初年時教育に対する危機感を背景にしていると読み取れることは、近年、ますます大きな問題となりつつある高大接続改善という視点からも、きわめて意義の大きいものであると考える。

センター試験と個別試験 筆者は図1を見て強い印象を受けた。これはまさに、私が学内の入学者選抜方法研究委員会の報告書等で見慣れた形だからである。すなわち、散布図の右上部分に点が集中し、左下にはほとんど点がなく、全体としては（中程度の）正の相関があるといった分布の形状である。これは、マークシート方式の試験で高得点でも、記述式の得点は高くない受験者が存在する一方、記述式の得点の高い受験者の中にマークシート方式の得点が低い者はほとんどいないことを意味している。こうした結果は、数学等の理系科目に関して、多くの大学で繰り返し発見されている事実であり、マークシート形式のテストで測られる能力は、記述式のテストで測られる能力の必要条件だが十分条件ではないと表現することができよう。

入試センター試験で測られる能力が、個別試験のそれとは全く無関係であるとする議論を時々見かけるが、これでわかるように決してそうではない。また、実際の試験の分析結果にお

いて見られる現象が、ここでも見られることは、この実験の妥当性を示す1つの証拠となつていられると思われる。

確率絞り込み型ランダム・ゲッシング しながら、本研究において、中心的に扱われているのは、記述型試験における解答のプロセスよりは、マークシート型の問題項目に対するそれである。

マークシート方式の試験に対するよくある批判として、選択肢の中には必ず正答が存在することから、その項目において問われている能力が全くなくても偶然で正答できるのが不合理だとするものがある。もちろん、それは極論であり、実際には項目数を増やすことによって、偶然による正答（いわゆるまぐれ当たり）の効果は減らすことができる。

しかしながら、これを全く無視することができないことは、たとえば、桜井らが引用している池田（1985）において、明確に示されている。そこでは、

1. その問題項目で問われている能力を持っている受験者（池田は真の正答者と呼ぶ）は、必ず正答の選択肢を選ぶ、
  2. その能力を持たないものは、すべての選択肢を項目間で独立に等確率で選ぶ、
- という前提の下で、実際の得点分布が、真の正答者だけが得点できるような形式の問題からなるテストの（仮想的な）得点に比べて平均値が上昇し、標準偏差が縮小することを理論的に導き、幾つかの典型的ケースについてそれを図と表で示している。

もっとも、マークシート方式と言っても、解答の候補をいくつか示し正解を選ばせる多肢選択形式ばかりではなく、数学で採用されているように、解答となる数値を選ばせるもの（これによって、ランダム・ゲッシングによる正答確率は大幅に低下する）もある。また、多くの科目では、いくつかの項目が1つの大問の下に組

織的に配列され、前の小問に対する解答が、後の小問で利用されるようなものもある。そうした出題形式は、項目間に人工的な相関を作り出す（いわゆる共倒れである）ために、解答行動の検討やテストの信頼性の正確な推定を行うには適当でない。「物理」の問題項目が、伝統的に5肢選択形式の離散点型（相互に論理的に関係しない問題項目をこのように呼ぶ）であったことは幸いであった。

さて、そうした条件の下で櫻井らの得た結果は、低学力の受験者たちは回答選択肢の範囲を2つから3つに絞り込んだ上で、ランダムな解答をしていると仮定すると、かれらの得点分布が説明できるという驚くべきものであった。

入試センター試験で測られている能力 この結果は、池田（1985）における想定が2つとも成り立たないことを示している。櫻井らはこれについて、「マーク方式が『全くのランダムゲッティング』ではなく、正解の可能性が高い選択肢を絞り込むという操作の中で、・・・ある種の学力を測定している」と述べているが、それは適切な解釈なのだろうか。

ある確率をもつ2項分布が当てはまるということは、全受験者がその確率ですべての問題項目に「ランダムに」解答しているという想定と矛盾しないデータが（低学力の受験者に限ってではあるが）得られたことを意味する。確かに、すべての選択肢を等確率に選ぶという意味でのランダムではないとしても、絞込み確率に個人差がないということは、通常の意味で「測っている」ことにはならないのではないか？

実際、項目間相関がほとんどないことにも言及されているが、これは、ここで対象にされた

テストの得点に、低学力者を対象とする限り、古典的テスト理論の意味での内的整合性が欠けていることを意味している。そして、内的整合性は何らかの意味のある個人差を測定する最低限の条件である。

このことは、受験者たちがどうやって選択肢を絞り込んでいるのか、という疑問につながる。何らかの意味のある知識なり技能なりが用いられているとすれば、その習得の度合いに（たとえ低学力者の間であっても）個人差があると考えるのが自然であろう。全員が等しく身に付けている能力とはどのようなものなのであるか？

かつて、数学は暗記科目であるとする議論が受験関係者から提起され物議をかもしたことがあった。低学力者が選択肢を絞り込む能力とは、相互に関連のない断片的な知識、あるいは単なる受験技術の機械的記憶なのであるか（それでも、個人差がないことの説明にはならないが）。そうした砂を噛むような記憶のトレーニングがセンター試験のための準備教育でないことを、切に望みたいものである。

そうした疑問に答えるためには、やはり、受験者の解答過程をより具体的に明らかにするような更に泥臭い研究が必要であろう。

終わりに 繰り返しになるが、こうした研究が大学入試センター研究開発部の手によって始められたことは、まことに心強いものがある。今後、より一層、入試センター試験の性質を明らかにするような研究が行われることを期待したい。