健康文化　28 号
2000 年 10 月発行

健康文化

試験は常に不完全である

村上 隆

１．試験をめぐるスキャンダル

あまり愉快な話題ではないが，このところ歯科医師国家試験の問題漏洩がニュースになっている。テストのことがマスコミで話題になるのは，大学入試センター試験や個別大学の入試の出題ミス，それに，同じくセンター試験の科目間得点較差など，よからぬことばかり目立つのは，心理・教育測定を専門としているものにとって，大変残念なことである。

日本では，試験の問題が正しく作られ，不正行為がないように実施されれば，それで問題なしという考え方が強い。もちろん，出題ミスも不正もあってはならないことだが，それ以外にも，試験を利用するにあたって注意しなければならない事柄は多い。たとえば，本稿でとりあげる試験の得点の信頼性である。

２．得点の誤差と信頼性

どんな測定にもランダム誤差が伴うことは，自然科学においては常識であるが，能力（ability，あるいは，proficiency）や（教育課程における）到達度（achievement）を測る試験の得点もまた，ランダム誤差を免れない。これは，試験の測定装置としての性能であり，試験問題の作り方や問題項目数によって異なるが，ランダム誤差の大きさは，通常，予想されるよりずっと大きい。

試験の得点の信頼性は，概念的には，同一の測定を 2 度反復して得られる２つの得点の間の相関係数として定義される。これを信頼性係数（reliability coefficient）と呼ぶ。もちろん，記憶をもっている人間に対して，同じ試験を全くそのまま反復しても意味はないから，2 回の測定に用いられる試験は平行テスト，つまり，問題自体は異なっているが，測定の内容に関しては同質な２つの試験という意味にとっておいていただきたい。実際には，2 度試験を実施するような手間をかけている暇は，（受験者の側に）ないから，１つの試験の問題を２つに分けて採点し，その間の相関係数から求める値を推定する等の方法がとられる。いずれにせよ，ここで相関係数が１であれば，その試験の得点にはランダム誤差が全く含まれず，信頼性は完璧である。このとき，試験の得点は，そ

のとおりに受け取って，資格認定や選抜の目的に用いてよい。

しかしながら，相関係数が１になることは，次の３つの理由だけを考えてもありえないことである。まず，多くの公的試験が現在とっている客観形式，通常，多肢選択形式（４つまたは５つの選択肢の中から正解を選ぶ形式）は不可避的に誤差を作り出す。(試験の「客観性」とは単に，この多肢選択形式のように，誰が採点しても同じ結果が得られるということを指す。)

かつて，このような多肢選択形式の項目からなる試験について，「もし正解したとしても，本当にわかっていたのか，まぐれ当たりなのかわからないから不合理だ」という批判があった。実際，この批判は１つの問題項目についてだけ考えれば，まったく正当である。ある受験者が，ある問題に正解していたことから，その個人が「わかって」いるかどうかを推測することは（その解答だけしか情報がない限り）不可能である。それは，他の（多数の）問いに対する解答と照らし合わせて初めて（ある程度）明らかになるにすぎない。そこで，まぐれ当たりは試験の得点の誤差の避けられない原因となる。

では，同じ問題項目に対して誤答した受験者は，全くわかっていないと断定してよいだろうか？　そうはいかないであろう。

解答に当たって不注意な誤りは起こりうるし，意図したものでなくても，正しく理解している受験者を惑わして誤答に導くような手がかりが，設問に含まれてしまうことも避けられない。つまり，あからさまに出題ミスでなくても，問題項目には誤差を引き起こすような要因が必ず混入する。測定誤差が発生する第 2 の理由は，この作題上の不備である。ただ，試験の信頼性に関して改善の余地があるとすればこの部分であり，試験の事後的な分析結果を検討すること等を通じて，作題者のスキルを高めていく努力が必要である。

第 3 に，試験の問題を無限に増やすことはできないことが問題になる。無限といわずとも，十分な時間が掛けられれば，試験したい領域のすべてを網羅する試験問題が作れるであろう。しかしながら，有限の（通常，必要とされるよりかなり短い時間で）終了させなければならない試験の項目数では，問題はどうしても一定の偏りをもたざるを得ない。受験者は，領域による得手不得手があるから，平行テストであっても，得点は毎回変動する。

さて，それでは実際のテストで目指されるべき信頼性とはどの程度か，ということであるが,先にのべた相関係数にして,0.85 から 0.90 程度とされている。大学入試センター試験の「数学」と「英語」は，だいたいこの水準に達しているが，入学者がよく文句を言っている「国語」は，これをかなり下回っている模様である。予備校の模擬試験と 1 ヶ月おいて実施される本試験の自己採点結

果との相関から推測される。

相関係数が 0.80 を超えると言えば，研究成果発表では，「極めて高い相関が得られた」と誇らしく語られるところであろう。しかし，試験に関しては，とてもそんな風に言えるものではない。図 1 は，相関係数が 0.85 であるような試験を反復した場合の仮想的な散布図である。

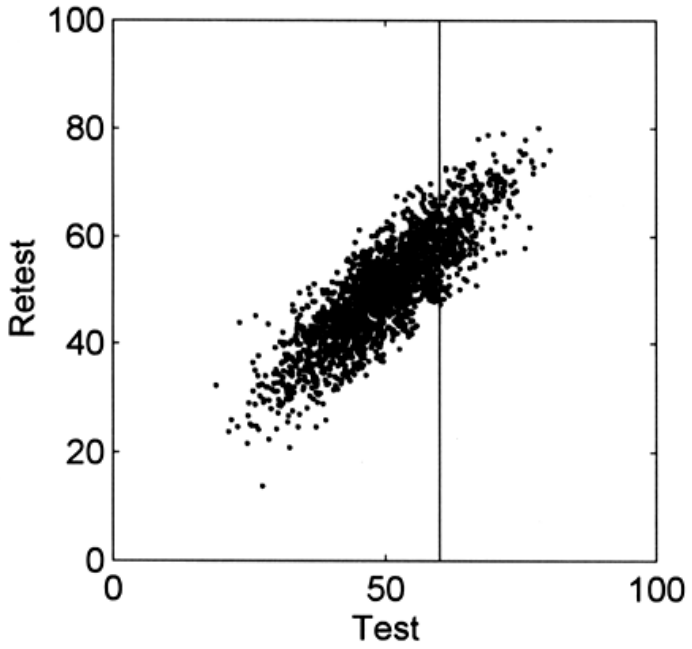

図 1　信頼性係数が 0.85 のテストを 2 度反復したときの得点間の相関関係を示す散布図。縦の線をたどってみれば，1 回目の得点が 60 点だった受験者が 2 回目で取る得点の分布状況（誤差の大きさ）がわかる。

測定をもう一度反復したときの得点の標準偏差（推定の標準誤差）は，

$$s_e = s_x\sqrt{1-r^2}$$

によって与えられる。（完全な信頼性，すなわち，$\gamma = 1$ なら，もちろん標準誤差は 0 である。）ここで，$\gamma = 0.85$ を代入してみると，

$$s_e \approx 0.5 s_x$$

であり，このことは，測定誤差の標準偏差は，信頼性係数が 0.85 程度のテストでは，測定値全体の標準偏差の半分程度にしかならないということである。図 1 のように得点を偏差値（平均値が 50，標準偏差が 10）では，標準誤差は 5 点に及ぶ。つまり，能力が変わらないまま，同質のテストを何回も受け続ければ，

その得点は一定どころか，標準偏差 5 点の分布を成すということである。

何か日常的な例をあげてみよう。我が家にある体重計の信頼性が 0.85 であるとしてみよう。体重の標準偏差は 5 kg として，標準誤差は約 2.5 kg である。これでは，今一つ実感が湧かないが，ある体重計の信頼性係数が 0.85 であったとして，たまたま 65kg を示した 20 名の人の体重をもう一度測り直したとすれば、次のような結果が得られるであろう。

65.5，64.0，63.0，64.0，66.0，69.0，59.0，66.0，65.0，66.0
65.5，61.5，61.0，67.0，67.5，65.5，66.5，66.0，64.5，63.5

ただし，0.5 kg 単位で読み取ったものとし， 正規乱数を用いてシミュレーションした。

体重計の表示がこれほど変動するとすれば，「しばらく続けていたダイエットが，やっとちょっと効果を表わしてきた」などということは，ほとんど判断できないであろう。しかし，これが現状において，かなりよくできた試験の実態に近いものなのである。つまり，日常生活の中にはこれほど誤差の大きい測定は見当たらず，したがって，われわれは試験の誤差を考慮に入れて選抜や資格認定の合否を判定することが，うまくできないのである。

３．公的試験の信頼性は開示できないか？

筆者は，医療関係の公的試験にはタッチしていないので，確かなことは言えないが，その実状がここで述べたのと，そう違わないことは予想できる。

試験の信頼性のような本質的な問題について話題にできないのは，試験に関する重要な情報がほとんど開示されていないからでもある。わが国では，試験問題の開示に向けての圧力が強く，大学入試センター試験など，実施された試験の問題が，ただちに翌日の新聞に正解つきで掲載される。その一方で，得点の統計量も，科目別の平均値等，最小限のものにとどまっている。

もちろん，情報開示はそれについて一定の正確さで理解でき，冷静に議論できる人々がいて初めて有効なものになる，というのも事実である。このまま開示すれば，短絡的な廃止論につながりかねない（かわりに提案されるのは，信頼性の観点からは，間違いなく，さらに改悪になるような案である）ような現状からすれば，安易に開示できないと考える実施主体の態度も解せないわけではない。問題は一朝一夕に解決できるようなものではない。だが，ここにもわれわれの健康との関わりが無視できない大きな問題があることを指摘しておきたいと思う。

（名古屋大学教育学部教授）