

Computer Assisted Personal Interview における ランダムマイゼーションの是非

西澤 由隆

要旨

コンピュータを利用した意識調査では、評価項目を自動的に入れ替える「ランダムマイゼーション」が一般的である。そのことで、評価項目の順序によるバイアス（順序効果）が、集計値としては排除される。ところが、個人単位で見たときには、アンカリング効果のため、質問の「信頼性」を損ねることが判明した。Web 調査を利用してその改善策を検討したが、質問形式の操作でアンカリングの効果を排除するのは難しい。したがって、ランダムマイゼーションは選択的に利用される必要があるとの結論を得た。

I. はじめに

世論調査の手法が社会科学に導入されてから半世紀がすでに経過している。その間に、社会科学の1つのツールとして揺るぎない地位を世論調査は確立した。日常的に実施されている市場調査、あるいはマスコミ各社が実施する意識調査、そして学術的な全国規模調査まで、世論調査の実施例は多い。また、最近では、国境を越えた比較研究も盛んで、その多くが大規模な比較世論調査を実施している。

ところが、有権者の意識や政治態度の実態をどの程度正しく世論調査が「炙り出して」いるかということについては、いまだに未知数な部分が多い。質問文における言葉の選択・回答肢の提示順・質問の書式など（いわゆる「コンテクスト」）が、回答に影響を与えることは確認されているが、それらに対する具体的な処方についての検討は、十分に進んでいるとは必ずしも言えない。とりわけ日本では、これまでに実施されてきた調査の総数に比較して、調査手法についての実験的な研究は少数であると言わざるをえない。世論調査を用いて検討されようとしている政治学的概念について、研究者が想定しているとおりに回答者が質問を「正しく」理解して、その意図に沿った回答を常にしているのだろうか。

しかも、近年では、パソコンを利用した世論調査（computer assisted personal interview、CAPI）が盛んとなってきている。それは、従来の紙とペンによる調査（paper and pencil interview）では困難であったより複雑な質問形式を可能とした。そしてそのおかげで、「真の世論」の表出に各種の世論調査が近づいているかのようでもあるが、一方で、これまでにはなかつ

た新しいバイアスを導入している可能性も否定できない。普及が急速であるからこそ、新しい技術についての慎重な評価が望まれるところでもある。

そこで、本稿では、学術的な調査ではすでに「標準」となりつつあり、また、民間の調査でも浸透が進む「評価項目の表示順に対するランダム化」について、それがもたらす新たな問題点と対応策を紹介しながら、「新しい技術」についての慎重な評価の必要性を提起したい。

Ⅱ. 選択肢のランダム化

面接・電話・郵送を問わず、いかなる形式の調査にせよ、質問文の書式が回答に及ぼす影響についてはすでに多くの研究がある (Sudman and Bradburn 1982, Tourangeau *et al.* 2000)。たとえば、いくつかの選択肢の中から回答を1つを選ぶような場合、選択肢リストの最初に来る項目が、他の項目に比較して選択される確率が高い。最高裁判所の裁判官の国民審査の際に、リストの最初に掲載された裁判官に「×」印が多く集まる傾向は、よく知られた事実である。

このように、選択肢の順序が回答に与える影響を「順序効果」と呼ぶが、それを「解消」する1つの対策が、回答肢の提示順におけるランダム化である。回答肢の表示順を固定せず、全ての可能な組み合わせ(「順列」)の表示パターンの中から、回答者ごとに無作為に選ばれた1組を用いて質問をするのである。

質問紙による従来の調査では非現実的であった「選択肢の提示順のランダム化」が、世論調査にパソコンを利用することできわめて容易となった。たとえば、ある選挙での投票政党を1つ選ばせるような質問の場合、政党の表示順をランダムに入れ替えて質問票(あるいは回答カード)を印刷することは不可能でも、CAPIであれば、それはプログラム上の操作として容易に実現ができる(図1参照)。実際、このようなケースでは「ほとんどの場合、自動的にランダム化をします」と、ある調査会社のプログラマーが証言をするくらいである¹⁾。

表示される項目順を無作為に変えることで、その「悪影響(バイアス)」を回答者間でばらつかせることになる。そのことを通して、全体としての回答分布を集計する際に、そのバイアスを相殺させようというものである。

Ⅲ. 評価項目のランダム化における新たなバイアス

上記の議論は、1つの問をいくつかの評価対象に繰り返し行うような質問形式についてもあてはまる。複数の評価対象(たとえば、テレビ・新聞・警察・病院など)に対して、「あなたはどの程度、信頼しますか」と問う「制度信頼」の質問は、その典型的な質問形式である(図2参照)。そして、ここでは、評価項目をどの順で提示するかが問題となる。「テレビ」についての評価から回答を求めると、それとも「新聞」からなのか。そして、第2番目に表示するのは、「警察」なのか、それとも「大企業」なのかという問題である。なお、その際の回答肢は、「信頼する」

The screenshot shows a survey interface titled "Cyber Panel" with a sub-header "社会的な組織や団体についてのアンケート". The question is Q15: "もし、今日、衆議院選挙があるとしたら、政党名で投票する比例区で、あなたはどの政党に投票しますか。この中から1つだけお選びください。(1つだけ)". The options are randomized and include: 公明党, 民主党, 社民党, 国民新党, 自民党, 共産党, その他の政党, 白票を投じる, 棄権する, and 言いたくない. A "次へ" button is visible at the bottom right, and the logo for "NR 日本リサーチセンター Member of Gallup International" is at the bottom left.

図1：選択肢のランダムマイゼーション、投票政党の場合

から「信頼しない」の尺度である（図3参照）。したがって、回答肢は、評価項目ごとに固定されていて、その表示順序はここでは問題とならない。代わりに、評価の対象の表示順が問題となるのである。そして、表示順により生じるバイアスを排除するために、評価項目順もランダムマイズすることがありえるし、また、これも一般的となりつつある。

ところが、評価項目のランダムマイゼーションの場合は、それがデータの「信頼性（reliability）」を損ねる可能性が考えられる。それは、世論調査に答えようとする回答者の、少なくとも3つの認知心理的な特性（「尺度の部分利用」・「相対評価の傾向」・「アンカリング」）から生じる問題である。

「制度信頼」などの質問は、「信頼しない」から「信頼する」まで、「0点」から「10点」の11点尺度で評価を求めることが多い（図3参照）。ところが、11点のモノサシの全幅を利用して評価をする回答者は少なく、その尺度の一部を使ってのみ評価をするのがむしろ一般的である。これをここでは「尺度の部分利用」と呼ぶ。

また、回答者は、評価項目のそれぞれを個別に絶対評価をするより、その直前の項目との相対比較で評価をする傾向が見られる。「テレビを6点としたのだから、それよりは少しは信頼のできる新聞は7点」というような評価の仕方である。いわゆる「順序効果」が観察されるのは、そもそもこの「相対評価の傾向」があるからである。

そして、評価項目のランダムマイゼーションにおいては、これら2つの傾向性が相互に補完し

合い、「回答者の評価項目全体に対する評価基準点が、第1評価項目に依存する」という「アンカリング(狭義)」の問題をより深刻にする。回答者単位でのリスト全体に対する評価の平均値が、表示リスト順(とりわけ、先頭の項目)によって異なる可能性がある。

たとえば、ある回答者Aについて、たまたま本人が好意的に(信頼できると)評価している項目(たとえば「裁判所」)から質問がスタートしたと仮定しよう。その場合は、「評価基準点」が信頼尺度の高い方で「アンカリング」され、それを中心にその他の項目の評価が続くことになる。一方、もう一人の回答者Bについては、非好意的(たとえば、「国の政治」)からスタートしたとする。すると、「評価基準点」が尺度の低い方に「アンカリング」され、それを中心にその他の項目を評価することになる。

このような場合、回答者A・Bそれぞれにおいて、項目間の相対的な評価は信頼に足るものであると言えるかもしれない。たとえば、警察には「5点」を、病院には「7点」を回答者Aが与えたとしよう。また、それぞれに「3点」・「5点」を回答者Bが与えたとしよう。回答者A・Bのいずれもが、警察より病院を2ポイント高く信頼しているという判断は、このデータから可能である。

ところが、同一評価項目に対する評価点を回答者A・B間では比較ができないことになる。先の例では、警察に対する評価が、回答者Aの方がBよりも2ポイント高いとは必ずしも言えない。回答者Aと回答者Bの信頼評価が、絶対的にはまったく同じであったとしても、第1表示項目によるアンカリングにより、それぞれの「評価基準点」が異っている可能性があるからである。同様の理由で、全評価項目に対する評価の平均値を回答者A・B間で比較することもできなくなる。

そして、これは、同一の回答者に、同じ質問を繰り返すパネル調査についても、重大な意味合いを持つことになる。ある回答者は、2つの調査時点間で、制度に対する評価はまったく変わっていないとしても、たまたま評価項目の第一位に表示された項目が、第1波と第2波で非好意的なものから好意的なものに変わったことで、制度全体に対する評価が好転することになる。もちろん、その逆も起こりえる。これでは、「信頼のおける(reliableな)」データとは言えない。実証的な検討が求められるところである。

IV. データ

上記の問題点を具体的に検討するために、本稿では、Web調査を利用した。Web調査は、通常の無作為抽出による世論調査(通常調査)とは異なり、対象者に代表性はない。したがって、その調査データをもとにした分析から「一般論」を引き出すことはできない。ただし、「コンテキスト」の違いについての効果を測定しようとする当該研究のような場合、対象者の「代表性」を犠牲にしても致命的な欠陥とはならないだろう。しかも、1回当たり数千万円の費用がかかる通常調査と比較して格安の経費でデータを得ることができる。経費を抑えることで、限られた予算でも複数回の調査を実施することができる。コンテキストを少しずつ変化させながら、その効果を比較することが可能となる。

表1のとおり、3回のWeb調査（「基本パネル調査」・「尺度調整調査」・「先頭固定調査」）を実施した。また、調査ごとにサンプルを分ける「スプリット調査」とした（スプリット数は、順に、2・3・3）。したがって、8つのサンプルを得ることになった（表1の[A・B]・[C・D・E]・[F・G・H]にそれぞれ対応）。さらに、「基本パネル調査」では、1週間の時差を設けて同じ回答者に回答を求めるパネル方式を採用した²⁾。これらの合計10の調査を基に、以下、検討を進めていく。

表1 信頼度評価の平均値と評価の幅、サンプル間比較

サンプル	基本パネル調査				尺度調整調査				先頭固定調査 裁判所・国の政治			
	A		B		C	D	E		F	G	H	
	第1波 R*なし	第2波 Rなし	第1波 Rなし	第2波 Rあり	3組で Rあり	任意の 最高・ 最低	最高 10 / 最低 0	16 項目	裁・国 固定	裁・国 Rあり	裁・国 Rあり 2/3位	
範囲	平均値	5.62	5.46	5.71	5.29	5.96	5.98	10	5.68	5.88	5.54	5.66
	最小値	0	0	0	0	0	0	10	0	0	0	0
	最大値	10	10	10	10	10	10	10	9	10	10	10
平均点	平均値	4.68	4.71	4.72	4.57	4.53	3.99	4.37	4.29	4.50	4.47	4.45
	最小値	0.00	0.00	0.17	0.06	0.00	0.00	0.56	0.00	0.00	0.44	0.00
	最大値	8.61	8.89	8.61	8.39	7.83	7.50	8.33	8.75	7.78	8.28	8.00
	N	392		769		490	446	483		525	516	519

* [R] はランダムマイゼーションを表す

V. 分析結果 1：認知心理的な傾向性の確認

まず、「基本パネル調査」を用いて、世論調査の質問に答えようとする回答者の「癖」について、順にデータを見ていきたい。

V.1 尺度の部分利用

調査票を設計する際、「評価尺度の幅を最大限に活用しながら、評価対象のそれぞれを回答者には評価してほしい」と設計する側は考える。評価の分散が大きいほど、説明力が上がるからである。とりわけ、ここで取り上げている「制度信頼」などの質問では、評価対象の項目数が一般的に多いのでなおさらである。たとえば、評価項目が18もあるのに、5点程度の狭い尺度で評価を求めたとしても、多くの項目につき同じ点数で評価されることとなり、項目間の微妙な評価差を測定できないことになる。

では、回答者は、調査設計者側の意図どおりに、たとえば「11点尺度」の11点をフルに「活用」しながら回答をしているのだろうか。どうもそうではないようである。

表1の「範囲 平均値」の行を見ていただきたい。18の項目（具体的な評価項目については、表2参照）について評価を求めた当該調査の場合、回答者単位で見たときの最高点と最低点の「差（幅）」は、いずれも6点にも達しない。つまり、ある回答者の平均的な信頼評価が仮に4点であったとすると、その回答者は、1点から7点の間で18項目を評価していることになる。そして、0

点、あるいは8点以上の両端の評価は使わなかったことになる。10回のすべての調査で、その傾向は一貫している³⁾。つまり、11点もの幅を設けたとしても、その幅の全部を使う人はそれほど多くはない。「評価尺度の部分利用」が一般的であることがわかる。

V.2 相対評価の傾向（あるいは、順序効果）

では、有権者は、用意された尺度を、絶対的にとらえて各項目を評価しているのだろうか。それとも、直前の項目との相対比較で評価をするものなのだろうか。そもそも、「順序効果」が問題となるのは、後者（相対評価）の傾向があるからである。ここでは、18項目に対する信頼評価点を因子分析することで、相対評価の傾向性を確認したい。

表2をご覧いただきたい。5組の因子分析の結果が報告されている。表では、評価項目と因子との関係を視覚的にとらえやすくするために、因子負荷量が0.5を超える場合を便宜的に網かけにしている。

因子分析1・2は、サンプルAの第1・2波に、因子分析3・4は、サンプルBの第1・2波に対応している。なお、サンプルAでは、同じ質問が、表示順も固定されたままで両波で聞かれている。また、サンプルBの第1波（因子分析3）は、サンプルAの第1波と同時に、また、同じ表示順で質問をしている。そして、サンプルBの第2波（因子分析4）では、評価項目がランダムマイズされたうえ提示されている（図4参照）。

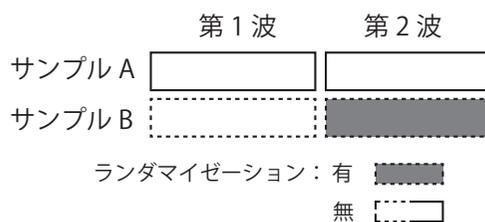


図4：サンプルA・Bとランダムマイゼーション

そして、順序効果（相対評価）がもっとも明確に見て取れるのが因子分析5である。因子分析5は、回答者ごとに「第1番目に表示された項目に対する信頼度」・「第2番目に表示された項目に対する信頼度」・「第3番目に表示された項目に対する信頼度」の順に、それぞれの評価点の最後までを因子分析したものである。当然のことながら、「第1番目に表示された項目」は回答者によって異なる。「第2番目に表示された項目」も同様である。要するに、評価項目の具体的な内容は無視をして、評価点のみ回答順に見たものを因子分析したわけである。内容を無視しているのであるから、本来なら、因子構造が検出されるはずがない。

ところが、表2（因子分析5）にあるとおり、第1表示項目から第7表示項目までが第2因子を構成し、第8表示項目を折り返し点として、残りが第1因子を構成している。表示内容としてはまったくランダムな項目同士が、その前後で因子を構成するのは、まさしく「順序効果」によるものであるとしか考えられない。ある項目に対する評価が、その直前の項目への評価点、

あるいはそのまた1つ前の項目への評価点を基準にされているから、それらがあたかも関連があるように（つまり、因子を構成しているかのように）「因子分析」には映るのである。

じつは、ランダムマイゼーションをしていない因子分析1から3についても、順序効果（相対評価）の傾向が見て取れる。いずれをとってみても、網掛けの部分が、項目リストの上から順に、第3因子・第2因子・第1因子と整然と移動している。つまり、第3因子と関連の見られる最初の2項目について、よく似た傾向の評価がされ、続いて、今度は第2因子と関連の見られる「警察」から「裁判所」が、また、近い評価がされる。そして、最後に第1因子と関連が見られる残りの項目について、また、同じような評価がされているわけである。

もちろん、「順序効果というより、実質的な評価として因子構造が表れている」との反論も可能である。そもそも、この質問を設計した側が、マスメディア・社会的な組織・政治的な組織（「日本の経済体制」は除く）の3グループを評価対象群として想定しているわけである。そして回答者もまた、そのような3グループに対して大まかな評価をまず行い、その内部でさらに細かな評価をしているとしたらどうだろうか。評価項目リストが、マスメディア・社会的な組織・政治的な組織の順に用意されているのだから、表2にあるような分析結果となるのはむしろ自然である。

実際、因子分析4（ランダムマイゼーションあり）が、順序効果のみによる見せかけの因子構造ではないことを裏付けている。じつは、因子分析4の網掛けの部分を見るかぎりでは、その因子構造は因子分析1から3とはずいぶん異なるように見える。ところが、因子分析4の第1因子と第2因子を入れ替えてみると（そして、第一因子の「労働組合」・「中央官庁」・「市区町村役場」・「あなたの住んでいる地域の政治」の網掛けの有無をひとまず無視しするとしたら）、因子分析4が、因子分析1から3ときわめて近い結果を示していることがわかる。因子分析4でも、やはり「テレビ」・「新聞」が1つの因子を構成し、「警察」から「裁判所」がもう1つの因子を、「国会」から「国の政治」までが最後の因子を構成している。これは、因子分析1から3の分析結果と基本的に同じである。繰り返すが、因子分析4では順序効果は起こりえない。にもかかわらず、それと同じ構造を示しているのであるから、因子分析1から3が「順序効果のみの反映である」との可能性は明らかに排除される。

これらの一連の因子分析の結果から、次のようなことが言えそうである。まず、因子分析5により、「順序効果」が確実に存在することがわかった。では、ランダムマイゼーションをしない場合に見られる因子構造は、順序効果のみによる「見せかけ」の因子の検出かということ、必ずしもそうではない。順序効果を排除した場合（因子分析4）と同じ因子構造がランダムマイゼーションなしでも検出されることから、そのことが確認された。回答者は、順序効果とは独立に、ある種の構造を持ちながら、それぞれの項目を認知しているわけである。

その一方で、「順序効果」が確認されたということは、とりもなおさず、絶対評価ではなく、相対評価を回答者がしていることを意味する。前の項目との相対的な比較で、当該の項目を評価しているという事実は、次節で説明するとおり、評価バイアスの観点からは重要な意味をもつ。

V.3 評価バイアスの存在

順序効果への対応策として有効なランダムマイゼーションではあるが、それが新たなバイアスを導入することはないのだろうか。システムティックな問題がもしもあるとしたら、その導入は慎重でなければならないし、少なくとも、(私がインタビューした調査会社のプログラマーが証言するように)「自動的に適応します」という「慣行」には再検討が必要である。

とりわけ、ここで注目するのは「評価の安定性」である。ランダムマイゼーションすることで、そうでない場合に比べて「評価が変わることがないか」という問題である。そこで、表3をご覧いただきたい。それは、「基本パネル調査」の2つのサンプル(A・B)のそれぞれで、第1波と第2波の2つの調査において、18の評価項目についての「サンプルとしての評価の平均点」がどの程度変化したかを確認したものである。「サンプルとしての評価の平均点」とは、たとえば「テレビ」の評価点についての、回答者全体の平均値を指している。それが、サンプルにより変わるのか変わらないのかを見ていきたい。なお、図4ですでに説明したとおり、サンプルAは、両波ともランダムマイゼーションはなく、表3の左端に表示された順序で固定した順に回答をしてもらっている。サンプルBについては、第1波は、サンプルAと同じように固定リストで質問をしているが、第2波では、ランダムマイズされた順で評価されている。第1波と第2波の時差はわずか1週間であったので、その間に信頼評価が大きく変わったとは思えない。

表3 信頼評価(平均値)の変化、(基本パネル調査)

	サンプル A				サンプル B			
	第1波 (Rなし)	第2波 (Rなし)	第2波 - 第1波		第1波 (Rなし)	第2波 (Rあり)	第2波 - 第1波	
			t値	危険率			t値	危険率
#1 テレビ	5.43	5.24	-2.953	0.003	5.37	4.70	-12.923	0.000
#2 新聞	6.33	6.23	-1.568	0.118	6.36	5.53	-14.647	0.000
#3 警察	5.91	5.73	-2.640	0.009	5.89	5.20	-12.087	0.000
#4 病院	6.31	6.20	-1.812	0.071	6.29	5.92	-6.790	0.000
#5 公立小中学校	5.51	5.42	-1.310	0.191	5.54	5.14	-7.551	0.000
#6 労働組合	4.67	4.57	-1.576	0.116	4.79	4.28	-8.762	0.000
#7 大企業	5.22	5.16	-0.910	0.364	5.17	5.00	-3.080	0.002
#8 銀行	5.42	5.37	-0.627	0.531	5.42	5.36	-0.906	0.365
#9 裁判所	6.58	6.65	0.958	0.339	6.64	6.07	-10.851	0.000
#10 国会	3.65	3.68	0.328	0.743	3.83	3.40	-7.113	0.000
#11 政党	2.96	3.07	1.709	0.088	3.06	3.17	1.762	0.079
#12 中央官庁	3.79	3.93	1.987	0.048	3.76	3.95	3.247	0.001
#13 市区町村役場	4.73	4.93	2.726	0.007	4.72	5.11	6.861	0.000
#14 年金制度	2.90	2.99	1.416	0.158	2.93	2.86	-1.407	0.160
#15 選挙制度	4.17	4.34	2.027	0.043	4.31	4.71	5.887	0.000
#16 日本の経済体制	3.71	3.81	1.330	0.184	3.72	3.80	1.423	0.155
#17 国の政治	2.87	3.08	3.148	0.002	2.98	3.39	7.392	0.000
#18 地域の政治	4.09	4.33	3.232	0.001	4.23	4.74	9.090	0.000

表によると、同じ回答者が、同じ順に表示された評価リストに対して回答したとしても、完全に同じ評価となるのは難しいようである。表3のサンプルAの「第2波 - 第1波」のコラム

をご覧いただきたい。18項目中、7項目において、第1波と第2波の平均値につき、統計的に有意な差（危険率5%）が確認された。しかしながら、第2波がランダムマイズされているサンプルBではさらに安定度が損なわれており、14項目について評価が変わっている。測度の安定度の観点から、ランダムマイゼーションの「危うさ」がまずわかる。また、次の表4では、項目ごとに、

表4 信頼評価の変化（悪化・変化なし・改善）、（基本パネル調査）

		グループ A	グループ B
#1 テレビ	悪化	10.2	25.6
	変化なし	81.4	69.4
	改善	8.4	4.9
#2 新聞	悪化	7.9	28.1
	変化なし	84.9	67.0
	改善	7.1	4.9
#3 警察	悪化	11.2	29.9
	変化なし	81.6	63.1
	改善	7.1	7.0
#4 病院	悪化	12.0	19.9
	変化なし	79.1	71.4
	改善	8.9	8.7
#5 公立小中学校	悪化	12.0	21.2
	変化なし	78.6	70.9
	改善	9.4	7.9
#6 労働組合	悪化	10.7	23.9
	変化なし	80.9	67.5
	改善	8.4	8.6
#7 大企業	悪化	11.7	16.9
	変化なし	77.3	71.4
	改善	11.0	11.7
#8 銀行	悪化	14.0	15.7
	変化なし	74.5	70.2
	改善	11.5	14.0
#9 裁判所	悪化	8.4	22.5
	変化なし	79.3	71.3
	改善	12.2	6.2
#10 国会	悪化	12.5	23.9
	変化なし	75.3	65.1
	改善	12.2	10.9
#11 政党	悪化	10.5	12.7
	変化なし	78.1	70.4
	改善	11.5	16.9
#12 中央官庁	悪化	9.2	12.4
	変化なし	78.3	69.2
	改善	12.5	18.5
#13 市区町村役場	悪化	9.2	8.1
	変化なし	76.0	70.6
	改善	14.8	21.3
#14 年金制度	悪化	8.9	13.5
	変化なし	77.3	74.8
	改善	13.8	11.7
#15 選挙制度	悪化	13.5	12.1
	変化なし	68.9	62.4
	改善	17.6	25.5
#16 日本の経済体制	悪化	11.5	13.8
	変化なし	76.8	67.5
	改善	11.7	18.7
#17 国の政治	悪化	7.4	7.9
	変化なし	80.4	70.9
	改善	12.2	21.2
#18 地域の政治	悪化	8.7	7.4
	変化なし	75.8	69.7
	改善	15.6	22.9

評点が変わらなかったか（あるいは変わってもプラス／マイナス1ポイントに収まったもの：「変化なし」）、2ポイント以上「信頼評価」が下がったもの（「悪化」）、2ポイント以上「評価」が上がったもの（「改善」）の割合を、A・Bサンプルで比較した。項目ごとにバラツキはあるものの、「変化なし（つまり「安定的」でバイアスのない回答）」が、サンプルBでは5から10ポイント一貫してAに比べて減少している。ランダマイゼーションの「危うさ」がここでも露呈された。データの信頼性（reliability）の観点から、見逃すことのできない結果である。

V.4 アンカリング

では、その「不安定さ」の原因はどこにあるのだろうか。

前2節で確認をした2つの認知心理的な「癖」から想起される原因の1つは、「第1表示項目によるアンカリング」である。そもそも、評価尺度の全幅を利用した評価をする回答者は多くない。そこに、先の評価項目に対する評価点を基準に次の項目を評価するという相対的な評価の仕方が重なると、最初に表示させる項目が重要な意味を持つことになる。

実際、サンプルBの第2波について、第一評価項目に対する評価点が「4以下」・「5」・「6以上」の3つのグループにわけ、それぞれのグループの信頼評価の平均値を比較してみた。順に、4.0・4.5・5.3で、信頼度が上がっている⁴⁾。アンカリングの影響を示すデータである。

そこで、サンプルBにおいて「第1波・第2波のズレの総量」に対する、第2波での第1表示項目の「表示バイアス」の影響を重回帰分析を用いて検証してみよう。

表5は、18項目すべてについての「ズレの総量」を従属変数として、それに影響を及ぼすと考えられる要因を独立変数に設定をした回帰分析の推定結果である。ここでの「ズレの総量」とは、項目ごとの第1波と第2波との評価点の差を、18項目すべてについて加算をした値である⁵⁾。

表5 「両波における評価差の総量」説明モデル（回帰分析、サンプルB）

	標準化係数	t 値	危険率
表示バイアス（第2波の第1項目の両波評価差）	0.414	11.366	0.000
第1波の第1項目（テレビ）評価点	-0.113	-3.251	0.001
第2波の第1項目評価点（本人）	0.029	0.635	0.526
第2波の第1項目評価点（平均値）	0.098	2.339	0.020
第2波の第1項目評価点（5からの距離）	0.018	0.525	0.600
政治関心	0.025	0.774	0.439
政治満足	0.105	3.147	0.002
性別*			
年齢*			
教育程度*			
所得*			
定数	-	-2.661	0.008
調整済み R ²		0.219	
N		769	

*最終的なモデルからは除外

ここでキーとなる説明変数は、「表示バイアス（第2波の第1項目についての両波評価差）」である。これは、第2波の評価項目リストの先頭に表示された項目についての「第1波から第2波の間での評価差」で定義している。たとえば、Aさん・Bさんの二人について、第2波では、それぞれ「国の政治」と「裁判所」から質問がスタートしたとする。そこで、Aさんについては、「国の政治」の第1波と第2波の評価差（第2波評価点 - 第1波評価点）を「表示バイアス」スコアとする。Bさんについては、「裁判所」について同様の計算をするというわけである⁶⁾。

表5にあるとおり、「表示バイアス」スコアが、圧倒的に大きい標準化係数を示している。ランダムマイズの結果として生じた「先頭項目における評価差」が「ズレの総量」に及ぼす影響の大きいことがわかる。

なお、その他のコントロール変数（「第1波の第1項目（テレビ）評価点」から「第2波の第1項目評価点（5からの距離）」まで）は、第1波の先頭項目として固定されている「テレビ」に対する評価や、第2波の第1項目（こちらは、回答者により異なる。先の例では、Aさんの「国の政治」・Bさんの「裁判所」）などへの評価を含めることで、「表示バイアス」の純粋な影響を測定しようとしたものである⁷⁾。

また、政治関心や政治満足度も、評価の安定度に寄与する可能性があるので説明変数に投入したが、前者は統計的な影響が認められなかった。政治関心の高い人の方が、絶対的な評価基準を持ち合わせる確率が高く、その分だけ評価も安定するかと予測したが、そうとは限らないようである。一方、政治満足度の高い人の方が「ズレの総量」が大きい。これも解釈が難しいが、政治に対して不満な人の方が政治的アクターに対する信頼度をつねに厳しい視線で見ているために、より安定的に（そして、おそらく低い評価を）一貫して行っているのかもしれない⁸⁾。

いずれにしても、上記の分析から、「信頼性（reliability）」を損ねるという点で、「表示バイアス」がけっして無視できないものであることが判明した。

VI. 「表示バイアス」への対処の試み

では、それに対する対処法はあるのか。本稿では、「尺度調整方式」と「先頭項目固定方式」の2つの方法を検討した（表1参照）。ところが、これらの書式変更では、期待した効果は得られなかった。この節では、それぞれの「対処方法」の意図と、期待した効果が得られなかった理由についての推論を紹介したい。

サンプルD・Eについては、次のような質問構成で制度信頼を尋ねた。

- 1) 図2と同じ画面で18の評価項目を提示した上で、a)「もっとも信頼できる」項目と、b)「もっとも信頼できない」項目をあらかじめ選択をさせる（図5）。
- 2) それらの項目について、サンプルDでは、既存の11点尺度で評価をさせる（図6）。
- 2) サンプルEでは、「もっとも信頼できる」項目については10点の評価点を、「もっとも信頼できない」項目については0点の評価点を与えるようにと回答者に指示をする（図7）。
- 3) 続く質問で、残りの16の項目について、同じように11点尺度で評価をさせる。なお、これ

Cyber Panel

社会的な組織や団体についてのアンケート（本調査）

いくつかの組織・団体・制度についてお伺いします

B1 ここにあげるような組織や団体や制度を、どの程度、あなたは信頼していますか、(1)もっとも信頼できるものと、(2)もっとも信頼できないものを、それぞれ1つずつお選びください。

(1)もっとも信頼できるもの (1つだけ)	(2)もっとも信頼できないもの (1つだけ)
<input type="radio"/> 1. テレビ	<input type="radio"/> 1. テレビ
<input type="radio"/> 2. 病院	<input type="radio"/> 2. 病院
<input type="radio"/> 3. 大企業	<input type="radio"/> 3. 大企業
<input type="radio"/> 4. 新聞	<input type="radio"/> 4. 新聞
<input type="radio"/> 5. 公立の小中学校	<input type="radio"/> 5. 公立の小中学校
<input type="radio"/> 6. 銀行	<input type="radio"/> 6. 銀行
<input type="radio"/> 7. 労働組合	<input type="radio"/> 7. 労働組合
<input type="radio"/> 8. 警察	<input type="radio"/> 8. 警察
<input type="radio"/> 9. 中央官庁	<input type="radio"/> 9. 中央官庁
<input type="radio"/> 10. 政党	<input type="radio"/> 10. 政党
<input type="radio"/> 11. 市区町村の役所	<input type="radio"/> 11. 市区町村の役所
<input type="radio"/> 12. 国会	<input type="radio"/> 12. 国会
<input type="radio"/> 13. 裁判所	<input type="radio"/> 13. 裁判所
<input type="radio"/> 14. 国の政治	<input type="radio"/> 14. 国の政治
<input type="radio"/> 15. 年金制度	<input type="radio"/> 15. 年金制度
<input type="radio"/> 16. あなたの住んでいる地域の政治	<input type="radio"/> 16. あなたの住んでいる地域の政治
<input type="radio"/> 17. 日本の経済体制	<input type="radio"/> 17. 日本の経済体制
<input type="radio"/> 18. 選挙制度	<input type="radio"/> 18. 選挙制度
<input type="radio"/> 19. 「1つだけ」は選べない	<input type="radio"/> 19. 「1つだけ」は選べない
<input type="radio"/> 20. あてはまるものはない	<input type="radio"/> 20. あてはまるものはない

NR 日本リサーチセンター
Member of Gallup International

次へ

図5：「尺度調整方式」 - 「最大・最少」選択画面

らの16項目については、ランダムマイゼーションを行った。

つまり、ここでの試みは、最高点と最低点をあらかじめ「固定」した上で、その範囲内で残りの項目を評価されるという形式で、それを仮に「尺度調整方式」と呼ぶことにする。「もっとも信頼できる」項目と、「もっとも信頼できない」項目で、使用する認知上の尺度の「両端（あるいは、「幅」）」をアンカリング（固定）することで、回答の分散を大きくすることを狙った。そのことでランダムマイゼーションの第1項目によるアンカリングの効果を相対的に抑制できると考えたからである。なお、サンプルDとサンプルEの違いは、前者は、回答者自らが持つ認知上の評価幅で相対化しようとしたのに対して、後者では、尺度の最大幅である「両端（0と10）」を基準に相対化させようとした⁹⁾。

サンプルFからHは、最初に表示する2項目をあらかじめ調査設計者の方で決めておくことで、

Cyber Panel

社会的な組織や団体についてのアンケート（本調査）

B2では、「信頼できない=0点」から「信頼できる=10点」の10点満点で「信頼度」を評価するとしたら、それぞれ何点くらいになりますか。

(1) 「もっとも信頼できない」としてお選びになった「テレビ」と
 (2) 「もっとも信頼できる」としてお選びになった「警察」
 について、点数でお答えください。（それぞれ1つだけ）

	信頼 できない	0	1	2	3	4	5	6	7	8	9	10	信頼 できる
(1) テレビ	<input type="radio"/>												
(2) 警察	<input type="radio"/>												

NK 日本リサーチセンター
Member of Gallup International

次へ

図6：「尺度調整方式」 - 最大値・最小値測定画面

Cyber Panel

社会的な組織や団体についてのアンケート（本調査）

C2では、「信頼できない=0点」から「信頼できる=10点」の10点満点で「信頼度」を評価するとします。そして、今、

- ・「もっとも信頼できない」としてお選びになった「日本の経済体制」を「0」点
- ・「もっとも信頼できる」としてお選びになった「あなたの住んでいる地域の政治」を「10」点

と、評価していただいたと仮定してください。
 そのうえで、次にあげる組織や団体や制度を、同じように0から10の点数で評価するとしたら、何点くらいになりますか。
 画面右下の「次へ」をクリックしてください。

NK 日本リサーチセンター
Member of Gallup International

次へ

図7：「尺度調整方式」 - 最大値・最小値の10点固定画面

回答者全員に対して同一のアンカリングをしようと狙うものであった。図8を参照していただきたい。サンプルFでは、これまでの調査において一般的に信頼度が高いとされてきた「裁判所」と、信頼度が低いとされてきた「国の政治」の2項目を、第1項目・第2項目として「固定」し、残り16項目をその後にランダムマイズして表示した。サンプルGは、サンプルFと同じように、「裁判所」・「国の政治」で固定するが、「裁判所」・「国の政治」のいずれが先になるかはランダムに決めた。残りの16項目については、同じくランダムマイズする。最後のサンプルHでは、「裁判所」・「国

	サンプル		
	F	G	H
第1表示項目	「裁判所」	「裁判所」・ 「国の政治」で ランダムイズ	
第2表示項目	「国の政治」		「裁判所」・ 「国の政治」で ランダムイズ
第3表示項目	残りの16項目を ランダムイズ	残りの16項目を ランダムイズ	第1表示位置と併 せて、残りの16 項目をランダム イズ
第4表示項目			
...			
...			
第18表示項目			

図8：「先頭項目固定方式」-「裁判所」・「国の政治」の固定位置とランダムイゼーション

の政治」の2項目についてのランダムイズを、サンプルGより1項目だけ後にずらした位置で実施をした。その上で、第1表示項目は、残りの16項目からランダムに選択する。つまり、任意の1項目と「裁判所」・「国の政治」の計3項目で、先頭のアンカリングをしようとするものであった。残る15項目については、やはりランダムイズをしている。

これらを「先頭項目固定方式」と呼ぶことにする。アンカリングの効果を相対化させようとする点では「尺度調整方式」とその狙いは同じであるが、回答者全員について同じ項目で相対化し、回答者間での比較を可能とした点が「先頭項目固定方式」の特徴である。

VII. 分析結果2：質問形式変更の効果

さて、これらの2組の質問形式の変更が、制度信頼についての回答パターンに変化をもたらしただろうか。結論から述べるなら、いずれの方式も、期待した効果が確認できなかった。

「尺度調整方式」の狙いは、「もっとも信頼できる」項目と、「もっとも信頼できない」項目でアンカリングをすることで、回答の分散を大きくし、そのことでランダムイゼーションの第1項目によるアンカリングの効果を相対的に抑制しようとするのであった。そして、表1の「範囲 平均値」にあるように、確かに「尺度調整調査 (D)」では5.98と、「基本パネル調査」のそれら(5.62・5.46・5.71・5.29)より評価幅が大きくなっている。一定の効果が認められる。

なお、サンプルEは、最高点を10、最低点を0に「強制」しているわけであるから、「範囲 平均値」が10となるのは当然である。そこで、残りの16項目への「効果」を確認する必要がある。

サンプルEの「16項目」のコラムに注目していただきたい。これは、「もっとも信頼できる」項目と、「もっとも信頼できない」項目を選ばせた後の、残る16項目についてのみの評価幅を示しているが、それは5.68と、Dよりも逆に小さくなっている。つまり、回答者が用いる認知

上の尺度幅を調査主体の方でコントロールしようとしても、それは難しいことを意味している。じつは、「もっとも信頼できる／できない」項目を選ばせた質問と、残りの16項目についての質問とは、回答者にとっては「別の質問」だったのではないだろうか。調査側の方で、「最小が0、最大が10」の尺度を想定くださいと依頼したところで、後続の質問においてその仮定の尺度を回答者が実際に利用する保証はない。アンカリングは、同一の質問内でその効果を持続するとしても、次の質問に移動するや「リセット」されると考えるのが妥当なようである。

「先頭項目固定方式」であるが、これは、先頭項目が回答者間で変わらないようにすることで、サンプル間の平均値の標準化を図ろうとしたものであった。表1の「平均点 平均値（サンプル F・G・H）」にあるように、それらには確かにほとんど差がない。したがって、当初の狙いは達成したかに見える。ところが、別の分析から、F・G・Hいずれのサンプルにおいても、「裁判所」以外の項目を高く評価する回答者が半数程度あったことが判明している。また、「国の政治」以外の項目を低く評価した回答者がいずれのサンプルでも80%も存在することが分かった。「裁判所—国の政治」が、もっとも評価幅を大きくする2項目であると設計者としては想定していたわけであるが、多くの回答者にはそれが当てはまらなかった。つまり、成功したかに見えた平均値の標準化は、評価項目の固定の結果では必ずしもなかったことになる。

VIII. 結論

上記の分析から導き出される結論は次のとおりである。

・ランダムマイゼーションによる「表示バイアス」は、無視することができない

最初に表示された項目に対する評価によって、その他の項目に対する評価がアンカリングされることが確認された。つまり、好意的に（信頼できると）感じている項目から質問がスタートする場合と、非好意的に（信頼できないと）感じている項目から質問がスタートするのでは、同じ回答者の評価であっても、その平均値が異なることになる。測定値の平均値が、第1評価項目として何が表示されるかという「偶然」により変わることは、尺度の信頼性（reliability）を損ねることになり、これは深刻な問題であると言わざるをえない。

・評価項目順についてのランダムマイゼーションにより、回答者個人の単位でのランダムな「順序効果」を導入していることにも留意が必要

評価項目のランダムマイゼーションにより、固定リストに固有な「順序効果」を排除することができる。そのことにより、母集団のパラメーターを集計値として推定する場合には、ランダムマイゼーションが有効である。

ところが、ランダムマイゼーションを導入しても、個人単位での順序効果が排除されるものではない。複数の項目に対して評価を求めるとき、回答者は常に相対評価をするのであるから、どの順に提示したとしても順序効果から逃れられるものではない。それどころか、「固定リスト（ランダムマイゼーションなし）」であれば、1組の順序効果パターンの影響を回

答者全員が同様に受けているのに対して、評価項目のランダムマイゼーションの場合は、評価項目の階乗 (N!) 分存在する多数の「順序効果」パターンをランダムに個人に対して導入していることになる¹⁰⁾。この事実は、これまであまり注目をされていない。

・変数間の関連性を検討しようとするような場合は、ランダムマイゼーションは避けるべきである

集計データとして調査結果を用いる場合は、「順序効果」を相殺することができるのでランダムマイゼーションは有効であるが、その変数と他の変数との関連性を回答者単位で確認しようとするような場合には、ランダムマイゼーションは勧められない。

たとえば、「政治制度や社会の組織を信頼する傾向にある人が、政治システム一般もサポートする傾向がある」というような命題を確認しようするときには、ある回答者が「高信頼者」なのか「低信頼者」なのかが重要な指標となる。ところが、「高信頼者」・「低信頼者」の判定が、ランダムに表示される「第1項目」によって決まるとしたら、これは重要な問題である。

そのような事例では、順序効果の影響がたとえあるとしても、固定リストの方が優位であろう。少なくとも、回答者全員が同じ項目（固定リストの最初の項目）によりアンカリングされているので、その意味において、回答者間の比較が可能となる。仮にも、評価項目のランダムマイゼーションがなされたデータセットを用いて、上記のような変数間の関連性を分析をする場合は、第1表示項目（回答者ごとに異なることに注意）に対する評価点でコントロールをすることが最低でも必要だろう。

・パネルデータの場合は、回答者単位での表示順は、すべての波の調査で固定するのが望ましい

仮に、ランダムマイゼーションの導入が優位であると判断された場合でも、それがパネルデータの場合は、少なくとも、回答者単位での表示順をすべての波の調査で固定するべきだろう。パネル毎に、第1表示項目が変わったとしたら、個人単位のパネル間の測定値の「変動」が、真に個人の態度変容によるものか、第1表示項目によるアンカリングによるものかの判断がつかない。第一波での表示順を記録しておき、それと同じ順序のリストを後続の調査でも用いることにより、この問題を排除することができる。

・アンカリングの弊害を、質問文の作り方で解消するのは容易ではない

「尺度調整方式」や「先頭項目固定方式」など、質問文の工夫でアンカリングの影響を軽減しようと試みたが、少なくともここではそれらの効果は確認できなかった。

「尺度の部分利用」の傾向がアンカリングの1つの誘因となっていることに鑑み、いずれの試みも、評価尺度を最大限に利用させることを狙った。11点尺度をフルに利用すれば、第1項目によるアンカリングの影響が軽減されると考えたからであった。そして、「尺度調整方式」は、回答者一人ひとりにとっての最高評価点と最低評価点を回答者に意識をさせることで、尺度の幅を操作しようとした。一方の「先頭項目固定方式」は、回答者全体として好意的に評価されている項目と、非好意的にしか評価されていない項目をリストの先頭に提示することで、認知上の尺度の幅を操作しようとした。

ところが、前者の場合、最高評価点と最低評価点を聞き出す質問群に加えて、残りの項目に対する評価点を聞き出す質問本体との少なくとも2組の質問が必要となる¹¹⁾。ところが、

質問間を移動するうちに、回答者の認知上の尺度の幅はリセットされてしまうようで、「尺度の最大利用」への誘導は成功しなかった。また、後者の場合も、「平均的な回答者」が最高点（あるいは、最低点）を与える評価項目でも、それが「個々の回答者」にとって最高評価項目（あるいは、最低評価項目）である保証がないことから、やはり「尺度の最大利用」へ誘導できなかった¹²⁾。

つまるところ、評価項目のランダムマイゼーションを実施するかどうかは、その質問群を調査票に含める際の「調査目的」によって個別に判断がされるべきだろう。少なくとも、技術的に可能であるから、あるいは、順序効果を必ず排除するべきだからということで、「自動的に」評価項目のランダムマイゼーションを導入する「慣行」があるとしたら、そのことには注意を喚起したい。

順序効果による弊害と、アンカリングによる弊害はトレード・オフの関係にある。したがって、当該の調査の目的に照らして、どちらの弊害がより深刻かを考慮して評価項目のランダムマイゼーションの導入を判断することになる。新しい概念について、その背後にある認知構造を探索する試みのような場合は、順序効果をまずは取り除く必要があるだろう。一方、その質問群と他の変数群の関連性を因果論的に検討することが主たる目的であれば、アンカリングの弊害の方が深刻であろう。

ここで取り上げた「制度信頼」に関する質問群などは、すでに政治学・社会学関係の各種の世論調査で質問が繰り返されている。その点では、因子構造についての理解などはある程度すでに進んでいると考えられる。さらに、その他の政治学・社会学的な要因と「制度信頼」の関係性や、回答者単位の「政治信頼」の時系列的な変化に関心があるのであれば、順序効果を一定に保っておくことがむしろ賢明であろう。具体的には、調査間で、あるいは回答者間で評価項目の順番を変えない、つまり、ランダムマイゼーションはしないという判断は十分にありえる。

世論調査への CASI の導入により、調査手法上の新しい「工夫」が可能となってきた。最近では、実験室実験の手法を世論調査に応用する試みなどもなされている。世論調査の「可能性」がますます高まっている（Mutz 2011）。その一方で、新しい「技術」が新しいバイアスを生んでいる可能性にも、私たちは自覚的でなくてはならないだろう。本稿では、評価項目のランダムマイゼーションの是非について検討をしてきたが、課題の多いことが判明した。そもそも、ここでの知見は、Web 調査に登録をしているモニターに対する回答をもとにしており、「一般化」が難しいことはすでに指摘したとおりである。ただし、1つの「サブグループ」においても問題が確認されたのであるから、全有権者からの無作為抽出による対象者についても、その利用は慎重であることが求められる。いずれにしても、ランダムマイゼーションを含めて、その他の多くの「工夫」に対する検証の必要性を喚起する1つの契機にこの小論がなれば幸いである。

注

本稿は、科学研究費補助金（研究期間：2008年-2010年、課題番号：20530123）による研究成果の1つである。

- 1) 世論調査会社（東京都内）のプログラマーに対するインタビュー、2008年12月26日。
- 2) いずれの調査も日本リサーチセンター（NRC）に業務委託をした。回答者は、首都圏（東京都・神奈川県・千葉県・埼玉県）在住で、NRCにモニターとしてあらかじめ登録をした20歳以上60歳未満の男女。
- 3) 後に説明をするが、サンプルEについては、最初の2項目を除く「16項目」についての幅を参照。
- 4) それぞれ該当数は、281（38%）・145（20%）・312（42%）。

- 5) $\sum_{i=1}^{18} (trust2_i - trust1_i)$ で定義される「差」の総量である。なお、*trust1* は、ある評価項目 (*i*) についての第1波の評価点（信頼度）を、*trust2* は、同じその項目についての第2波の評価点を表している。また、「差」については絶対値を取っていない。第1表示項目の評価点により、評価項目の全体として、上方、あるいは下方のどちらにバイアスが発生するかを確認するのが目的であるからである。

詳細は、私のホームページ (<http://ynishiza.doshisha.ac.jp>) の「Please Download」で公開をしているSPSSのシンタックスを参照いただきたい。

- 6) 数式で表すとすれば、*trust2displayedfirst* - *trust1correspond* となる。なお、*trust2displayedfirst* は、第2波でランダムに表示された評価項目に対する評価点（信頼度）で、*trust1correspond* は、前者に対応する項目の第1波における評価点を表す。
- 7) 「第2波の第1評価項目評価点（平均値）」には、第2波の最初の表示項目（回答者ごとに異なる）に対するサンプル全体の平均評価点を当てた。評価項目ごとに同じ点数（平均値）が全回答者に与えられている。また、「第2波の第1評価項目評価点（5からの距離）」については、第2波の最初の表示項目に対する回答者ごとの評価点について、中央値（5）からの距離を取っている。

前者は、ある項目に対する評価の高さによって「表示バイアス」の大きさの影響が異なる可能性をコントロールしようとするものである。たとえば、回答者の多くが強く信頼している項目については、ランダムマイゼーションをしたとしても安定的に好意的に評価され、そのために「表示バイアス」も少ないかもしれない。あるいは、その逆に、回答者の多くが不信を持っている項目は、やはりいつも非好意的な評価しか受けないかもしれない。

一方、後者は、「表示バイアス」のおこりえる「マージン」の大きさをコントロールしようとしている。そもそも、11点尺度の中央付近で評価されるもののほうが、ランダムマイゼーションによる評価のズレが大きくなると考えられる。たとえば、ある回答者が「テレビ」に対して9点を第1波で与えたとしよう。すると、第2波では、それより右には1ポイントしかズレのマージンがない。一方、「国会」に対して、5点を第1波で与えたとしよう。その場合には、左右に5ポイントずつズレのマージンが存在する。「表示バイアス」へのその影響を排除することを狙った。

- 8) なお、標準的な属性カテゴリー（「性別」から「所得」まで）の影響も検討はしたが、いずれもシステムティックな影響が確認できなかった（したがって、最終的なモデルからは除外した）。
- 9) サンプルCは、基本的には、サンプルAと同じ構造を取っている。ただし、サンプルAでは、18項目全体でランダムイズをしているのに対して、サンプルCでは、18の評価項目を3つのグループに分け、グループ毎でランダムイズを実施した。そのことの効果については、本項の課題とは直接に関係がないの

でここでは言及しない。また、サンプル D・E と同時に調査をしたので、「尺度調整調査」群に表 1 では整理しているが、サンプル C については「調整」はない。D・E に対するコントロールとして調査に組み込んだままである。

- 10) 評価項目が 18 項目あれば、18 の階乗 (6,400,000,000,000,000) パターンのうちのいずれかの「順序効果」を導入していることになる。
- 11) 実際は、より複雑である。紙面の関係で、詳細は紹介できないが、最大・最小を尋ねる質問 (図 5) に対して「19.『1 つだけ』は選べない」・「20. あてはまるものはない」を選択する回答者に対しては、「あえて言うなら・・・」とフォローアップの質問が必要になる。したがって、2 段階ではなく、3 段階の構造を実際には取った。質問票の構造が複雑となる (あるいは、質問数が増える) ことから、コストパフォーマンスもよくない。

詳細は、私のホームページ (注 7 参照) で公開をしている質問票を参照いただきたい。

- 12) そもそも、一般的な回答者における「最高評価項目 (あるいは、最低評価項目)」が、あらかじめわかる場合ばかりではない。すでに多くの世論調査で質問がされている「制度信頼」などは、むしろ例外である。その点でも、「先頭項目固定方式」は、一般的な解決方法とは言えない。

参考文献

- Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton and Oxford: Princeton University Press.
- Sudman, Seymour and Norman M. Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass Inc., Publishers.
- Tourangeau, Roger; Lance J. Rips; and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.