社会調査の入力ミスの発生率について

PROCESSING ERROR AT A MAIL SOCIAL SURVEY

吉村 治正[1]・小久保 温[2]・澁谷 泰秀[3]・渡部 諭[4]

[1]奈良大学社会学部
[2]青森大学ソフトウェア情報学部
[3]青森大学社会学部
[4]秋田県立大学総合科学教育研究センター

Processing error of mail-returned survey questionnaires is measured in experimental settings. Two of three pairs of examinees showed unexpectedly high error ratio – about 7% of all entry –, but, fortunately, most of the errors were systematic and relatively easy to detect. Simple typographic error, which is very hard to find, was around 0.5%. Coding error about respondent's occupation and industry was about 15%, which posed another very serious issue for survey researchers.

Key Words: *social survey, processing error, non-sampling error*

1. 問題の所在

社会調査における入力ミスの発生頻度はどのぐらいなのか。研究者が集う席でこういう問いかけをすると、驚いたような顔をされる。適切な標本の抽出や質問項目の設定に技術的な知識が必要であるという認識は、社会調査に携わる研究者の間で広く行き渡っている。だからこそ、社会調査の教科書が何種類も出版されているのである。ところが、調査票を回収した後は単純作業と見なされる。回収した調査票からデータファイルを作成するのは単純作業であり、適切に処理されるのが当然で、そこにミスが生じるなどというのはあってはならないこと、というのが共通した認識であろう。実際、筆者らの周囲を見回しても、調査業者に任せずに自分でデータを集めようとする同業者は、決して多数ではないがいる。しかしながら、その彼らですらデータ入力は業者に任せきりで､入力エラーを調べることなど考えない｡筆者などは、回収票を自身の机に積み上げてエラーチェックをしていたら、同僚に呆れられた経験がある。

だが、この問題は決して軽視していいものではない。海外の研究事例を見れば、ごく単純な入力ミスですら、熟練の作業員でも 0.1%、経験の浅い者だと 1.6%の発生率を示すという報告がある(Weisberg, 266)。一般的な社会調査では調査項目数は少なくとも 100 を越え、大規模な訪問調査などでは 300 近くに達する事もある。仮に 180 項目の調査で 1.2%の入力ミスだとすると､回答者一人あたりの入力ミス発生の期待値は 2.16 件。つまり、どの回答者についても、２カ所程度どこかで間違った情報が記載されている事になる。これはごく単純な数値入力の場合であって、作業が複雑になれば、当然その発生率は上昇する。例えば、アメリカで職業を自由回答で聞き、それを職業分類にしたがってコーディングした際に、全体の 20%近くが作業に携わったスタッフの間で不一致だったという報告もある(Weisberg, 265)。こうなると事態は深刻である。

入力ミスを含め、調査終了後の作業におけるミスその他の問題の発生に対する危機意識は近年急速に高まっており、作業誤差(processing error)として非回答誤差や測定誤差と同様に研究対象として位置づけようという動向も生まれている(Weisberg 2005; Biemer & Lyberg 2003; 矢野 2005)。しかしながら、これらの著書で一様に指摘されるのが、研究事例の少なさである。実際、海外でもこのプロセスに研究者が関与する事は極めて稀であり、実務家としての経験から作業誤差の問題にアプローチした研究例として Bourque & Clark (1992)があるぐらいで、社会学・心理学的なフレームワークに準拠した例は、現時点で存在していない。本邦に目を向けると研究事例はさらに稀少で、小林・雨森・山本(2008)が調査後の作業プロセスを解説しているが、これは社会調査実習の授業を想定した記述となっており、具体的にどのようなエラーがどのような場面でどの程度生じているかといった分析は行われていない。つまり、現状では作業誤差の発生理由もその発生確率も、何もわかっていないに等しい。

そこで本稿では、筆者らが実施した社会調査を機会と

して、この作業誤差を実験的に調べてみた。筆者らは2013年の1月に五市町村（北海道一市、東北二市、近畿二市）で、合計1000人を対象に郵送およびインターネットによる職歴調査を試みている[1]。この調査の回収率等については別途報告していくが、ともかく、この調査によって、筆者らの手元には郵送によって回収された調査票342票が残された。これを学生アルバイトとして募集した被験者に入力させ、エラーの発生率を測定しようというわけである。

2. 調査票回収後の作業

社会調査が実施され、調査票が回収された後に行われる作業は、一般的には以下のようにまとめられる(Weisberg, 264; Biemer & Lyberg, 216)。

1 修正箇所の目視による発見：データ入力の前に、回収票を目視し、二重回答（複数の選択肢にマークがついている）や欄外記入などを見つけ、入力に支障が出ないように修正、あるいは欠損として扱うことを決定する。この段階では、主として回答者（自記式の場合）あるいは調査員（訪問調査の場合）によるミスが修正される。
2 コーディング：自由回答項目について回答のカテゴリーを定義し、数値化する。
3 データ入力：回収票に記載された内容をデータとして入力する。一般的には手による入力が行われるが、光学的機器を用いる事もある。
4 修正：論理的に矛盾する回答や選択肢にないはずの回答の発見と修正。スキップフレームに従っていない回答などもチェックされる。この段階では、主として入力作業を行った作業員によるミスが修正される。
5 ファイル管理：データとして扱いやすい形式に変換、保存する。ケースウェイトの設定なども含む。

こうした手順は、調査の方法や状況によって多少とも変わってくる。例えば、訪問調査の場合は、最初の「修正箇所の目視による発見」を各調査地点の管理者が個々の調査員に口頭で確認しながら行う事ができるのに対し、郵送調査では回収された調査票しか確認の手がかりがない。また、コーディングは、調査の実施前にカテゴリーが決められている場合と、回答者の回答の出方と頻度によって選択肢を定義していく場合とがある。一般的に職業や産業は事前にカテゴリーが定義されている事が多い。さらに重要なのは、こうした手順が調査者一人で全てなし得るものではなく、複数のスタッフが分担して作業を行うという点である。大規模な調査になるほど、これに関わる人間の数は増える。作業に関わる全員がこうした一連の作業内容を詳細に至るまで理解していると想定するのは困難であり、したがって実際の場面では、技術的な内容を含め作業全体を把握し各担当に指示を与える管理者の存在と、分業体制の確立が必要不可欠となる。

データの入力に際しては、小林・雨森・山本(2008)が推奨するように、コードブックを作業に先立って作成しておくことが不可欠になる。コードブックとは、コンピュータファイルの中に入っているデータの定義（例えば変数名やカラムの桁数、欠損の場合の値など）を一覧にしたもので、一般的にはデータの二次利用の段階で必要とされる。これがデータ利用の以前の段階、データ入力に際して必要になるという指摘は重要なものだが、筆者らの経験では、入力作業の段階ではコードブックよりも細かいデータの定義が必要となる。つまり、「数値」の入力と言っても、その内容は

1 選択肢の番号を入力する場合
2 世帯員数や年齢など、回答された数値を入力する場合
3 複数回答項目で、各項目について０か１で入力する場合

と分けられる。この種類ごとに、問題となるケースの出方が異なる。例えば、友人の数などを聞く場合、「2、3人」や「15～6 人」といった回答が必ず現れる。勤続年数の場合なども、「3 ヶ月」や「0.5 年」といった回答が出てくる。このように、選択肢の番号を入力する場合であれば一律にエラーとなるような回答であっても、実数値を入力する場合には正当な回答とみなされることがある。複数回答項目については、調査票の該当部分を見ただけでは、通常の項目と区別がつかない事が多い。また、このような形式でなければ入力に支障が出るということを作業員が理解しているとは限らない。したがって、実際場面では、質問項目ごとにこうした個々の入力形式を作業員に指示してやる必要がある。筆者らは、この指示を一覧表にしたものを便宜的に「入力ポリシー」と呼んでおり、入力作業に際しては、必ず作業者に提示することにしている。

3. 実験の設定

実験状況のセッティングは以下の通りである。まず、今回の調査はスプリットバロット方式となっており、回収された調査票のうち170票は「タイプ１」、172票は「タ

イプ２」と呼ばれるものであった。この二種類の違いは、職業および勤務先の産業をあらかじめカテゴリー化して回答者に提示し、そのカテゴリー番号で回答を求めたもの（タイプ１）と、職業および勤務先の産業を自由回答で記入してもらったもの（タイプ２）にある。したがって、タイプ１の調査票については、データ入力の段階でコーディングを行う必要がなく、タイプ２については、作業員がコード表を見ながらコーディングしていく必要があった。

被験者は、２名１組で３組、合計で６名をアルバイトとして募集した。なお、これらの各組は、お互いに面識がなく、また、他で同じ作業が行われているとは知らない。各組ごとに ID 番号を添付した調査票と入力ポリシーが与えられ､貸し与えられたノート PC に Excel で入力するように指示をした。作業に際しては、二重回答など回答者によるミスを事前に修正することも考えたが、今回はこれを行わない状態で被験者に手渡し、彼らがどのように対応するかを観察する事にした。つまり、被験者は､「調査票の目視による修正」と「コーディング」（タイプ２の場合)、そして「データ入力」の３つの作業をこなすことを求められた。

入力ミスのチェックは、以下のような手順で行った。まず一組目（便宜的に A 組とする）の入力結果と二組目（B 組）および三組目（C 組）の結果を照合し、どちらか一方であっても、合致しないセルが出て来たら、そのセルの位置をマークさせた。次に、被験者とは別に募集した 2 名の学生アルバイトに、マークされたセルの正しい値を調査票の原本を見ながら確認させた。被験者が作業を行う際は、入力ポリシーとこれに関連する指示をメモにして渡し、作業には監督をつけないで、不明な点が出たら指示書を読んで自分たちで判断させるようにしたのに対し、入力ミスの判定作業に際しては筆者らが同席し、逐一指示を与えながら作業を行った。こうして入力ミスと判断されたセルを、①単純な入力ミス（打ち間違い)､②インストラクションにしたがわない事によるミス、③コーディングのミス、④回答者に起因するミス（判別が困難な数字などの読み間違い）と、大きく四種類に分けて数えていった。B 組、C 組についても、同様の手順を繰り返した。

4.　結果

タイプ１の調査票の項目数は 150 項目、回収調査票数が 170 件、タイプ２は 149 項目 172 件回収なので、非該当や無回答のセルを含め、入力セル数は総数で 51128 となる。タイプ２の産業および職業のコーディングエラーを除くと、上記四つの種類の入力ミスは、A 組で 4186 件、B 組で 4034 件、C 組で 277 件となり、これを発生率に直すと、A 組が 7.98%、B 組が 7.58%、C 組が 0.35%という数字になる（表１)。驚くべきはその数字のばらつきで、未経験の学生であることを考慮し二人一組で作業させているにも関わらず､8%もの入力ミスを発生させているのが３組中２組もある。これだけ間違いを大量発生されては、データとしては実用に堪えない。

表１：入力ミスの発生件数および発生率

	A 組	B 組	C 組
(件数)			
単純な打ち間違い	17	315	81
指示に従わないミス	4034	3508	55
コーディングのミス*	4	6	1
回答者のミスに起因	28	53	41
入力エラー発生総数	4186	4034	277
(発生率)			
単純な打ち間違い	0.03%	0.62%	0.16%
指示に従わないミス	7.89%	6.86%	0.11%
コーディングのミス*	0.05%	0.10%	0.08%
回答者のミスに起因	0.05%	0.10%	0.08%
入力エラー発生率	7.98%	7.58%	0.35%
入力エラー発生率**	0.14%	0.82%	0.32%

*タイプ２の産業および職業欄を除く

**指示に従わないミスを除くエラーの発生率

だが、これを四つの種類に分けてみると、問題がはっきりする。つまり A 組および B 組で入力エラーが大量に発生しているのは、②の「インストラクションにしたがわない事によるミス」であり、これを除いた①の「単純な打ち間違い」と③「コーディングのミス」(産業および職業の項目を除く)、④の「回答者に起因するミス」は、全て合計しても 0.5%程度に留まる。A 組と B 組で大量発生している｢インストラクションにしたがわないミス」とは、そのほぼすべてが複数回答項目の入力方式に関するものである。つまり、複数回答項目については、各項目についてマークがあれば１、なければすべて０として入力するようにという指示が与えてあるにもかかわらず、マークのついた項目の番号が入力されていたり、０の部分を空欄にしたりといった入力が続いている。これは要するに作業に携わる被験者が「勘違い」をしているもので、しかもその事に本人が気づいていないため、大量に同じミスが発生しているのである。

幸いなことに、このようなミスは体系的に発生するために、入力後の修正の段階で発見することが比較的容易である。これに対し、本当の意味で注意すべきは、①の

「単純な打ち間違い」の発生である。単純な打ち間違いは、インストラクションにしたがわない事によるミスと比べれば、発生確率は低い。ただし、ここでも被験者ごとのばらつきの大きさが目立つ。特にB組の場合は単純な打ち間違いとしか見なし得ないミスが 0.6%と他の２組に比べ群を抜いて高い。ところが、この種のミスは単純であるが故に、対策も発見も難しい。実際問題として、この種のミスを発見するためにはデータを逐一照合するしか方法がなく、それは入力作業を二重におこなう、つまり二つの組がそれぞれ同じデータを入力する必要があることを意味している。これは要するに、作業コストが２倍になるということである。

次に、タイプ２で設定された職業および産業に関する自由回答のコーディングエラーについて検討したい。今回の調査は職歴をテーマとしており、回答者が学校卒業後に最初についた仕事（初職）から離転職を何回経験して現在の仕事（現職）に到達するか、経験した各々の仕事についてその内容と勤務先の事業内容、さらには入職・離職の年齢や職位を、最大で７回まで聞いている。したがって、産業および職業について、最大 1204（172人×7 回）セルの入力があり得ることになる。ただし、実際には６回も離転職している回答者はむしろ稀であり、転職経験がない、１回だけ転職したという回答者の方が多くなる。この場合、回答者個々人の離転職回数を越えたセルについては、非該当として空欄となる。つまり入力がない。実際に入力されたセルの数を数えると、勤務先事業所の産業については 446、回答者の職業については 442 となる。つまり平均すると、回答者一人あたり 2.5 か所での勤務経験があるということになる。

産業および職業のコーディングエラー発生率は、予想通りというべきか、深刻なものであった（表２）。産業・職業ともに、エラー発生率 17%という組が１組、残る２組も 10 から 12 %のコーディングミスがある。こうした間違いの発生は、回答者が不明瞭な記載をしたために分類に失敗したのか、それともコーディング作業員がコードを正確に理解していなかったために生じたものなのか。回答者があいまいな記述をすることで、コーディングができないという事態は実際に生じている。勤務先事業所の産業として判別不能と判断せざるを得なかった回答としては、例えば「一般の株式会社」、「民間」、「一部上場」などといったものがある。職業についても、「職人」や「一部上場」といった記載から具体的な仕事の内容を推測するのは、無理と判断せざるを得ない。ただし、こうした判別不能として扱われた例は決して多くはない。今回の調査では、産業については 11 件、職業については 3 件だけが、最終的に判別不能と判断された。これは、割合から見れば 2.5%と 0.7%に過ぎず、作業員によるコーディングミスの発生率の 1/4 から 1/10 程度にすぎない。つまり、自由回答のコーディングで生じやすいのは、回答者に起因するミスよりも作業員が誤ったコードを与えてしまうミスであると言える。

表２：産業および職業のコーディングミス発生率

	A 組	B 組	C 組
（勤務先事業所の産業）			
入力済みセル数	446	446	446
コーディングミス数	51	76	46
産業コードエラー率	11.4%	17.0%	10.3%
（職業）			
入力済みセル数	442	442	442
コーディングミス数	52	76	53
職業コードエラー数	11.8%	17.2%	12.0%

今回、これだけのコーディングミスが生じた原因として、入力ミスの確認作業に立ち会った経験から、分類カテゴリーが大きすぎたのではないかという反省がある。回収した調査票の記載欄を見ると、「自動車ディーラーの修理工」、「仕出し弁当屋での調理と盛り付け」といった、具体的な記述が目立つ。こうした記述であれば、日本標準産業分類・職業分類での小分類あるいは中分類で該当する産業・職業を探すのは、さほど難しくない。だが、今回はタイプ１で産業・職業カテゴリーを回答者に示す必要があったため、大分類をベースにして、さらに分類を大きくまとめた６カテゴリ（産業・職業とも）に、派遣社員を想定した「不特定」と「判別不能」を加えた８カテゴリーを設定した。その結果、逆説的ではあるが、作業員の間で混乱が起こったようである。例えば自動車の修理工を「生産労務職」ではなく「専門・技術職」にコーディングしたり、介護職を「専門職」とするか「サービス職」とするかで悩んだり、あるいは農協を「農林水産業」とするか「サービス業」とするかで迷っている例が非常に多い。実際、入力ミスの確認の段階でも、照合に用いたのは主に小分類項目の一覧であった。分類カテゴリーは大きい方が理解しやすいし間違いも少ない、と筆者らは当初予想していたのだが、実際には必ずしもそうではないということが明らかとなった。

5.　考察

作業員によるデータ入力が終わると、一般的にはソフトウェアを用いたデータの論理チェックが行われる。定義域にない数値が入力されていないか、他から大きく外れた回答がないかなどを度数分布表を作って確認し、スキップフレームに従っていないケースを探していく。今

回の実験で入力ミスと判断されたものの大半は、この論理チェックの段階で把握できる性質のものと言える。だが､それでもチェックできないミス､単純な入力ミスが、作業員にもよるが、0.2%から 1%程度は残存する。これを深刻な問題とみなすか否かは、個別の研究者の判断による。

単純な入力ミスは、発生が単純であるだけに発見が難しい。確実な方法としては、複数の作業員による入力と照合であるが、これは作業に要する時間と人件費を跳ね上げる。CAPI(compute assisted personal interview)などは、この作業員による入力ミスをなくすという点で画期的な方法ではあるが、必要機材の購入が必要になる。また、本実験のように自記式の調査でこれを適用しようとすると、つまるところはウェブ調査になり、今度はウェブアクセス能力を持つ標本をどのように抽出するかという、カバレッジ(coverage)の問題が発生する。実際、作業段階における誤差の発生については、データのクオリティという問題だけでなく、経費の問題、調査インフラの問題としても考える必要がある。

いずれにせよ忘れてはならないことは、社会調査は調査票を回収したら終わりということではない、ということである。データ入力という単純作業においても作業をきちんと管理することは不可欠だし、そこにおいて技術的なノウハウが必要とされることもある。調査の基本デザインや実施状況に応じて適切な対策を講じていくことが、結局のところ、総合的なデータの「質」を向上させることにつながると言えよう。

6. 文献一覧

Biemer, P. & L. Lyberg. 2003. *Introduction to Survey Quality.* Wiley Interscience.

Bourque, L. & V. Clark. 1992. *Processing Data.* Sage University Paper.

Weisberg, H. 2005. *The Total Survey Error Approach.* University of Chicago Press.

小林久高・雨森聡・山本圭三. 2008. 「社会調査データの入力とチェックの方法」, 『同志社社会学研究』, 12.

矢野宏. 2005. 『誤差のおはなし』. 日本規格協会.

[1] 本調査の実施にあたっては、文部科学省科学研究費平成 23～25 年度採択課題（基盤 C）『郵送・インターネットによる実験的な職歴調査の実施』（課題番号 23530623、代表研究者：吉村治正）ならびに平成 24～25 年度の財団法人電気通信普及財団助成研究「インターネット社会調査と多様化する情報端末―スマートフォン、タブレット端末時代の社会調査に向けて」（採択者：小久保温）による助成を受けた。

PROCESSING ERROR AT A MAIL SOCIAL SURVEY

Harumasa YOSHIMURA[1], Atsushi KOKUBO[2], Hirohide SHIBUTANI[3], Satoshi WATANABE[4]

[1]Faculty of Sociology, University of Nara
[2]Faculty of Software and Information Technology, Aomori University
[3]Faculty of Sociology, Aomori University
[4]Research and Education Canter for Comprehensive Science, Akita Prefectural University

郵送によって回収された調査票の入力エラーの発生率を、２名１組にした被験者３組を用いて測定した。３組中２組の入力ミスが 7%前後という驚きの結果が得られたが、このほとんどが体系的なミスで、論理的エラーチェックの段階で比較的容易に発見できるものであった。極めて単純な入力ミスは 0.5%程度だが、この種のミスは発見が極めて困難である事から、事前の対策の必要性が指摘された。なお、産業・職業に関するコーディングエラーの発生率は 15%前後となり、こちらも看過できない問題である事が指摘された。

キーワード：社会調査，郵送法，非標本誤差