Journal of Official Statistics, vol. 29, n. 2 (2013)

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 223–231

The 2012 Morris Hansen Lecture: Thank You Morris, et al., For Westat, et al.

Kenneth Prewitt[1]

This article, delivered as the 22nd Memorial Morris Hansen lecture, argues that the contract houses, typified by Westat, are uniquely situated in the cluster of institutions, practices, and principles that collectively constitute a bridge between scientific evidence on the one hand and public policy on the other. This cluster is defined in *The Use of Science as Evidence in Public Policy* as a policy enterprise that generates a form of *social knowledge* on which modern economies, policies, and societies depend (National Research Council 2012).

The policy enterprise in the U. S. largely took shape in the first half of the twentieth century, when sample surveys and inferential statistics matured into an information system that provided reliable and timely social knowledge relevant to the nation's policy choices. In ways described shortly, Westat and other social science organizations that respond to "request for proposals" (RFP) from the government for social data and social analysis came to occupy a unique niche.

The larger question addressed is whether the policy enterprise as we know it is prepared for the tsunami beginning to encroach on its territory. Is it going to be swamped by a data tsunami that takes information from very different sources than the familiar census/survey methods?

Key words: Policy enterprise; RFP; scientific integrity; scientific productivity; boundary organizations; big data.

1. What's The *et al* In the Title About?

In my title, Morris is of course Morris Hansen, and though his contributions are properly celebrated, he was one of many who helped establish the dozen or so flagship contract houses in the U.S that produce a large share of the survey based social knowledge on which I focus. Other important contributors at Westat, for example, include Ed Bryant, Joe Hunt and Joe Waksberg. Well before Westat's founding in 1963, however, came the National Opinion Research Center (NORC) founded by Harry H. Field in 1941 and located at the University of Chicago. Field was advised by three social science giants of the period: Gordon Allport, Hadly Cantril, and Sam Stouffer. Other NORC luminaries are Paul Sheatsley, Peter Rossi, and Norman Bradburn. The Institute of Social Research (ISR) founded in 1946 at the University of Michigan brings to mind Rensis Likert, its storied first director, and also in its early days Charles Cannell, George Katona, Leslie Kish, and Dorwin Cartwright, who were soon followed by Donald Campbell, Philip Converse and many others. Clark Apt, and his Apt Associates, pioneered the for-profit base of contract

[1] Professor at Columbia University, 535 West 116th Street, New York, NY 10027, U.S.A. Email: Kp2058@columbia.edu

research. Among another dozen or so flagship institutions of the sort under discussion are Mathematica, RAND, and RTI.

These institutions are more than “contract houses,” but I will use that term to draw attention to the RFP mechanism that shaped the growth of a new way to collect social information in the second half of the twentieth century.

The new way was much more than taking to scale the methodology of survey research, though it involved that. It was also and perhaps more consequentially a new way to structure the relation between science and public policy. It is this structure that needs re-engineering for the twenty-first century information environment.

2. How Institutions Matter

Ian McNeely, writing with Lisa Wolverton (2008), observes that in the western tradition “organizing knowledge became as important as knowledge itself” (p. xix). In fact, “‘the west’ is better defined by its institutions for organizing knowledge than as a set of cultural values or a region of the world” (p. xiv). Libraries – stretching back to Alexandria and Timbuktu – were places where written knowledge could be stored in one place, made accessible, and added to, helping to establish the idea that knowledge is cumulative. Monasteries used multiple copies of the same text to standardize religious instruction that built a church with priests and parishioners scattered across a large region of the world. Museums, basically an Enlightenment creation, collected flora and fauna, which established a natural science based on taxonomy and comparison. Another major step was the nineteenth century invention of the research universities, initiated in the United States when Johns Hopkins, Clarke University, and the University of Chicago combined two traditions: Germany’s great research institutes with England’s great teaching colleges. These new institutions promoted scientific specialization, giving us the familiar disciplines housed in departments.

Subtract libraries, monasteries, museums, or research universities from western history and its substantive knowledge would of course look very different. I do not promote contract houses into the distinguished company of these institutions, but I do hold that they should be seen as new institutional forms that helped establish the conditions for social knowledge production from the 1950s to today. It has been a specific type of knowledge – largely quantitative and intended for use in public policy. The importance of this is apparent under the next heading.

3. The Policy Enterprise

America, fresh from its victories in World War II, and especially appreciative of the role of science – radar, penicillin, and of course the atomic bomb – latched on to the idea that America’s universities could build a knowledge base relevant to economic growth and national security. This spawned the National Science Foundation (NSF), Defense Advance Research Projects Agency (DARPA), government laboratories such as Lawrence-Livermore and the National Institutes of Health (NIH) laboratories. Big science had arrived. By the mid-1960s, social science was in the big science game. The Coleman Report (Coleman 1996) with its database of 600,000 students and 40,000 teachers in 6,000 schools addressed a “big question” – racial inequality in America’s schools. By the standards of social science in 1964, this was big science.

The policy enterprise was underway. Its institutional base includes: policy think-tanks, now estimated at more than 5,000 worldwide, with 1,800 in the U.S. and about 450 of these in Washington; public policy schools and programs in higher education, providing career training for thousands of positions in the policy enterprise; for-profit consulting firms drawing on social science; advocacy groups and professional lobbyists organized around particular policy goals, repackaging social science to their purposes; government units for social and economic analysis in the executive branch, Office of Management and Budget (OMB) of course, but also in the intelligence agencies and many domestic departments such as education, health, and human services; other units attached to Congress, including the Congressional Budget Office and the Congressional Reference Service; and, of course, the statistical agencies and their arrangements to make data widely available for analysis. New research fields emerged, obviously around the statistical underpinnings of probability sample surveys but also fields such as science and technology studies and what came to be known as the knowledge utilization specialty.

I will not dwell on this history; in her discussion Margo Anderson goes into greater historical depth and advances an important structural hypothesis, with which I agree.

In short, the policy industry is a multi-billion dollar cluster of institutions bringing social knowledge to bear on policy design, implementation, and evaluation. It has given birth to evaluation research, social indicators, ranking schemes, performance metrics, evidence-based policy and practice, accountability measures, best practice, and, more generally, the quantification of policymaking.

Of course none of this is unique to this country. You can identify some version of a policy enterprise in at least 150 countries – China to Ghana, Brazil to Jordan, Britain to Australia.

I return below to the unique place of the contract houses in the policy enterprise, but here insert a historical footnote of considerable importance. The contract houses that began to take shape in the 1940s and 1950s, some rooted in research universities (NORC and ISR especially) and others drawing personnel from the census and statistical agencies (Westat being a prime example), were built on a foundation of scientific and academic principles. This was equally true for the for-profit as for the non-profit institutions.

For instance, there was the expectation that the contract houses would produce science of a quality that matched what was found in universities. In fact, when it made sense, they would go beyond current survey practices. NORC, from the day of its founding, insisted that interviewers be treated not as casual workers, but as specialists key to the quality of the survey effort (see Sheatsley 1981–82). Interviewers were trained before being sent into the field. More generally, the competition for contracts led to constant quality innovations– area probability sampling and randomized digital dialing being dramatic examples (Bryant 1997). Personnel in the contract houses would be active in professional academic societies; in the case of the America Association for Public Opinion Research (AAPOR), Henry Field and his NORC colleagues were among its founders.

The pioneers also insisted on openly produced and publicly accessible knowledge. As early as 1947 the ISR was turning down funders who would not agree to make results publicly available, including a study proposed by the Ford Motor Company. ISR did not refuse corporate sponsors, but applied two criteria: Was it a problem of social importance and would the results be published (Frantilla 1998).

Much more could be said on the principles that guided the founding of the contract houses, whether nonprofit or for profit, but the principles continue today and thus are familiar to and probably taken for granted by readers of this Journal.

4. The Integrity and Productivity of Social Knowledge

I borrow the terms integrity and productivity from David Guston (2000). He points out that when the government purchases scientific knowledge, it needs a guarantee of the integrity and the productivity of the science. Integrity refers to the absence of fraud or other substandard practices.

Productivity is a more complicated concept. Certainly it involves cost-effective performance, but the term has a broader meaning. Productive knowledge is that which meets the exacting criteria of “usefulness” to its public sponsors.

Both integrity and productivity are reasonable demands. A government that wasted public funds on fraudulent or useless social knowledge puts its own legitimacy at risk. Fraud is fairly straightforward, but productivity takes us into tricky territory. This is clear if we compare social knowledge to engineering or biological knowledge. What constitutes fraud is comparable across these knowledge sources – the deliberate effort to get the user to accept as true something which is false.

No such comparable standard is available for productivity. What do the engineers or the natural sciences promise when they claim their knowledge is “useful”? Engineers promise bridges that won’t fall down; physicists a missile that will hit its targets; biochemists vaccines that prevent diseases.

But what does the policy enterprise promise? The promise is not, or at least should not be, better data. It is not, or at least should not be, better policy. On first reading you may find this counterintuitive. But better social data does not, in the policy world, translate into more usable data. And better policy begs such questions as for whom, under what conditions, over what time frame. These are political more than scientific questions. For additional explanation, see *Using Science as Evidence in Public Policy* (National Research Council 2012).

This does not mean that we lack grounds on which to claim that social knowledge is productive. Social science can promise to describe social conditions, and whether they are changing, in what direction, and with what velocity. That is, we can describe an aging society and its many features that might require policy attention. Social science can also make estimates about what is likely to happen if a policy intervention occurs, and, post hoc, what did happen as a consequence of that intervention.

If the policy maker wants to know how rapidly the population is ageing and what that means for the social security system, we can provide useful knowledge. If he or she asks what is likely to happen if the age at which social security starts paying benefits is moved up or down, we can provide useful knowledge.

Incidentally, this limited but workable definition of productivity has nothing to do with basic versus applied science, terms that tell us nothing about whether or not the policy maker will find the science useful. And for the policy maker it is beside the point whether scientists call their knowledge discipline-based or interdisciplinary. These distinctions might be of interest among scientists, but to the policy maker what matters is whether the knowledge can be put to productive use in the context of the available policy choices.

I return below to the issue of usefulness, but first I want to return to the period in which contract research was getting underway.

5. A Tacit Agreement Fades

In the postwar period, science policy was heavily influenced by Science the Endless Frontier (Bush 1945), especially its strong assertion on behalf of government investment in basic science. The National Science Foundation was the most visible result on the federal scene; peer review science was ascendant. Science policy in the U.S. was generally based on the assumption that science could perform most productively if free of government control, though of course not free of public obligations. Science, solely concerned with the truth, did not need to be tightly regulated or directed. Its internal policing mechanisms would guarantee scientific integrity and its culture of responsibility would guarantee productivity.

As Don Price (1965) makes clear, this tacit agreement was short lived. The generalized trust in science was gradually replaced by incentives, oversight, performance measures, and related institutional arrangements by which the government assures itself that publicly funded science meets the criteria of integrity and productivity. The current reflection of this is the concern in science policy circles with metrics that can assess “broader impacts” of the government’s investment in science.

Principal-agent theory helps us see what the issues are. The principal – the government – lacks the expertise to produce knowledge it needs. It needs to delegate to an agent the task of producing expert knowledge, that is, scientific research. If the government trusts the integrity and productivity of its agent, nothing else is called for. The problem of science policy is solved.

If, however, the government worries that perhaps not all of the knowledge it is purchasing is free of fraud or rent seeking, and worries even more that scientists have an inclination to be more concerned with peer approval than in producing what society needs, the government will monitor its expert agents and create incentives to influence their behavior in desired directions.

RFP is an obvious mechanism. The basic RFP design specifies in detail what research is to be carried out and uses price as a prime criteria for awarding the contract. This is not the place to assess whether the particularities of the RFP optimize productive knowledge, though thoughtful participants believe it does not. At one point there was hope that the RFP might be designed differently – that is, the government agency would define its objectives, make clear how much it intended to spend, and then judge proposals in terms of their scientific merits within the budget constraints provided. This idea was never seriously considered.

My interest is a particular consequence of RFPs in structuring social knowledge intended for use in public policy. RFPs uniquely position the contract houses on the boundary between science and government. In fact, the contract house can be understood as a “boundary organization,” using the term in the specific way it has been developed in science and technology studies, especially by Shelia Jasanoff (1990). She writes that bringing the scientific community into “decisionmaking produces a stronger consensus than any achievable through the agency’s in-house expertise alone” (p. 237).

As elaborated at greater length in Dan Gaylin’s discussion, a boundary organization, in conducting research guided by an RFP, is responding to what the government has defined for itself as productive knowledge. This allows the government to claim that they are holding their agents accountable for the integrity and productivity of their work. Simultaneously, the contract house is advancing scientific goals. If the product is statistical data, the data will meet standard scientific criteria. If the product is analysis, the reports and the methods by which it is produced will enter the stream of public social knowledge. Stated most simply, the contract houses identified above are a successful example of institutions that respect the government’s need for reliable and productive social knowledge and do so without compromise to scientific principles.

This is not a trivial observation. If the policy enterprise alluded to above is a four to five billion dollar annual effort, the contract houses are responsible for perhaps half of the public funds involved.

On integrity – I know of no instance of scientific fraud associated with any of our flagship contract houses. Certainly there have been cost over-runs and an inability to meet deadlines or other targets such as high response rates. It does not trivialize these failures to note that the occurrence is low and the magnitude is nothing at the scale routinely reported about the government’s purchase of weapon systems.

The issue of productivity is, of course, more difficult to assess – but the growth in dollar volume, in size of studies, in methodological innovation, in timely delivery, and most other metrics we might mention is indirect evidence that the contract houses have proven their worth against the exacting criteria that principle-agent theory puts on the table. They have been productive.

At a moment of skepticism, even cynicism, about the contribution of science to public policy, it is reassuring to have this “success” story available. It is even more important to have this asset in place as we turn to a challenge unimaginable to the founders of NORC, ISR, RTI, Rand, Mathematica, and, of course, Westat.

6. The Digital Data Tsunami

The tsunami is the large and growing supply of social information from digital sources – credit card transactions, surveillance cameras, internet search patterns, social media, with many more technologies yet to come. Before addressing the challenge this poses to the policy enterprise, I want to emphasize what is probably the greatest achievement of the practices and principles we associate with that enterprise.

It has produced a high quality, *shared* information base for the nation’s polity, economy, and society. This has made for healthy democratic debate about policy choices. This is most clearly seen in the arrival of the Great Society policies of the 1960s and 70s, and the critique and partial dismantling of them in the decades since. The critical point is that the information order used to design and implement the Great Society policies was used to evaluate and then challenge them as negative unintended consequences were documented – welfare dependency; the hidden taxes in government regulations; the poor record of urban school reform leading to demands for choice; the mixed record of affirmative action as it benefited upper income African Americans and immigrant groups, especially Asians and West Indians, for whom it was not originally intended.

But it is not the policies that interest us here. It is the fact that America created social knowledge facilitating robust debate about public policies. The statistical underpinnings of the debates had characteristics so familiar that we assume their permanence – high quality, peer reviewed, publicly disseminated, theoretically guided, data representative of the entire population, all provided as a public good.

Whether the digital data tsunami now on its way will disrupt the system carefully assembled over the last century is unclear. I have no crystal ball. But at least in its early days we know much of the digital data from the private sector to be proprietary, of unknown quality, guided less by social theory than commercial benefit, largely unconcerned with privacy/confidentiality, unrepresentative in troubling ways and only incidentally provided as a public good.

These weaknesses notwithstanding, it is difficult for the government to let digital data go unused. It is too much, too cheap, too easy. What is already underway in the national security sector will surely migrate to the domestic policy sector on which today's policy enterprise is largely focused. On March 29, 2012, the federal government announced the "Big Data Research and Development Initiative," with the Office of Science and Technology Policy challenging "industry, research universities, and non-profits to join the Administration to make the most of the opportunities created by Big Data" (Office of Science and Technology Policy 2012). A report issued a few months later, titled the "Big Data Gap", observes that the "promise of big data is locked away in unused or inaccessible data" and that most federal "agencies are still years away from using it" (Big Data Gap 2012). We are not surprised to learn that government wants to use the tsunami of digital data, nor to learn that it is ill-prepared to do so. One concern is that it is the IT departments and not the program units that "own" the data.

The questions are obvious. What institutional platform will collect, manage, house, and analyze the tsunami of digital data? Where will "social knowledge" live in the new information order? Will today's statistical agencies remain involved? What parts of government will draft the RFPs? Who will bid? Today's academically rooted contract houses or Google, Apple, McKinsey, and commercial firms yet to be invented? For readers of this Journal, deeply committed to knowledge for the public good, the question is whether you will be in the game or marginalized – important to the second half of the last century but historical relics by the second half of this century.

Five issues are worrisome: ethics, quality, representativeness, theory, and productivity. This is not an exhaustive list of worries, but enough to alert us to the challenge.

Ethics. The policy enterprise worked out some difficult ethical challenges – informed consent, privacy, confidentiality, and access. How can millions of surveillance cameras and billions of electronic censor devices offer informed consent? Some of you have heard me complain about the failure to distinguish between privacy and confidentiality in the census/survey world (Prewitt 2011). Privacy is the public saying, "you don't have a right to know that about me" – don't ask. Confidentiality is the public saying, "don't share my information in any way that it could be used against me" – don't tell. I doubt that privacy protection has a future in the digital data world, but confidentiality protections are more plausible – though only if taken seriously. We have figured out how to protect confidentiality and still provide robust access for legitimate policy analysis – migrating the arrangements from census/survey data to digital data should be possible, but will require alert attention.

Data quality. We know what quality means for census/survey data – sampling error, respondent burden, cognitive bias, imputation, response rates, external validity, attrition in panel studies, and on and on.

What are the equivalent quality issues for digital data? There are few professional meetings with hundreds of papers debating the error structure of digital data. There is no generally accepted understanding of what constitutes errors when it is machines collecting data from other machines and passing the data along to algorithms for analysis and on to clouds for storage and dissemination.

Representativeness. The probability sampling method basic to the census/survey-based information order provides deep theory about what groups are represented in any given study. Representativeness has not been a major issue in the analysis of digital data. On the one hand, the entire Facebook population or users of a specific browser are, in principle, available. On the other hand, persons who are not online and not candidates to be enrolled – too old, too young, too poor, and so on – are of marginal interest to businesses with a product to sell. Again, the unrepresentativeness of digital data is a problem that can be addressed, but not likely by the current providers of digital data.

Theory. I noted Jim Coleman's pioneering study of racial equality. It was less its size than its finding that we recall today – he reported that family characteristics had as much bearing on educational outcomes as school characteristics – cost per pupil, classroom size, and so on. We now examine the out-of-school versus in-school influences on school performance with batteries of questions, longitudinal data sets, and powerful statistical tools. But some readers will recall that Coleman added but two simple questions: do parents go to Parent-Teacher Association (PTA) meetings; is there an encyclopedia in the house. Coleman brought sociological theory to bear.

Survey data may be case poor but they are variable rich. It is our control over the variables that provides theory-derived social knowledge. Digital data are case rich and variable poor, and insofar as the variables are constructed by theory it is likely to be theory drawn from marketing concerns with consumer behavior. The theory deficit in digital data is a problem that can be fixed, but it will take an active three-way partnership that replicates what the contract houses helped build: government defining its policy needs, theory-designed measurement strategy, and data collection expertise.

This twentieth-century partnership of government, academic social science and contract house created a productive social knowledge base because each of the three players brought relevant technical expertise to the table and a shared understanding of what was required. It is a model for what needs to be created as new data, in very large quantities, becomes available.

Productivity. I opined above that the failure to persuasively explain productivity – useful knowledge that was used in public policy – weakened public confidence in the social sciences. The contract houses have been an important corrective, especially in government circles that appreciate the value of high quality data. But this time around – given the hovering presence of commercial players with deep pockets and lobbying skills – even that might not be enough.

7. References

Big Data Gap. (2012). Based on a survey of 151 Federal government CIOs and IT Managers in March, 2012. Available at: http://www.meritalk.com/bigdatagap (accessed October 15, 2012).

Bryant, E.C. (1997). Westat: Still Counting. Maryland: Rockville.

Bush, V. (1945). Science The Endless Frontier. A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development. Washington, DC: United States Government Printing Office.

Coleman, J.S. (1966). Equality of Educational Opportunity. Washington DC: US Department of Health, Education and Welfare, Office of Education.

Frantilla, A. (1998). Social Science in the Public Interest: A Fiftieth Year History of the Institute for Social Research. Ann Arbor, Michigan: The University of Michigan, Bentley Historical Library, Bulletin No. 45. September 1998.

Guston, D.H. (2000). Between Politics and Science: Assuring the Integrity and Productivity of Research. Cambridge, United Kingdom: Cambridge University Press.

Jasanoff, S. (1990). The Fifth Branch: Science Advisers as Policymakers. Cambridge, MA: Harvard University Press, 237.

McNeely, I.F. and Wolverton, L. (2008). Reinventing Knowledge: From Alexandria to the Internet. New York: W.W. Norton & Company. Norton paperback edition, 2009.

National Research Council (2012). Using Science as Evidence in Public Policy. Committee on the Use of Social Science Knowledge in Public Policy. Division of Behavioral and Social Sciences and Education, K. Prewitt, T.A. Schwandt, and M.L. Straf (eds). Washington, DC: The National Academies Press.

Office of Science and Technology Policy (2012). Big Data is a Big Deal. Available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (accessed October 15, 2012).

Prewitt, K. (2011). Why It Matters to Distinguish Between Privacy & Confidentiality. Journal of Privacy and Confidentiality, 3(2), Article 3. Available at: http://repository.cmu.edu/jpc/vol3/iss2/3/.

Price, D.K. (1965). The Scientific Estate. Cambridge, MA: Harvard University Press.

Sheatsley, P. (1981–82). NORC: The First Forty Years. Chicago: University of Chicago. NORC Report 1081–82.

Received December 2012

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 233–239

Discussion

Margo Anderson[1]

Kenneth Prewitt's Morris Hansen Lecture provides us with a provocative analysis of the importance of what he calls the "contract houses" for the production of high quality, scientific, credible official statistics. His larger subject is the relationship between social science and social policy, or social science and politics. He focuses on the development of the contract houses of the past half century and how they function as boundary organizations supporting the integrity and productivity of scientific knowledge, that is, its quality and usefulness for policy and politics.

As an historian, I found myself asking whether the growth of the contract houses was a logical development in the larger development of democratic policy making, or whether there was something special about Morris Hansen in particular, something of a butterfly effect that is actually a bit of a surprise. I will suggest that the connection between the development of social science and American political development is linked. We should not be surprised.

Let me break the world of politics in two, though in practice they don't really separate that well. Merriam-Webster (2013) has several definitions:

(1) "the art or science of government;"
(2) "the art or science concerned with guiding or influencing governmental policy;" and
(3) "the art or science concerned with winning and holding control over a government."

Let's lump the first and second definitions together, and leave the third definition aside for the moment while we embark on a quick American history lesson.

In the 1770s, the Americans who declared independence from Britain had to establish a structure for their revolutionary government (Morgan 1988; Wood 1998). The foundational documents they drafted, the Declaration of Independence, the various state constitutions, the Articles of Confederation, and the 1787 Constitution, articulated the theory of the state at the time, and developed mechanisms for providing for governing. Chief among the principles were that the power of government derives from the "consent of the governed." Government should, in the language of the 1787 Constitution, "establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity."

These documents do not derive political authority or power from God, ancient traditions, a monarch, or an established nobility or propertied class. The Declaration of

[1] University of Wisconsin, Milwaukee, History Department, Milwaukee, WI 53201, U.S.A. Email: margo@uwm.edu

Independence famously posits the self- evident “truth. . .that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.” God may “endow” rights, but “Governments are instituted among Men.”

Operationalizing these tenets was no easy task. Accordingly, accompanying the documents themselves is a huge literature of interpretation, advocacy, and argumentation. The founding “fathers” or framers mandated that government function in public. Since government derived its authority from the “people,” a vigorous free press and institutions of public debate became essential to the functioning of the state. “People” had to make the system go, and the support of the “people” were metrics of success, which led to the development of both the art and science of policy making and the art and science of democratic electioneering. Governing and policy-making involved embracing the goal of developing knowledge, or “science” – ultimately “social science.”

Americans came to count and describe the “people” in the census (Anderson 1988). The constitutional requirement for open records and public debate, and a ‘state of the union’ report from the President to Congress, resulted in the availability of national public records of governance, finance, taxation and expenditure. By the early nineteenth century, they could be compiled into time series of data, published by both the government in official documents and privately in almanacs and statistical compilations. Serendipitously, detailed, regular, relatively reliable, public data poured out of the new government, revealing the dynamics of social life (Anderson 2010).

The big story of the United States in the nineteenth century was growth and expansion. The population grew from 3.9 million in 1790 to 76 million in 1900. The population growth rate was 30–35% a decade until 1880. The nation expanded to the Pacific coast. The concomitant expansion of the economy was also well documented in the data. Though nineteenth century Americans did not yet have a concept of GDP, they knew that economic growth was explosive. Historians now estimate that per capita GDP grew (in 1996 dollars) from $1,163 in 1790 to $4,204 in 1900 to $32,579 in 2000 (Carter et al. 2006, Part C, Ch Ca, Series Ca9–19). Within this context of growth and expansion, a number of historical examples of the interplay of social science and politics reveal the longer trajectory that frames Prewitt’s analysis.

1. Measuring Race, Ending Slavery

This growth and expansion was not without controversy and crisis. Most notably, the foundational documents of the American revolutionary era left the problem of race-based slavery for future generations. Americans have wrestled with issues of race and inequality ever since. Both abolitionists and the defenders of slavery turned to statistics and the social sciences to inform and justify their policy recommendations. “Race science,” a theory designed to defend the institution of slavery, was a response to the political crisis of the future of the nation, and has not held up as “science” in later years.

For example, the 1840 census seemed to show dramatically higher rates of insanity among free blacks in the North than among slaves in the South. Secretary of State and South Carolinian John C. Calhoun oversaw the administration of the census and claimed the results demonstrated why slave emancipation was impossible. Northerners accused him of fudging the numbers. Massachusetts physician and statistician Edward Jarvis undertook a detailed examination of the local census results to understand what had occurred in the enumeration. He prepared an analysis showing that the data were faulty, though they were never officially corrected (Cohen 1982).

Interestingly, Congress responded in the late 1840s by improving the census, passing new legislation for the 1850 census to assure the errors would not be repeated, and investing in technical innovation and new statistical agencies (Anderson 1988). As Prewitt notes, Americans have been able to use the "shared information base" for fighting over contentious policy differences. Even by the middle of the nineteenth century in the terrible days leading up to the Civil War, all politicians had come to recognize that the half century of data generated by the American state was valuable for the nation.

A second example from the Civil War illustrates that commitment to information-based policy making. In this sesquicentennial year of the Emancipation Proclamation, one might want to visit the U.S. Senate galleries and see Francis Bicknell Carpenter's painting, *First Reading of the Emancipation Proclamation of President Lincoln*. Off in the corner in the painting, on the right on the floor is a map. It is a population density map of the slave states, showing the density of the slave population by county. Cartographers in the U.S. Coast Survey drew it in September 1861, four months into the Civil War, using the recently compiled 1860 census data. The map had a place of honor in Abraham Lincoln's office throughout the war, and played a major role in his conceptualization of military strategy and emancipation. The painting and the map were reproduced and sold popularly throughout the war, and provide powerful visual knowledge of the challenges of emancipation (Schulten 2012).

The painting may be viewed at the U.S. Senate website:
http://www.senate.gov/artandhistory/art/common/image/Painting_33_00005.htm
The map may be viewed at the Library of Congress website:
http://hdl.loc.gov/loc.gmd/g3861e.cw0013200
Schulten (2012) has created a companion site to her study with additional copies and information on these maps and the development of statistical mapping:
http://www.mappingthenation.com/
The Civil War maps are in Chapter 4:
http://www.mappingthenation.com/index.php/chapter/index/4

So what do these developments say about the role of social science and policy?

First, professional social science organizations, starting with the American Statistical Association founded in 1839, were always both knowledge producers and advocates for high quality official statistics and data systems. Social scientists were prominent in the founding of the American Association for the Advancement of Science (1848). The American Geographical and Statistical Society (now AGS) was founded in 1851. The men who founded these organizations also joined international efforts. Joseph C. G. Kennedy, Census Superintendent in 1850 and 1860, for example, was a prominent participant in the early meetings of the International Statistical Institute of the 1850s. Organizations of economists, political scientists, and sociologists followed from the 1880s through the early 1900s.

Second, the advisory board of experts is also an old institution. Congress created a "Census Board" in 1848 to devise new methods for the 1850 census. Then Congressman and future President James Garfield convened a study of census methods in 1869 in preparation for rewriting census legislation for 1870. The early twentieth century saw Congress or the White House create numerous study commissions, for example, the Industrial Commission (1898–1902), Immigration Commission (1907–1911), and a Commission on Industrial Relations (1913–1916), with social science expertise. A permanent Census Advisory Committee, with members from the American Statistical Association and American Economic Association, was established in 1919. It has served as the model for additional committees and has functioned since.

Third is the creation of the "spinoff organization" by the 1920s. Whether for lack of funding, or simply because the social science was still untested, officials within government began to create structures outside the state where the knowledge work could continue. By the early 1900s, young social scientists took positions within government and then moved to or returned to permanent university or research positions to continue the work. Walter Willcox, for example, was already on the Cornell University faculty when he took the position of Chief Statistician for Methods and Results for the 1900 census. He returned to Cornell and remained involved in census policy and a prominent advisory committee member for the rest of his career. He lived to 103 and was still advising on census matters in 1960! Wesley Mitchell, another young census staff in the early 1900s, was a founder of the National Bureau of Economic Research in the 1920s.

My last example, on the development of the measurement of unemployment, illustrates the work of those precursors.

The problem of unemployment measurement presented very new challenges when it emerged as an economic and social issue in the late nineteenth century (Duncan and Shelton 1978). See Figure 1 which displays the pattern from 1890–1990. Before the late

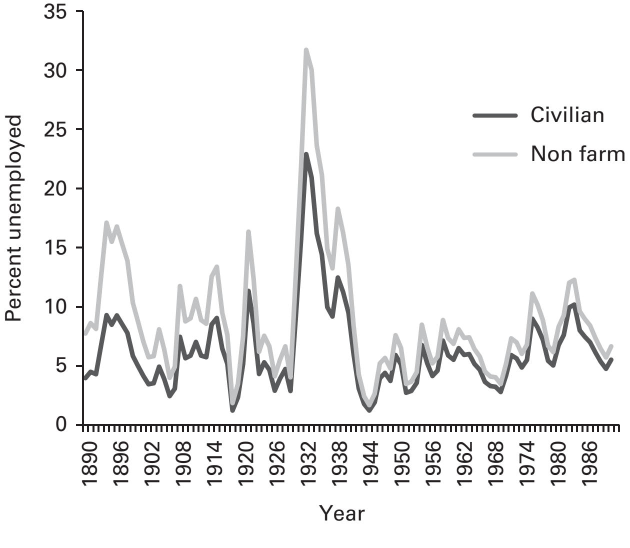


Fig. 1. U.S unemployment rate 1980–1990

1930s, there was no credible official measure of unemployment in the United States, and the data points in this figure for the years before 1935 are estimates developed by historians retrospectively (Carter et al. 2006, Part B, Ch Ba, Series Ba470–477). Data on unemployment, we know now, requires rapid and repeated measurement because the underlying phenomenon is volatile. The graph reveals ragged swings of unemployment after 1890, and more dramatic swings in the non-farm labor force which was growing rapidly as a proportion of the overall economy.

We know Morris Hansen played a major role in the development of the sampling methods used in the 1937 unemployment census, and then improved in the Monthly Report on the Labor Force, now the Current Population Survey.

But let's back up to some earlier developments to see who else was involved.

Here is Mary Van Kleeck (1923, pp. 344–362, quotation at 344) reporting in *Business Cycles and Unemployment*, an NBER report from the President's Conference on Unemployment. This report was prepared after the short intense business depression at the end of World War I:

> If the facts [data on employment and unemployment] are to be useful . . . they must be widely enough scattered geographically not to be over-influenced by condition which may be merely local in one section of the country; they must be made available by some central agency which can correlate and interpret them; and, perhaps most important of all, they must be made public with sufficient promptness to be approximately true measures of the state of employment at the time when they are issued. Thus the problem of extending and improving employment statistics is less statistical in its nature than it is administrative. It demands a machinery strong enough and simple enough to work smoothly and rapidly without breakdowns.

This is quite a mandate. The data have to be current; accurate; credible; geographically distributed to reflect national diversity. Van Kleeck proposed data collections on payrolls and number of employees from employers in manufacturing, trade, mining, railroad transport, utilizing state labor-reporting mechanisms where they existed. She recognized that such a method would omit large portions of the labor force, but had no mechanism to reach the remaining portions of the economy. She sacrificed coverage for efficiency and speed of reporting. Unfortunately, once the economy improved, the pressure to develop the statistics waned. When the next spike in unemployment hit in 1929, the data systems and statistical theory had not advanced.

Then the political problems of improving the statistics hit the statistical system with a vengeance. Government budgets were cut, including those in statistical agencies. For almost eight years, neither President Herbert Hoover nor President Franklin Roosevelt could see any political benefit in developing a statistic that would highlight the administration's failures. So they obfuscated. Roosevelt was famous for confusing journalists by pointing out that when the "breadwinner" lost his job, perhaps his wife or children went looking for work. Three or more people might be looking for work, when all that was really needed was to put the head back to work (see, for example, Roosevelt 1938).

Employer reports failed provide the necessary information. The conceptual definition of what needed to be measured sharpened. Over time, debate shifted to measuring the household situation, which in turn required surveying a much larger respondent base.

Morris Hansen and his colleagues at the Census Bureau recognized they could solve the respondent universe problem with sampling, and they knew it several years before they received authorization from the White House for the 1937 unemployment census. They had to wait out the 1936 presidential election cycle before Roosevelt would authorize the survey.

Now what are the lessons for our discussions of the contract houses?

The social scientists worked independently of the policy makers. In the unemployment case, a parallel process was securing the funding for the new survey.

The policy makers did not necessarily know what they needed. Indeed, they sometimes resisted developing the information and dreaded the analysis they would get.

In all these cases, the social science knowledge creators were relatively unknown to the political establishment or the general public.

In sum, when Morris Hansen came to the Census Bureau in 1935, there were a wide array of extant structures supporting the interaction of social science and federal public policy. Outside government, a statistical revolution was underway, which had yet to penetrate the day to day activities of the statistical agencies or the administrative agencies that produced large amounts of quantitative information. The environment was ripe for new structures, and by the 1940s, the new contract houses appeared on the scene.

Those organizations also benefited from the presence of social scientists and statisticians, like Morris Hansen, who remained within the federal statistical system. He arrived at the Census Bureau at the age of 24, remained for a 30 + year career, joined Westat in 1968 and started another two decades of work. He was in the room, so to speak, with the founders of the contract houses – at professional meetings, and founding new professional organizations such as the American Association of Public Opinion Research, AAPOR (Sheatsley and Mitofsky1992), or when the staff of the contract houses served on federal agency advisory committees. There were models for upholding professional integrity at hand, as well as social scientists who had had interactions with policy agendas in the crucial post-World War II years when the contract houses were getting established.

Prewitt concludes that the social science/government research environment is now facing new challenges, from "big data" and, I might add, from the emergence in the United States of more overtly partisan social science think tanks in the 1970s and 1980s (Smith 1991; Ricci 1993; Rich 2004). This new organizational form has added that third definition of politics to the social science and policy intersection, that of gaining and maintaining control over the apparatus of government, rather than simply providing policy guidance for legislators. Both big data and the rise of the partisan think tank will challenge social scientists and government policy makers alike to rethink the issues of integrity and productivity that Prewitt described. But there is a rich tradition from which to draw to address these new challenges.

2. References

Anderson, M. (2010). The Census and the Federal Statistical System: Historical Perspectives. Annals of the American Academy of Political and Social Science, 631, 152–162.

Anderson, M. (1988). The American Census: A Social History. New Haven: Yale University Press.

Carter, S., Gartner, S.S., Haines, M., Olmstead, A.L., Sutch, R., and Wright, G. (2006). Historical Statistics of the United States, Millennial Edition. New York: Cambridge University Press.

Cohen, P.C. (1982). A Calculating People: The Spread of Numeracy in Early America. Chicago: University of Chicago Press.

Duncan, J. and Shelton, W. (1978). Revolution in United States Government Statistics, 1926–1976. Washington, D.C. Government Printing Office.

Merriam-Webster Dictionary, online edition. (2013). "Politics." Available at: http://www.merriam-webster.com/dictionary/politics (Accessed April 9, 2013).

Morgan, E. (1988). Inventing the People: The Rise of Popular Sovereignty in England and America. New York: Norton.

Ricci, D. (1993). The Transformation of American Politics: The New Washington and the Rise of Think Tanks. New Haven: Yale University Press.

Rich, A. (2004). Think Tanks, Public Policy, and the Politics of Expertise. New York: Cambridge University Press.

Roosevelt, F.D. (1938). Excerpts from the Press Conference, May 13, 1938. Available at: http://www.presidency.ucsb.edu/ws/index.php?pid=15641 (Accessed October 2012).

Schulten, S. (2012). Mapping the Nation: History and Cartography in Nineteenth-Century America. Chicago: University of Chicago Press.

Sheatsley, P. and Mitofsky, W.J. (1992). A Meeting Place: The History of the American Association for Public Opinion Research. Ann Arbor, MI: AAPOR. Available at: http://www.aapor.org/A_Meeting_Place.htm (Accessed October 2012).

Smith, J.A. (1991). The Idea Brokers: Think Tanks and the Rise of the New Policy Elite. New York: The Free Press.

Van Kleeck, M. (1923). Charting the Course of Employment. In Business Cycles and Unemployment: Report and Recommendations of a Committee of the President's Conference on Unemployment. Washington, D.C.: Government Printing Office. Available at: http://www.nber.org/chapters/c4676 (Accessed September 2012).

Wood, G. (1998). The Creation of the American Republic, 1776–1787. Chapel Hill: University of North Carolina Press (originally published 1969).

Received December 2012

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 241–247

Discussion

Daniel Gaylin[1]

This article, delivered as remarks to the 22nd Memorial Morris Hansen lecture, modestly expands on a few themes from Kenneth Prewitt's lecture. The article provides some context on the interrelationship between the federal statistical agencies and the contract houses, and offers some preliminary thoughts about what it means to respond to Prewitt's charge that we cannot rest on our laurels.

1. The Relationship Between Contract Houses, Boundary Organizations, and Think Tanks

To begin the discussion, it is worth briefly reviewing Sheila Jasanoff's (1990) scholarship on boundary organizations and boundary work. Jasanoff's main argument is that by creating sharp lines between science and policy, scientific boundary organizations create legitimacy and "cognitive authority" for their scientific work products. As Prewitt noted, this is a key element of the independence that the contract houses will say they have from the government agencies that fund their activities.

With that in mind, it is also important to place the contract houses into the broader context in which they exist. An imperfect but useful term for these contract houses is "think tank." Two broad definitions of think tanks are as follows:

- A research institute or other organization providing advice and ideas on national or commercial problems (Oxford English Dictionary 2012).
- An institute, corporation, or group organized for interdisciplinary research such as technological and social problems (Merriam-Webster 2012).

Although these very general definitions do not fully capture the core elements of the contract houses, most of the research and scholarship on think tanks clearly categorizes the contract houses as clearly categorizes the contract houses as a particular type of think tank. Indeed RAND, the organization for which the term think tank was invented, is one of the contract houses. James McGann (2007), in his work with the University of Pennsylvania Think Tanks and Civil Societies Program, describes a broad taxonomy of think tanks. I have simplified the taxonomy here into three main types of think tanks:

- Academic think tanks resemble academic institutions and are staffed by academics. They foster academic culture and organization, and follow established academic

[1] Executive Vice President, Research Programs, NORC at the University of Chicago, Chicago, IL U.S.A. Email: gaylin-dan@norc.org

disciplines. They set their own agendas and determine which questions they wish to study. Research conducted by these think tanks generally has longer time horizons and is published in the form of books, journal articles, and monographs. They do not typically issue reports or policy briefs. Two examples of academic think tanks are the Brookings Institution and the Center for Strategic and International Studies.

- Contract think tanks conduct the majority of their research for government agencies using contracts. These organizations have a close working relationship with agencies but are independent and objective, offering data collection and analysis. They are more likely to produce short-term reports and policy briefs. In contract-type think tanks, the researcher's freedom to set research agendas is limited, and usually set by the agency. A few examples of contract think tanks are NORC, RAND, and RTI.
- Advocacy think tanks have a central goal of advancing a cause or ideology. They are usually driven by an issue, philosophy, or constituency and are organized to promote their ideas. They are skeptical of academic, technocratic methods of policy analysis, and cultivate a culture and organizational structure that resembles an advocacy organization. A few examples of advocacy think tanks are the Cato Institute and the Institute for Policy Studies.

In the U.S., the more traditional think tanks are the first two types, but the number of advocacy think tanks has grown in the past few decades.

We will address advocacy think tanks presently, but focusing for now on the first two types, the main point is that for think tanks in the U.S. that focus on objective knowledge generation there are two paths – the federal funding path or the private funding path. Academic think tanks tend to pursue the private funding path, whereas contract think tanks by definition follow the government funding path.

As shown in Figures 1 and 2, there are pros and cons to each path. Academic think tanks might argue that they are more scholarly, have more academic freedom and that they are less subject to bias. Contract think tanks might argue that what they do has the potential for greater impact because they are working directly with the government, that accepting government funds is no more or less biased than accepting private money for research, and

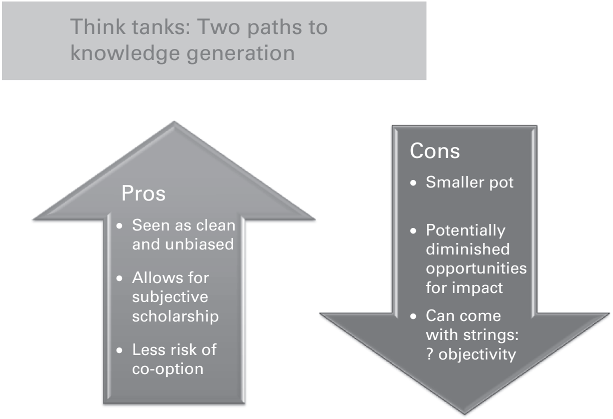

Think tanks: Two paths to knowledge generation
Pros
• Seen as clean and unbiased
• Allows for subjective scholarship
• Less risk of co-option
Cons
• Smaller pot
• Potentially diminished opportunities for impact
• Can come with strings: ? objectivity

Fig. 1. Path 1 – Focus on Private Funding

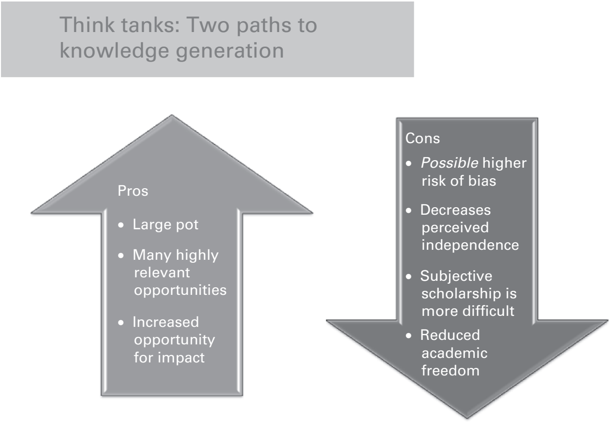

Think tanks: Two paths to knowledge generation
Pros
• Large pot
• Many highly relevant opportunities
• Increased opportunity for impact
Cons
• *Possible* higher risk of bias
• Decreases perceived independence
• Subjective scholarship is more difficult
• Reduced academic freedom

Fig. 2. Path 2 – Focus on Government Funding

that they have a larger playing field. However, when working directly with and for the federal government, contract think tanks need a regime for dealing with the real and perceived risks to their objectivity and independence.

Figure 3 displays a range of approaches by which an organization can mitigate these risks.

Coming full circle, taken as a whole, the approaches in Figure 3 constitute a convincing example of the boundary work that Prewitt references.

2. Enter the Advocacy Think Tank

Despite the successful boundary work by traditional think tanks, an increasing risk to the perceived objectivity of both the academic and contract type think tanks is posed by the extraordinary proliferation of advocacy think tanks that has accompanied increased political polarization in the United States. Andrew Rich (2005), in his book on think tanks, notes that advocacy think tanks are a departure from the commitment to objectivity and independence that is the defining ethos for traditional think tanks. Rich argues that the known ideologies of many, especially newer, think tanks contribute to a situation in which think tanks as a whole, including the more traditional types, are often perceived as promoting points of view and compromising on scientific rigor to do so. As a result, their credibility is undermined and they fail to achieve the substantive impact that they might have. In effect, the scientific boundary work regime is no longer effective because there is *a priori* doubt about the organization promulgating it.

This contributes to the larger milieu of science denial and attacks on the usefulness and credibility of social science and social science data. Examples include the threats faced by the American Community Survey (Groves 2012; Prewitt 2012b; Silver et al. 2012; Webster 2012) and the social sciences arm of the National Science Foundation (Flake 2012a; Flake 2012b). As the line between fact and opinion gets blurred and biased information is promulgated through advocacy think tanks and further disseminated through media outlets with aligned perspectives, the information provided and consumed

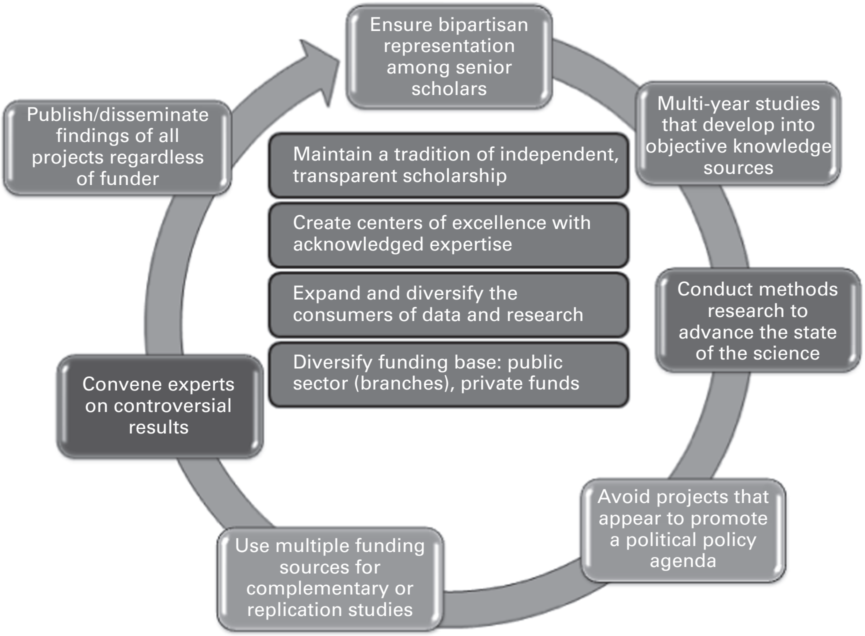

Ensure bipartisan representation among senior scholars
Multi-year studies that develop into objective knowledge sources
Conduct methods research to advance the state of the science
Avoid projects that appear to promote a political policy agenda
Use multiple funding sources for complementary or replication studies
Convene experts on controversial results
Publish/disseminate findings of all projects regardless of funder
Maintain a tradition of independent, transparent scholarship
Create centers of excellence with acknowledged expertise
Expand and diversify the consumers of data and research
Diversify funding base: public sector (branches), private funds

Fig. 3. Avoiding the Pitfalls of Path 2

in the public domain is no longer objective or grounded in science. Hence the role of contract houses as collectors of truly objective information and arbiters/presenters of “truth” proves ever more important. Moreover, government has a steadfast need for objective information upon which to base decisions, devise programs, and design or refine policies. Indeed one of the essential pillars on which the contract houses are based is that the government has this fundamental need for independent and objective, scientifically grounded information. As this pillar starts to crumble, the future of the contract houses becomes less certain.

3. Systematic Limitations of the Contract House Model

In addition to these more exogenous challenges, there are endogenous challenges too. While there may not be major failures of the contract organizations, as Prewitt notes, there are some inherent problems with the system.

Prewitt refers to rent-seeking behavior on the part of the contract houses: What is good for the contract house is not necessarily good for the federal government or the taxpayer. The contract houses carefully steward their incumbencies on long-term recurring projects. The competitive nature of our industry and our business models demand it, but it creates the possibility of conflicts. For example, there is the risk that contract houses will not be as quick as we could be to identify efficiencies in what we do. A good contract house continually assesses this risk and takes steps to mitigate it.

Similarly, there can be problems on the government side. Redundancies and regulations that do not allow data sharing create an environment that results in the government not being as efficient as possible, thereby costing the tax payers more. Anyone who has taken part in the federal budget process has observed the extraordinary effort a federal agency will expend protecting and defending its programs, often despite clear redundancies or substantial overlaps with another agency's programs.

The Office of Management and Budget is a mediator against this *vis-á-vis* government agencies. Other bodies (the National Academy of Sciences, the Committee on National Statistics, etc.) are mediators against it from the contract houses and the agencies, as are the Federal Acquisition Regulations, external review panels, Inspector General's offices, and others. However, there are still weaknesses in the system. A few illustrative examples:

- Problems of focusing on content instead of sample, and vice versa. Despite the use of advisory panels and the dedicated efforts of federal staff, optimizing this tradeoff is a challenging goal in most large data collection efforts.
- Sticking with status quo too long – for example, the overly slow incorporation of cell phone sampling in our surveys (Keeter et al. 2007).
- Successes (but limited successes) with the continuing and recurring efforts at survey integration and data harmonization (U.S. Department of Health and Human Services 1996).
- Lack of means to make data available and useful and the de-legitimization of those data that then ensues (as a type of "sour grapes" by researchers or justification for the limited access by agencies) (Orszag 2009; United States Congress 2012).
- Significant data gaps: we can all agree that we have important holes in the knowledge base, and in many instances we do not have particularly good ideas for filling them.

These examples above are not meant to suggest that the government/contract house model is fatally flawed, but to acknowledge that in addition to the threats posed by advocacy organizations, the contract house model has some inherent challenges that we should continue to address as we look to the future.

4. The Road Ahead

Prewitt, in his article, warns the contract houses and their federal clients, who are "deeply committed to knowledge for the public good, the question is whether you will be in the game or marginalized." What does this suggest for us in terms of concrete action? How does the symbiosis between federal agencies and the contract houses need to evolve as the 21st century unfolds?

Prewitt referred to the digital data explosion: big data, social media, and sampling the internet. The common element here is making better use of existing, new and emerging sources of data, to answer pressing questions more completely and more comprehensively, and also at potentially lower cost.

The contract houses, and indeed the entire federal statistical system need to do a better job of showing the cost effectiveness of their efforts, and that includes both elements of that term. Efficiency is, of course, collecting important data in ways that are high quality, but that cost less. Effectiveness goes further – it requires us to address very crucial questions such as: Why are these data needed, and how are they used? What are good examples of ways in which they have informed policy and other important decision making? What would happen if we did not have them?

We must continue to innovate and expand the science base to incorporate new data sources and new methods for dealing with them into the new cost effectiveness regime that is the reality of 21st century federal budgets. Moreover, we need to do a better job of getting the data out to the data users. This is Todd Park's (2011a; 2011b) "data liberaccion" theme. Park uses that term to make clear that freeing the data, and getting it into the hands of users who can do useful things with it, is something to which we should pay much more attention.

That is the challenge that lies before us, and it is incumbent on all of us to meet it. Boundary work alone will not do it. We must reinvigorate our efforts to provide a clear rationale for the importance of what we do. That begins with the federal funders, but it continues on to the contract houses, who are their partners in this important enterprise.

What might some of the key components of this joint response look like? Prewitt (2012a) spoke at a conference at Stanford University earlier in 2012, delivering a speech in which he tackled some of this head on. In that speech, he discussed how we need a new science and methods base to help us understand the strengths and weaknesses, reliability and validity – or the fitness for use, as Bob Groves (Groves and Lyberg 2010) refers to it – of the new types of large but not necessarily representative data sets that our increasingly connected world creates. In effect, in the 20th century it was the partnership between government, the contract houses, and academia that created the science base which formed the core of our ability to talk about what high quality survey data is. Now we need similar investment and innovation for newer types of data, even as the possible types of partners broaden.

Defining the new methodological approaches is an important part of the overall challenge. But the larger question remains thus: What should a successful partnership between the contract houses and the federal agencies look like going forward, as we try to be more effective in the 21st century, and as the potential data sources to inform policy, and broader decision making, grow exponentially?

5. References

Flake, J. (R-AZ) (2012a). Amendment Offered by Mr. Flake. Congressional Record, 158:65 (May 9, 2012), H2543-44. Available at: http://www.gpo.gov/fdsys/pkg/CREC-2012-05-09/pdf/CREC-2012-05-09.pdf (Accessed October 2012).

Flake, J. (R-AZ) (2012b). Commerce, Justice, Science, and Related Agencies Appropriations Act, 2013. 112th Cong. (2011–2012), H. Amdt. 1094 to H.R. 5326. Available at: http://beta.congress.gov/amendment/112th-congress/house-amendment/1094 (Accessed October 2012).

Groves, R.M. (2012). Census Surveys: Information that We Need. The Washington Post, July 19, 2012. Available at: http://www.washingtonpost.com/opinions/census-surveys-provide-information-that-we-need/2012/07/19/gJQA66wWwW_story.html (Accessed October 2012).

Groves, R.M. and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74, 849–879.

Jasanoff, S. (1990). The Fifth Branch: Science Advisers as Policymakers. Cambridge, MA: Harvard University Press.

Keeter, S., Kennedy, C., Clark, A., Tompson, T., and Mokrzycki, M. (2007). What's Missing from National Landline RDD Surveys? The Impact of the Growing Cell-Only Population. Public Opinion Quarterly, 71, 772–792.

McGann, J. (2007). Think Tanks and Policy Advice in the U.S.: Academics, Advisors and Advocates, (Vol. 1). New York: Routledge.

Merriam-Webster Dictionary. (2012). s.v. "think tank." Available at: http://www.merriam-webster.com/dictionary/think%20tank (Accessed October 2012).

Orszag, P. (2009). Open Government Directive. Memorandum for the Heads of Executive Departments and Agencies, M-10-06, December 8, 2009. Washington, DC: Executive Office of the President, Office of Management and Budget. Available at: http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf (Accessed October 2012).

Oxford English Dictionary. (2012). s.v. "think tank." Available at: http://www.oed.com/view/Entry/200809?redirectedFrom = think + tank#ed (Accessed October 2012).

Park, T. (2011a). Advancing Social Impact Investments through Measurement. Panel 1 Transcript. Washington, DC: Federal Reserve. Available at: http://www.frbsf.org/cdinvestments/conferences/social-impact-investments/transcript/Park_Panel_1.pdf (Accessed October 2012).

Park, T. (2011b). Trends in HIT Innovation. Transcript. Washington, DC: National e-Health Collaborative. Available at: http://www.nationalehealth.org/ckfinder/userfiles/files/Trends%20Transcript.pdf (Accessed October 2012).

Prewitt, K. (2012a). Interactive Session at the First Annual CASBS Summit: Where Social Meets Science. Stanford, CA. June 19, 2012.

Prewitt, K. (2012b). Letters: Census Questions Fulfill Important Purpose. USA Today, July 24, 2012. Available at: http://usatoday30.usatoday.com/news/opinion/letters/story/2012-07-24/census-long-form-mlb-hall-of-fame-steroids/56466354/1 (Accessed October 2012).

Rich, A. (2005b). Think Tanks, Public Policy, and the Politics of Expertise. Cambridge, MA: Cambridge University Press.

Silver, H.J., Casey, R., and Brady, K. (2012). Written Testimony for the Record. Washington, DC: Consortium of Social Science Associations, Joint Economic Committee. Available at: http://www.cossa.org/advocacy/2012/COSSA-ACS-Testimony.pdf (Accessed October 2012).

United States Congress. (2012). Digital Accountability and Transparency Act, 2012. 112th Cong. (2011–2012), 2nd Sess., H.R. 2146. Washington: U.S. Government Printing Office.

U.S. Department of Health and Human Services (HHS). (1996). HHS Plan for Integration of Surveys. Washington, DC: HHS. Available at: http://aspe.hhs.gov/datacncl/srvyrpt1.htm (Accessed October 2012).

Webster, D. (2012). Opposing View: Census Survey Intrusive and Expensive. USA Today, July 15, 2012. Available at: http://usatoday30.usatoday.com/news/opinion/story/2012-07-15/Census-American-Community-Survey/56241350/1 (Accessed October 2012).

Received December 2012

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 249–259

Do Different Listers Make the Same Housing Unit Frame? Variability in Housing Unit Listing

Stephanie Eckman[1]

Housing unit listing is often used in countries that do not have household or person registries to create frames for household surveys. While several studies have reported the kinds of units and areas that are at risk of overcoverage and undercoverage in such frames, none has looked at variability in the listing process. This article explores this variability by comparing two frames created by trained field staff using the same methods and materials. The overall overlap rate between the two listings is 80%. In nearly all blocks, the listers created different frames, and in more than ten percent of the blocks, the two frames did not overlap at all. In this observational study, the overlap between the two frames is particularly low in the blocks listed using the traditional (from scratch) listing method. There is also evidence that sometimes one lister visited the wrong block. The results show that the listing process can introduce variance into survey data.

Key words: Listing; coverage error; coverage variance.

1. Introduction

The Total Survey Error framework for survey data summarizes the ways in which a survey estimate, for example a mean, might deviate from the true mean in the population. An estimate may suffer from measurement error if respondents do not answer the question accurately, or from nonresponse error if nonrespondents are different than respondents. An estimate might also suffer from coverage error if the frame from which the sample was selected does not match the population on the measured characteristic. Each error component can introduce bias and/or variance. If we imagine repeating the entire survey multiple times, some errors will always be the same across repetitions (biases) and some will differ (variances) (Andersen et al. 1979; Groves 1989; Lessler and Kalsbeek 1992; Biemer and Lyberg 2003). Measurement error literature has long focused on variances: see, for example, Fellegi (1964); O'Muircheartaigh and Marckward (1980); Schnell and Kreuter (2005). Nonresponse literature has recently begun to consider how the respondent sample changes across repetitions of the recruitment process (O'Muircheartaigh and

[1] Institute for Employment Research (IAB), Weddingenstraße 20-22 104, 90478 Nuremberg, Germany. Email: Stephanie.Eckman@iab.de

Acknowledgments: This article is based on a chapter of the author's dissertation, written at the Joint Program in Survey Methodology at the University of Maryland. She appreciates the assistance of her committee: Frauke Kreuter (chair), Katharine G. Abraham, J. Michael Brick, and Colm O'Muircheartaigh. The research for the chapter in question was supported by the Census Bureau Dissertation Fellowship, and Tommy Wright, Xijian Liu, Clifford Loudermilk and Aliza Kwiat of the Census Bureau were particularly helpful. Many thanks also to Brady West, Timothy Kennel, Jennifer Sinibaldi, Roger Tourangeau and four reviewers for comments. Any views expressed on statistical, methodological, or operational issues are the author's and not necessarily those of the U.S. Census Bureau.

Campanelli 1999; West and Olson 2010). This article takes a similar approach to the study of coverage error, exploring how listed housing unit frames vary over repetitions of the listing process.

In countries where no register of persons or households is available, listers often create housing unit frames by traveling around the areas selected for the survey and recording the address of every housing unit that they see. In North America, the National Survey of Drug Use and Health (Morton et al. 2006), the General Social Survey (Harter et al. 2010), the National Survey of Family Growth (Lepkowski et al. 2010), the National Children's Survey (Montaquila et al. 2010), the Canadian Labour Force Survey (Statistics Canada 2008), and the Current Population Survey (U.S. Census Bureau 2006) all use listing. Several countries participating in the European Social Survey do so as well (Jowell and the Central Co-ordinating Team 2003, 2005, 2007; Central Co-ordinating Team 2010).

With listed frames, survey researchers worry both about undercoverage, the exclusion of proper housing units from the frame, and about overcoverage, the inclusion of units that do not exist, are not in the selected area, or are not residential. Research has shown that units not occupied at the time of listing and those in small multi-unit buildings (two to nine units) are both undercovered and overcovered (Childers 1992, 1993; Barrett et al. 2002, 2003). Mobile homes are undercovered in listed frames (U.S. Census Bureau 1993; Childers 1993). Undercoverage is also more likely in low-income areas (Manheimer and Hyman 1949; O'Muircheartaigh et al. 2007) and rural areas (O'Muircheartaigh et al. 2007). In addition, households occupied by non-Hispanic black householders have a lower coverage rate than those with non-Hispanic white and other race householders (Barrett et al. 2003). No studies have explored how listed frames vary when the listing task is repeated. Although Kwiat (2009) explores the different actions two listers take when updating an existing frame, that is, whether they confirm, delete or add the same units, these studies do not explicitly compare the frames created by the two listers and do not include blocks where there was no existing frame to update.

Faced with challenging situations, listers may make different judgments and thus produce different frames. For example, in debriefings, six professional listers (none of whom collected the data used in this article) revealed that they use various techniques to count the number of residential units in buildings: some count the mailboxes, others the gas or electricity meters, others the doorbells (Eckman 2010, Appendix F). This article uses data from a repeated listing to explore how similar two frames created by two different listers are. The analyses here do not assert that either frame is more accurate, but instead investigate the overlap between them. The article also investigates the blocks where the two frames do not overlap at all.

2. Data

The data for this study come from two listings of a sample of areas carried out by the U.S. Census Bureau in 2007. Two different trained and experienced Census Bureau field representatives listed the housing units in each of these areas. The analyses in this article compare the overlap, or intersection, of the frames made by these listers.

The Census Bureau maintains a Master Address File (MAF), which aims to be a database of all housing units in the United States and serves as the sampling frame for the

American Community Survey (U.S. Census Bureau 2009). In 2007, the Census Bureau conducted a national evaluation of the MAF's coverage (see Loudermilk and Li 2009 for a discussion of the results). As an add-on, some of the blocks involved in that evaluation were selected for a second listing, carried out by a different lister using the same methods and materials. (Blocks are the smallest geographic units defined by the Census Bureau: they are bounded on all sides by streets, water, railroads or political boundaries.) The subsample gave higher probabilities of selection to blocks with a high rate of growth in the number of addresses available from the U.S. Postal Service, a major component of the MAF. Blocks with no growth were excluded from selection.

Although 301 blocks were selected, only 215 were listed a second time and were found to contain housing units in at least one of the listings (Kwiat 2009). Only these 215 blocks are analyzed in this article, and thus the sample used below is not nationally representative. Unfortunately, none of the housing units listed in this study was selected for a survey, and no data on occupancy status or the characteristics of the residents are available. Also unavailable are any data about which listers were assigned to which blocks, or how many different listers participated in the study.

The listing method used in each block depended on the number of addresses already on the MAF. When the MAF contained addresses for a selected block, the listers were provided with these addresses and updated the list in the field. They added units which lay inside the block but were missing from the MAF, and deleted those that were outside the block or were not housing units. Listers could also move units from one block to another, or simply verify that the unit was correct. This method of listing is called *dependent* or *update* listing. When the MAF contained no addresses for a block, listers traveled around the block and created a frame of housing units: this method is called *traditional* or *scratch* listing. Fourteen of the blocks in this study were listed using traditional listing because the MAF contained no housing units, and 201 were listed using dependent listing.

In each of the blocks, the two listers used the same listing methods and materials. That is, when the first lister in a given block used traditional listing, so did the second. When the first lister used dependent listing, the second lister did so as well, and the input to the second listing was identical to the input to the first: The second listing was not dependent on the first. The listing software provided listers with a map of the blocks they were to list and displayed the addresses already on the MAF, if any. Assignment of listers to blocks was not random but was driven by location and availability. The second listing was always completed within five months of the first.

The 215 double-listed blocks are located in 44 states and 147 counties and thus not tightly clustered geographically. However, there is one group of 22 contiguous blocks in a large west-coast city that was selected into the sample. All but one of these blocks contained no housing units in the 2000 Census, and they were combined into one cluster prior to selection to ensure that the group as a whole would contain some dwellings. However, these blocks grew substantially during the U.S. housing boom of the last decade, and an average of 97 unique units were listed in each of these blocks (range from one to 510). These 22 high-growth blocks were particularly troublesome for the listers, as discussed below.

Across the blocks, the first frame contained 59,365 housing units and the second contained 60,945 housing units. (In each block, the first frame is simply the one that was completed and returned to the central office first.) These counts do not include housing units already on the MAF that were removed by both listers, because these cases do not appear on the final frames. The first step in preparing the data for analysis was matching the two frames. Details on the matching procedures are given in the Appendix.

The two characteristics available for each listed housing unit were whether the unit was a *mobile home* and whether it was part of a *multi-unit* structure. Matching revealed some discrepancies in these variables, which required reconciliation. When adding or verifying a unit, a lister can indicate that it is a mobile home. The design of the Census Bureau listing software makes it more likely that listers will fail to flag a mobile home (a false negative) than falsely flag a non-mobile home (a false positive). For this reason, a matched unit was coded as a mobile home if either lister indicated it was. Only 4.3% of the listed housing units in this study were mobile homes, which is lower than the nationwide occurrence rate of 6.6% (U.S. Census Bureau; American Community Survey 2008).

Units with any text in the apartment field of the address (except those flagged as mobile homes) were designated as in multi-unit buildings. In one percent of the matched cases, the two listers disagreed about whether a unit should have an apartment designator; these units were marked as multi-units. In this dataset, 58.4% of the housing units were flagged as in multi-unit buildings, higher than the corresponding nationwide rate of 32.4% (U.S. Census Bureau; American Community Survey 2008), which points to the over-representation of urban high-growth areas in this study. For those units identified as in multi-unit buildings, information from both listings was used to determine whether the building was small (two to nine units) or large (ten or more units).

In addition, each housing unit on the final two frames can be flagged as either originally on the MAF, or as added by the lister(s). All units in the 14 traditionally listed blocks were by definition added units. In the dependently listed blocks, it was not always straightforward to determine which units were originally on the MAF, because listers sometimes delete a unit that is already on the initial frame and later add that same unit back. In matched cases where one unit was confirmed and the other was added, the unit was *not* marked as an added unit in the dataset. Overall, 10.2% of the units were flagged as added in the dataset.

3. Methods

The results and discussion below center around the overlap rate, defined as the ratio of the number of housing units in both listings to the number of unique housing units in either listing. Put another way, the overlap rate is the size of the intersection of the two frames divided by the size of the union. (Due to the exclusion of blocks where the two listers both listed no units, the denominator is always non-zero.) An overlap rate of 100% indicates that the two frames are identical. A rate of 0% indicates that they have no housing units in common. Thus a high overlap rate means low variability in the listing task and vice versa. (Note that standard measures of inter-rater reliability, such as kappa statistics (Cohen 1969), are not appropriate here, because the data set does not contain housing units in the fourth cell of the two-by-two table, those that neither lister included on the frame.) Due to the nonrepresentative nature of the sample, no significance tests are performed on the overlap rates.

4. Results

The overall overlap rate across all housing units and blocks is 79.9%. That is, more than 20% of all units listed in this study were included by only one of the two listers. It is clear that the two listers did in fact create different frames, and thus that there is variability in the listing process across replications. This finding is new in the literature.

Table 1 breaks this overlap rate down by housing unit and block characteristics. The overlap rate is higher for units already on the MAF (81.9%) than for those that were added (62.4%). There are several possible explanations for this finding. Because the added units were harder to match, as explained in the Appendix, matching errors may explain some of the lower overlap rate for the added units, though the careful matching procedures were designed to minimize such errors. The difference could also be related to failure-to-add confirmation bias. Eckman and Kreuter (2011) found that units not on the initial frame were 14.5 percentage points less likely to appear on the frame than those already included. This substantial reduction in listing propensity could explain why one lister added a unit and another did not, leading to a lower overlap rate for the added units.

The blocks in this study contained few mobile homes, and the overlap rate for these units (73.0%) is only slightly lower than for the non-mobile home cases (80.2%). The difference between the overlap rates for single-family (83.5%) and multi-family units (77.3%) is also small. However, breaking the multi-unit buildings into small and large shows that the overlap rate for units in small buildings is quite a bit lower (66.0%). These results are consistent with previous research that finds small multi-unit buildings to be difficult to list correctly (Childers 1993), as well as with the listers' own statements in the debriefings.

The overlap rate in this study does not differ between rural and urban blocks, despite previous findings that rural areas tend to be undercovered. This finding may mean that all listers are affected similarly by the challenges of rural listing, undercovering the same units. However, the sample used in this study itself underrepresents rural areas, and thus this result may be misleading.

Table 1. Overlap rates, by housing unit and block characteristics

	Overlap rate	*n*
On input list	81.9%	60,042
Added	62.4%	6,838
Mobile home	73.0%	2,868
Non-mobile home	80.2%	64,012
Single family	83.5%	27,825
Multi-Unit	77.3%	39,055
Small, 9 or fewer units	66.0%	4,087
Large, more than 9 units	78.7%	34,968
Rural block[a]	79.8%	12,802
Urban block	79.9%	54,078
Traditional listing	13.6%	523
Dependent listing	80.4%	66,357

[a] Census 2000 Summary File data

The 14 blocks listed via traditional listing have much lower overlap rates than the blocks listed with dependent listing (13.6% vs. 80.4%). However, listing method was not randomly assigned in this study and this result should be interpreted with caution. It could be other block attributes, rather than the listing method itself, which drive the low overlap rates.

The overlap rates do vary quite a bit across the 215 blocks in the study (Figure 1). There are only 20 blocks where the two frames overlap completely, in the upper right corner, and these are small blocks: None has more than 60 unique housing units, and 15 have ten or fewer units. In 116 blocks, the overlap rates are 80% or higher. However, these tend to be the largest blocks, and they represent 74% of the unique housing units in the dataset.

There are 25 blocks in the lower left corner of Figure 1 where the two listers created frames that do not overlap at all. In most (22) of these blocks, the overlap rate is 0% because one lister listed no units. In other words, one lister found no housing units in the assigned block, while the other found one or more. It is possible that both listers saw the structures, but one thought they were nonresidential and thus did not include them. Another possible explanation for the 0% overlap is that one lister was in the wrong block. Inspection of the pattern of housing unit numbers and street names against Google Maps strongly suggests that in at least ten of the 22 blocks, one lister included units outside the selected block. That is, in nearly 5% of the blocks listed in this study, one lister seems to have created a frame for the wrong block. When listers systematically list the wrong block, they overcover the housing units in the wrong block, giving them more than one chance of selection, and undercover the units in the right block, giving them no chance of selection.

The blocks listed via traditional listing are overrepresented among the 0% overlap blocks. Ten of the 25 blocks with 0% overlap rates were traditionally listed, and these ten are the majority of the 14 traditionally listed blocks in the study, which explains the low overlap rate for traditional listing in Table 1. Furthermore, of the ten blocks where evidence suggests that one lister was in the wrong area, eight were listed via traditional listing. The number of blocks and units listed via traditional listing is small in this study,

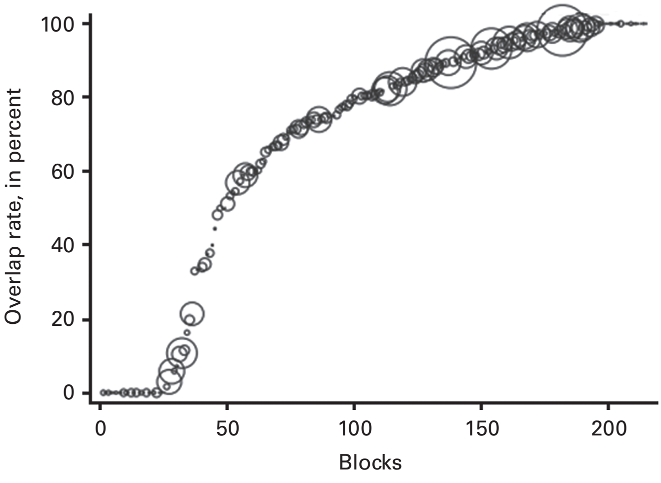


Fig. 1. Block–level overlap rates; horizontal axis is the 215 double-listed blocks, sorted by overlap rate; size of point is relative to the count of unique housing units

due to the nonrepresentative sample. Nevertheless, these results point to high variability in the traditional listing process.

In the debriefings, the listers provided a possible explanation for the connection between traditional listing and the wrong block problem. Several said that they find it easier to locate the selected block when using dependent listing: They simply look for the housing unit listed first on the existing frame and list the block it is in (Eckman 2010, Appendix F). When doing traditional listing, listers cannot use this technique, which may increase the likelihood that they list the wrong block, and thus explain the low overlap rate for traditional listing found in this study.

The 22 contiguous West Coast blocks are also overrepresented among the 0% overlap blocks. The overlap rate across these blocks is only 44.9%, and in thirteen of these blocks the two frames had no units in common. The housing unit stock in these blocks increased by more than 200% from the 2000 to 2010 census, indicating a good deal of growth at some point in the decade, and raising the possibility that real change in the field occurred between the first and second listings, which were at most five months apart. However, close inspection of the frames in these blocks in conjunction with online mapping resources revealed that one lister was confused about the block boundaries. An interstate highway runs through this neighborhood and cuts out very narrow strips of land between the highway and the frontage road on both sides. It appears that one lister did not recognize these strips as blocks and thus was one block off when listing portions of the area.

5. Discussion and Conclusion

This repeated listing study finds that different listers do produce different housing unit frames. Listers disagreed about the inclusion of added units in the frame more than those already on the MAF, and about those in small multi-unit buildings more than single family homes. There was also substantial variation in the overlap rates across the 215 blocks. Traditional listing was more likely to be associated with low overlap than dependent listing, and there is some evidence that listers using traditional listing were more likely to list the wrong block. Because the sample for this study is not nationally representative and the data are observational, the findings may not generalize broadly, but substantial variability in the housing unit listing process is a result not seen before in the literature, and should be of interest to all studies using listing.

The most troubling finding to come from this investigation is that listers may at times be in different blocks, particularly when doing traditional listing. If this finding is replicated in larger studies, several procedural changes could help prevent and detect such errors. First, lister training should include material and job aids on locating the assigned block based on the provided map. Second, quality control procedures should be revised to detect when the wrong block is listed. The Census Bureau's listing check procedure, which sends a senior field representative back to relist the blocks, could catch these mistakes if the second lister does not herself use the addresses listed by the first to locate the block to be checked. The National Survey of Family Growth (NSFG) reviews all of its listings by comparing the street names and numbers against external sources such as Google Maps. This sort of review may be more likely to catch the wrong-block errors, and is less costly than in-field relisting. Third, survey researchers may wish to avoid traditional listing when possible. Less traditional listing may mean fewer errors of listers misreading their maps and listing the wrong block. However, there are two important caveats to this recommendation. First, the addresses on the initial frame may themselves be in the wrong block, due, for example, to geocoding error (Eckman and English 2012). Second, dependent listing has its own vulnerability, namely confirmation bias (Eckman and Kreuter 2011).

This unique double listing dataset has demonstrated that replications of the housing unit listing process do result in different frames. If the residents of the housing units included by the first lister are different, in ways that are captured in the survey items, from those included by the second lister, then the variability in the frames would introduce coverage variance. This study was not able to estimate coverage variance, due to the lack of survey data for all of the listed units, but future studies should aim to do so. Future research should also experimentally manipulate the listing method to better understand how the two methods work. Taking a cue from research into interviewers' contributions to measurement and nonresponse error, coverage studies should collect information about the listers themselves to explore how their characteristics, such as experience, training, and attitudes, affect housing unit listing.

Appendix

The quality of the matching procedures is central to the results presented in the article. The 59,365 housing units on the first listings and the 60,945 unit on the second were run through a three step matching process. The guiding principle throughout the process was whether the same housing unit would be interviewed if the two addresses were selected. For example, unit A and unit 1 at the same address most likely refer to the same unit, and selecting either one would lead to the same unit being approached for an interview, thus these two units were matched. None of the matching steps made use of recent advances in statistical or probabilistic matching (Herzog et al. 2007). While these techniques are appropriate for large-scale matching projects, they are not necessary here where the dataset is rather small, especially within each block, and all addresses can be compared visually.

In the blocks where the listers used dependent listing, matching the units on the input list that each lister verified was straightforward. Every housing unit on the MAF has a unique ID, and the first step simply matched units on the two frames by this ID, identifying housing units that both listers verified. These are the majority of all of the matches identified, as shown in Table 2.

All housing units on the MAF that were not matched in step one and all those added by the listers moved onto step two, which consisted of seven matching routines programmed in SAS 9.1. These routines took advantage of the fact that listers parsed addresses into eight fields in the listing software. The first pass required that all of the address fields, plus block number and apartment identifier, match exactly. Subsequent passes dropped fields from the matching criteria. For example, the second pass did not require a match on the direction prefix field, so that 932 E Elm St would match to 932 Elm St. The seventh pass would match 115 Bryant Ave to 115 Bryant St if they were listed in the same block and still unmatched. Identifying more precise matches first ensured that a low quality match

Table 2. Matches identified, by matching step

Matching step	Housing units	% of all matches
Step 1: Units from MAF	44,753	83.8%
Step 2: Address matches, in SAS	8,428	15.8%
Pass 1: block, unit no., & all address fields	7,849	
Pass 2: drop direction prefix (N, E)	12	
Pass 3: also drop extension	0	
Pass 4: also drop direction (W, NE)	6	
Pass 5: also drop house no. suffix (A, 1/2)	30	
Pass 6: also drop street type prefix	0	
Pass 7: also drop street type (St, Ave, Dr)	531	
Step 3: Address matches, manual review	249	0.5%
Total matched units	53,430	
Total unmatched units	13,450	

would not crowd out a better match. All passes required that the block number, house number, street name, and apartment designator match exactly. These matching routines identified 8,428 matches, more than 90% of those in the first (most precise) pass (see Table 2).

Units still unmatched after Step 2 were output for manual matching. This step caught many spelling and parsing errors as well as different street names (Route 93 versus Main St) and apartment designators (A, B, C versus 1, 2, 3). (In several cases, the spelling and parsing errors in the datasets were fixed and the Step 2 matching routines rerun. The match counts in Table 2 reflect the results after these cleaning steps were applied.) When one lister included two units at an address and the other only one, the single unit was matched to the first unit and the second unit left unmatched. This step identified 249 additional matches (Table 2).

The distance between the geographic coordinates collected by the two listers for each of the matched pairs provides a quality check on the matching. The average distance between the points was 0.06 kilometers, the median was 0.03, and the maximum was 3.3 kilometers. However, the largest distances occurred among the most precise matches (Step 1). In the less precise steps, the distance between the units was always less than one kilometer. Variability in the geocoding of points (Sando et al. 2005; Listi et al. 2007) may explain some of the large distances between matched points.

6. References

Andersen, R., Kasper, J., and Frankel, M.R. (1979). Total Survey Error. San Francisco: Jossey-Bass.

Barrett, D.F., Beaghen, M., Smith, D., and Burcham, J. (2002). Census 2000 Housing Unit Coverage Study. Proceedings of the Section on Survey Research Methods: American Statistical Association, 146–151.

Barrett, D. F., Beaghen, M., Smith, D., and Burcham J. (2003). Census 2000 Housing Unit Coverage Study. Technical Report Census 2000 Evaluation O.3, U.S. Census Bureau, Washington, D.C.

Biemer, P.P. and Lyberg, L.E. (2003). Introduction to Survey Quality. New York: Wiley-Interscience.

Central Co-ordinating Team (2010). European Social Survey Round 3 2008/2009. Final Activity Report ESS4e03.0, City University London.

Childers, D.R. (1992). The 1990 Housing Unit Coverage Study. Proceedings of the Section on Survey Research Methods: American Statistical Association, 506–511.

Childers, D.R. (1993). Coverage of Housing in the 1990 Decennial Census. Proceedings of the Section on Survey Research Methods: American Statistical Association, 635–640.

Cohen, J. (1969). Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin, 70, 213–220.

Eckman, S. (2010). Errors in Housing Unit Listing and Their Effects on Survey Estimates. Ph. D. thesis, University of Maryland.

Eckman, S. and English, N. (2012). Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates. Field Methods, 24, 399–408.

Eckman, S. and Kreuter, F. (2011). Confirmation Bias in Housing Unit Listing. Public Opinion Quarterly, 75, 139–150.

Fellegi, I.P. (1964). Response Variance and Its Estimation. Journal of the American Statistical Association, 59, 1016–1041.

Groves, R.M. (1989). Survey Errors and Survey Costs. Hoboken, N.J: John Wiley and Sons.

Harter, R., Eckman, S., English, N., and O'Muircheartaigh, C. (2010). Applied Sampling for Large-Scale Multi-Stage Area Probability Designs. Handbook of Survey Research, P.V. Marsden and J. Wright (eds). (Second ed). Bingley: Emerald, 169–197.

Herzog, T.N., Scheuren, F.J., and Winkler, W.E. (2007). Data Quality and Record Linkage Techniques. New York: Springer.

Jowell, R. and the Central Co-ordinating Team (2003). European Social Survey 2002/2003. Technical Report ESS1e06.1, London City University.

Jowell, R. and the Central Co-ordinating Team (2005). European Social Survey Round 2 2004/2005. Technical Report ESS2e03.1, London City University.

Jowell, R. and the Central Co-ordinating Team (2007). European Social Survey Round 3 2006/2007. Technical Report ESS3e03.2, London City University.

Kwiat, A. (2009). Examining Blocks with Lister Error in Area Listing. Proceedings of the Section on Survey Research Methods: American Statistical Association, 2546–2557.

Lepkowski, J.M., Mosher, W.D., Davis, K., Groves, R.M., and Hoewyk, J.V. (2010). The 2006– 2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey. Vital Health Statistics, 2(150).

Lessler, J.T. and Kalsbeek, W.D. (1992). Nonsampling Error in Surveys. Hoboken, N.J.: John Wiley and Sons.

Listi, G.A., Manhein, M.H., and Leitner, M. (2007). Use of the Global Positioning System in the Field Recovery of Scattered Human Remains. Journal of Forensic Science, 52, 11.

Loudermilk, C.L. and Li, M. (2009). A National Evaluation of Coverage for a Sampling Frame Based on the Master Address File (MAF). Proceedings of the Section on Survey Research Methods: American Statistical Association, 1721–1734.

Manheimer, D. and Hyman, H. (1949). Interviewer Performance in Area Sampling. Public Opinion Quarterly, 13, 83–92.

Montaquila, J.M., Brick, J.M., and Curtin, L.R. (2010). Statistical and Practical Issues in the Design of a National Probability Sample of Births for the Vanguard Study of the National Children's Study. Statistics in Medicine, 29, 1368–1376.

Morton, K.B., Hunter, S.R., Chromy, J.R., and Martin, P.C. (2006). Population Coverage in the National Survey of Drug Use and Health. Proceedings of the Section on Survey Research Methods: American Statistical Association, 3441–3446.

O'Muircheartaigh, C. and Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. Journal of the Royal Statistical Society. Series A (Statistics in Society), 437–446.

O'Muircheartaigh, C. and Marckward, A.M. (1980). An Assessment of the Reliability of World Fertility Study Data. Proceedings of the World Fertility Survey Conference, 3, 305–379.

O'Muircheartaigh, C.A., English, E.M., and Eckman, S. (2007). Predicting the Relative Quality of Alternative Sampling Frames. Proceedings of the Section on Survey Research Methods: American Statistical Association, 551–574.

Sando, T., Mussa, R., Sobanjo, J., and Spainhour, L. (2005). Quantification of the Accuracy of Low Priced GPS Receivers for Crash Location. Journal of the Transportation Research Forum, 44, 19–32.

Schnell, R. and Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects. Journal of Official Statistics, 21, 389–410.

Statistics Canada (2008). Methodology of the Canadian Labour Force Survey. Technical Report 71–526-X, Statistics Canada.

U.S. Census Bureau (1993). Programs to Improve Coverage in the 1990 Census. Technical report. 1990 CPH-E-3.

U.S. Census Bureau (2006). Technical Paper 66: Design and Methodology, Current Population Survey. Technical report.

U.S. Census Bureau (2009). Design and Methodology American Community Survey. Technical report, U.S. Census Bureau, Washington, D.C.

U.S. Census Bureau; American Community Survey (2008). Selected Housing Characteristics: 2008, 1-Year Estimates. Generated by author using American FactFinder <http://factfinder.census.gov > (Accessed 28 December 2010).

West, B. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance?, Public Opinion Quarterly, 74, 1027–1045.

Received November 2011
Revised February 2012
Accepted January 2013

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 261–276

The Effects of a Between-Wave Incentive Experiment on Contact Update and Production Outcomes in a Panel Study

Katherine A. McGonagle[1], *Robert F. Schoeni*[1,2], *and Mick P. Couper*[1]

Since 1969, families participating in the U.S. Panel Study of Income Dynamics (PSID) have been sent a mailing asking them to update or verify their contact information in order to keep track of their whereabouts between waves. Having updated contact information prior to data collection is associated with fewer call attempts, less tracking, and lower attrition. Based on these advantages, two experiments were designed to increase response rates to the between-wave contact mailing. The first experiment implemented a new protocol that increased the overall response rate by 7–10 percentage points compared to the protocol in place for decades on the PSID. This article provides results from the second experiment which examines the basic utility of the between-wave mailing, investigates how incentives affect article cooperation to the update request and field effort, and attempts to identify an optimal incentive amount. Recommendations for the use of contact update strategies in panel studies are made.

Key words: Panel study; nonresponse; contact strategies; incentives; survey methods.

1. Overview

Since 1969, families participating in the U.S. Panel Study of Income Dynamics (PSID) have been sent a mailing asking them to update or verify their contact information in order to keep track of their whereabouts between waves of data collection. Having updated contact information prior to data collection is associated with fewer call attempts and refusal conversion efforts, less tracking, and lower attrition (Budowski and Scherpenzeel 2005; Couper and Ofstedal 2009; Calderwood 2010; Ribisl et al. 1996). Two experiments were designed with the goal of increasing response rates to this between-wave contact mailing. The first experiment in 2008 increased the overall response rate to the between-wave contact update by approximately 7–10 percentage points by manipulating the design of the mailing, its timing and frequency, whether a study newsletter was also mailed, and use of prepaid versus postpaid incentives (McGonagle et al. 2011). As reported in this article, a second experiment was undertaken in 2010 to determine the overall utility of the

[1] Institute for Social Research, Survey Research Center, University of Michigan, PO Box 1248, Ann Arbor, MI 48106, U.S.A. Emails: kmcgon@umich.edu and mcouper@umich.edu

[2] Department of Economics, 611 Tappan St., and Gerald R. Ford School of Public Policy, 735 S. State St., University of Michigan, Ann Arbor, MI 48109, U.S.A. Email: bschoeni@umich.edu

Acknowledgments: This work was carried out with support from the main sponsors of the Panel Study of Income Dynamics, including the National Science Foundation (SES0518943), the National Institute on Aging (R01AGO19802), and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (R01HD033474). We gratefully acknowledge the assistance of PSID staff members April Beaule, Noura Insolera, Dennis Kloska, Eva Leissou, Mohammad Mushtaq, and Ofelia Purkayastha.

contact update mailing. Families were randomly assigned to a condition in which no mailing was sent in order to examine the effect on data collection effort. A second goal was to examine the effect of providing an incentive on response to the mailing as well as on data collection effort, and to identify an optimal incentive amount. The experimental design randomly assigned PSID panel members to one of four treatment conditions: no mailing, a mailing that included no incentive, or a postpaid incentive of either $10 or $20 in exchange for returning the contact update postcard with a verification or update of their address and/or telephone number.

The goal of this article is to describe the effects of the new experiment on the provision of contact update information as well as its impact on data collection outcomes in Wave 37 of the PSID (2011). The differential responsivity to the incentives for returning the postcard by key socioeconomic characteristics of sample members is examined, and information is provided on the cost-effectiveness of the between-wave mailing.

2. Background

A substantial literature exists on the benefits of providing incentives in exchange for participation in surveys (e.g., Laurie et al. 1999; Laurie and Lynn 2009; Singer et al. 1999a; Singer et al. 1999b; Singer 2002). Research based on longitudinal studies generally finds evidence of a positive association between incentive amount and response rate (Fumagalli et al. 2010; Laurie 2007; Martin et al. 2001; Rodgers 2002) and data collection efforts such as number of calls to complete a case (James 1997; Rodgers 2002). Additional research has documented enduring effects of incentives provided at one wave that persist over time, reducing cumulative nonresponse over multiple waves (Laurie 2007; Mack et al. 1998; Scherpenzeel et al. 2002). Moreover, several studies find that economic characteristics of sample members, such as having low income or being in poverty, increase responsivity to financial incentives, possibly due to the greater value that incentives provide for those who need it most (Laurie 2007; Martin et al. 2001; Mack et al. 1998; Ryu et al. 2006).

In contrast to the large literature on incentives and survey completion, the experimental research on alternative between-wave contact strategies is limited. Fumagalli and colleagues (Fumagalli et al. 2010) found a positive association between incentive amount and return of an address change card. In our 2008 study (McGonagle et al. 2011), we attempted to improve upon the between-wave contact strategy used for many decades in the PSID, which has provided families with a postpaid incentive of $10 in exchange for their return of a contact update postcard. We found no effect of a prepaid versus postpaid $10 incentive on return of the contact update postcard, but we did find a positive effect of a second mailing sent to a subgroup of families who did not return the postcard in response to the first mailing.

The current study drew on these results and the existing literature to design an experiment examining the utility of a four-decade long practice in the PSID of providing incentives for a between-wave contact update mailing. The key research questions that this experiment seeks to answer are the following:

1. Does the between-wave contact update mailing lead to improved response rates to the contact update request and reductions in data collection effort, such as number of calls during production, the tracking of families, and provision of an interview?

2. Does the offer of a conditional incentive compared to no incentive yield similar improvements?
3. Does doubling the incentive from the amount used in many prior waves increase response rates and reduce data collection effort?
4. Do these treatment effects differ by key characteristics of panel members, including being young and having low income?

3. Methods

3.1. Sample and Dataset

The sample of families included in this experiment consisted of 8,690 families from the Panel Study of Income Dynamics (PSID), who had provided a completed interview in the 2009 wave and were eligible to be followed in the 2011 wave. The PSID is a longitudinal study of a nationally representative sample of U.S. families that began in 1968 (see McGonagle et al. 2012 for more information). The original 1968 PSID sample was drawn from two independent samples: a nationally representative sample of roughly 3,000 families designed by the Survey Research Center at the University of Michigan (the "SRC sample") and an over-sample of roughly 2,000 low-income African American families from the Survey of Economic Opportunity (the "SEO sample"). In 1997, 511 families who had immigrated to the U.S. after 1968 were added to enhance the national representativeness of the sample. The study is a genealogical panel, following the original 1968 panel members and the offspring in these households that grow up and form their own economically independent families (known as "split-offs"). Thus the active panel includes related families, with up to four generations of families participating in a given wave. Data were collected annually from 1968 to 1997, and have been collected biennially from 1999 through the most recent wave in 2011. The mode of data collection is via computer-assisted telephone interview (CATI) for approximately 97.5% of panel members, with computer-assisted personal interviewing (CAPI) for the balance. Data collection occurs over a nine-month period from March to December in odd-numbered calendar years.

3.2. Experimental Design and Assignment to Treatment Conditions

As shown in the first row of Table 1, approximately ten percent ($n = 876$) of the families were randomly assigned to receive no contact update mailing; ten percent ($n = 940$) were assigned to receive a mailing but no incentive ($0); approximately forty percent ($n = 3{,}460$) were assigned to receive a mailing and a $10 postpaid incentive (which has been the PSID status quo amount for many waves) and about forty percent were assigned to receive a mailing and a $20 postpaid incentive ($n = 3{,}414$). Because families in the PSID are related and may communicate with each other, all related families received assignment to the same treatment condition, which is why the proportion of families assigned to each condition is slightly variable. Table 1 also provides information on key characteristics of the families as of 2009, with approximately 28 percent reporting a high likelihood of moving, about 33 percent having a household head younger than age 35, and about 71 percent requiring four or more calls to finalize their 2009 interview. Approximately 63 percent of the families are from the SRC sample, 29 percent from the low-income SEO over-sample, and seven

Table 1. Sample sizes and characteristics of families for each treatment condition group (n = 8,690)

	Treatment conditions			
	Mailing sent			No mailing sent
	$0	$10	$20	
Number of families	940	3,460	3,414	876
Characteristics of families in 2009 (%)				
Likelihood of moving before 2011:				
Probably or definitely	28.5	27.3	28.4	28.0
None or uncertain	67.9	69.2	69.2	68.9
Missing	3.6	3.5	2.4	3.1
Family income is less than or equal to bottom quintile	19.6	20.3	19.4	21.9
Number of calls in 2009 to finalize the case:				
1–3	29.3	29.2	28.0	30.5
4 or more	70.7	70.8	72.0	69.5
Age of head of family:				
Less than 35	35.0	32.4	31.8	32.2
35 or older	65.0	67.6	68.2	67.8
Sample types:				
SRC	63.4	59.0	65.8	64.3
SEO	28.4	34.2	27.3	25.6
Immigrant	8.2	6.8	6.8	10.2
Split-off family	6.4	6.7	6.7	6.0

percent from the immigrant sample. About 6.5 percent of the families in total are "split-offs" who became eligible to participate in the PSID for the first time in the 2009 wave.

The contact update mailing consisted of a black and white postcard labeled with the last known address and telephone numbers of the respondent (see Appendix I). The postcard included prepaid postage to cover the cost of returning the mailing. The text of the mailing sent to families in the $10 and $20 incentive conditions read: "Here's how to receive your $10 ($20) check!" The text of the mailing sent to families in the no incentive condition read: "Here's how to update your contact information!" Families in the three mailing conditions were sent the initial contact update mailing in August 2010, approximately seven months before the start of 2011 production interviewing. Drawing on the success of including a second mailing for families who did not return the postcard in the 2008 experiment, the current experiment re-mailed the materials to all families who did not respond to the initial mailing within two months.

3.3. Measures

3.3.1. Outcome Measures

Results for two sets of outcome measures are reported. The first set is referred to as "contact update outcomes" and captures information about respondent behavior in returning the contact information postcard or providing a new telephone number. Analysis of this set is based on the group of families in the three treatment conditions that were all sent the contact update mailing ($n = 7{,}814$). Contact update outcomes consist of "postcard return" and "new telephone number". "Postcard return" is defined as a dummy variable coded as 'yes' = 1 for instances when the respondent returned the postcard and either verified the current information, updated the information, or fixed the current information (e.g., changed 'street' to 'avenue'), and 'no' = 0 for instances when no postcard was received back from the respondent. The overall rate of postcard returns was 68.5 percent, with 7.8 percent of returners providing a new address, 19.4 percent providing an address fix, and 72.8 percent verifying their contact information.

"New telephone number" is defined as a dummy variable coded 'yes' = 1 for instances of receiving a postcard back from the respondent that includes the provision of a new telephone number that was not available in the prior wave, and 'no' = 0 when a new telephone number was not provided. New telephone numbers were provided by 13.5 percent of the postcard returners. Having an accurate telephone number at the beginning of field production is important because the PSID completes more than 97% of its interviews over the telephone.

A second set of measures was designed to assess the effect of the treatment conditions on subsequent data collection outcomes. Analysis of this set is based on families in all four of the treatment conditions, including those sent the contact update mailing and the condition that was not sent the mailing. "Total calls in 2011" is a continuous variable from 1–270 capturing the full range of telephone calls that were made to reach the final disposition of the case during the 2011 field effort. The average number of calls is 15.3 with a median of 6.0. An indicator variable for "high calls in 2011" was constructed to examine whether the treatment conditions were related to a reduction in the number of high effort cases. Cases requiring calls above the mean of 15 were coded as 'yes' = 1 and those below the mean were coded as 'no' = 0. More than 15 calls were required to finalize 24 percent of all cases. A third variable captures information about whether the case needed to be tracked during the field effort, either due to the telephone number on record not being answered or being out of order, or the respondent having moved with no forwarding contact information ('yes' = 1 / 'no' = 0). Nearly 22 percent of the families required tracking in 2011. Finally, the effect of the experiment on the overall 2011 interview response rate was also examined ('yes' = 1 / 'no' = 0). The response rate for the families in this study during the 2011 wave was 94.3 percent.

3.3.2. Measures to Assess Differential Impact

Four key measures were constructed to investigate whether the treatment conditions had differential effects for key subgroups on contact updates and production outcomes. These variables were obtained from the public use data available at the PSID website (http://psidonline.org; Panel Study of Income Dynamics 2009) and were also included in all the models as covariates. First, a dummy variable for likelihood of moving over the next couple of years as reported in 2009 was created with 'no' = 0 indicating "none" or "uncertain" likelihood of moving and 'yes' = 1 indicating a "definite" or "probable" move. A second indicator variable for whether the total family income reported in 2009 was equal to or below the bottom quintile in family income was created ('yes' = 1/'no' = 0). A third indicator variable was constructed to signify whether the age of the household head was under 35 ('yes' = 1) or 35 and older ('no' = 0). Finally, a variable for sample type was included that coded families who were part of the original low-income SEO oversample as 'yes' = 1 and families who were part of the original SRC national probability sample as 'no' = 0.

3.3.3. Control Variables

Three additional variables known to be related to the outcome measures, thereby increasing the precision of the estimated effects of the treatment conditions, were included in all models. In addition to the variables for likelihood of moving, family income, age of household head, and sample type, these included whether the family was a member of the immigrant refresher sample, and whether the family was new to the study as of 2009. A dummy variable coded families who came from the 1997/1999 post-1968 immigrant refresher sample as 'yes' = 1 and those who did not come from this sample as 'no' = 0. A second dummy variable was included that coded families who were designated as new split-off families during the 2009 wave as 'yes' = 1 and non-split-off families as 'no' = 0. Finally, in order to control for high effort cases, an indicator variable for "high calls in the prior wave (2009)" was included, with cases requiring 4 or more calls to complete coded as 'yes' = 1 and those below 4 coded as 'no' = 0.

3.3.4. Analysis Strategy

The first step in the analysis is to provide a description of the bivariate results of the effects of the treatment conditions on the contact update and 2011 production outcomes. The second step describes results from multivariate regression analyses predicting each of the outcome measures from the four treatment conditions. Logistic regression was used to model the contact update outcomes: "postcard return," and "new telephone number;" as well as the 2011 production outcomes: "tracking," "high calls," and "completed an interview." Because of its skewed distribution as a count variable, Poisson regression was used to model "total calls."

4. Results

4.1. Contact Update and Production Outcomes after Treatment: Bivariate Results

Table 2 displays two sets of results. The top panel presents the proportion of families who were sent the mailing that returned the postcard and provided a new telephone number by each of the incentive conditions. The bottom panel includes families who were sent the mailing as well as those assigned to the no mailing condition, and presents the proportion requiring tracking, needing high calls, completing an interview, and the average number of telephone calls to finalize the case, by each of the treatment conditions. All of the statistical tests of mean differences reported in the table control for the comparison-wise error rate using Duncan's multiple range test (Duncan 1955).

Contact update outcomes. There are three findings of note in the top half of Table 2. First, among families sent the mailing, there is a consistent positive effect of the non-zero incentive conditions on contact outcomes. The $10 and $20 incentive conditions resulted in a significantly greater proportion of respondents returning the postcard (68.1% and 71.5%, respectively) and providing a new telephone number (13.6% and 14.2%, respectively) compared to the $0 condition (59.3% returning the postcard and 10.5%

Table 2. Contact update and production outcomes after treatment for each treatment condition group

	Treatment conditions			
	Mailing sent			No mailing sent
	$0 (A) $n = 940$	$10 (B) $n = 3,460$	$20 (C) $n = 3,414$	(D) $n = 876$
Contact update outcomes (n = 7,814)				
% Returning postcard	59.3 (b**,c**)	68.1 (a**,c*)	71.5 (a**,b*)	NA
On first mailing	42.8 (b**,c**)	54.7 (a**,c*)	59.0 (a**,b*)	
On second mailing	16.5 (b**,c**)	13.4 (a**)	12.5 (a**)	
% Providing new telephone number	10.5 (b**,c**)	13.6 (a**)	14.2 (a**)	NA
On first mailing	6.8 (b**,c**)	9.8 (a**)	10.7 (a**)	
On second mailing	3.7	3.8	3.4	
Production outcomes in 2011 (n = 8,690)				
% Tracking required	21.3	22.7	20.6 (d*)	24.4 (c*)
% High calls required to finalize case	22.4 (d*)	24.7	23.8	26.7 (a*)
Average number of calls to finalize case	14.9	15.4	15.3	15.9
% Completing interview	94.7	94.4	94.4	92.9

Notes: Duncan's multiple range test was used to test mean differences. Superscripts a, b, and c designate the groups for which the mean is being tested, with the level of statistical significance of the difference indicated by $^{**}p < =0.01$,$^{*}p < =0.05$.

providing a new telephone number). Second, respondents in the $20 incentive condition had significantly higher rates of postcard returns compared to those in the $10 condition (71.5% vs. 68.1%, respectively) but did not have significantly higher rates of new telephone number provision. The same pattern of results was observed for the percentage of respondents in each treatment condition who responded to the first mailing, with a greater proportion returning the postcard in the $10 and $20 conditions compared to the $0 condition, and those in the $20 condition returning the postcard at a significantly higher rate than those in the $10 condition. Similarly, there was no difference in the rate of new telephone number provision in the first mailing between respondents in the $10 and $20 conditions, but both groups responded to the first mailing with significantly more new telephone numbers than those in the $0 condition.

Third, the second mailing, during which families who had not responded within two months of the initial mailing were re-sent the request for contact update information, was successful in substantially increasing the initial postcard return rate for each incentive group. The second mailing was particularly effective for families in the $0 incentive condition, increasing the postcard return rate in this group by 38 percent (from 42.8% to 59.3%). Families in the $10 and $20 conditions returned the postcard from the second mailing at a significantly lower rate than those in the $0 condition (13.4% and 12.5% vs. 16.5%, respectively), but at rate that was still substantial – adding about 13 percentage points to the overall return rate for both groups. The significantly higher responsivity to the second mailing of families in the no incentive condition compared to those in the $10 and

$20 conditions by approximately three percentage points may simply be due to the greater pool of families in the no incentive condition who were sent a second mailing, given their higher rate of nonresponse to the initial mailing. While there were no statistically significant differences in the rate of new telephone number provision across the three groups in response to the second mailing, it generated a substantial increase in new telephone numbers for each incentive group. Again, the second mailing was particularly effective for the $0 incentive group, raising the overall new telephone number rate by more than 54 percent.

Production outcomes. Examination of the 2011 production outcomes in the bottom panel of Table 2 – which also includes respondents in the "no mailing" treatment condition – indicates that the mailing had beneficial effects on field effort. Respondents in the no mailing condition had significantly poorer results on two of the four production outcomes compared to respondents in the mailing conditions. Tracking rates in this group were significantly higher compared to respondents in the $20 condition (24.4% vs. 20.6%). Compared to families in the $0 condition, those in the no mailing condition were significantly more likely to require a high number of calls to finalize their case.

4.2. *Effects of Treatment Conditions on Contact Update Outcomes: Multivariate Results*

Multivariate regression models were estimated to examine the effects of the treatment conditions on the contact update outcome measures described above. Logistic regression was used to estimate models for obtaining a postcard return and a new telephone number. Each of the models included the set of control variables described earlier.

Table 3 presents the results of the effects of the treatment conditions on contact update outcomes for the three groups that were sent the mailing. The results from the main effect models are consistent with the bivariate results in Table 2 in showing that respondents in the $10 and $20 conditions returned the postcard and provided a new telephone number at significantly greater rates than those in the $0 condition. Respondents in the $20 condition also returned the postcard at a significantly higher rate than those in the $10 condition; there were no significant differences between these groups in the provision of new telephone numbers (results not shown). Interestingly, the control variables had quite different effects on the outcomes as well. Likely movers, families requiring four or more calls to finalize the case, younger heads, and the low-income SEO oversample all had significantly lower rates of postcard returns. The opposite pattern was observed in the models predicting the provision of new telephone numbers, with these characteristics all significantly predicting higher rates, most likely due to the association of these socioeconomic attributes with more frequent changes in telephone numbers. Families from the immigrant refresher sample and those who were newly split off each had significantly lower rates of postcard returns.

In order to test the differential impact of the incentives, a second set of multivariate models included simultaneous terms for interactions between each of the mailing conditions and characteristics of families, including: likely movers, young age of household head, very low family income, and membership in the original low-income SEO oversample. The models included the same control variables included in the earlier models. As shown in Table 3, families with very low income were more responsive to the

Table 3. Effects of treatment conditions on contact update outcomes for families sent a mailing: (n = 7,814)

	Contact update outcomes			
	Returned postcard:		Obtained new telephone number:	
	Main effect model OR	Interaction model OR	Main effect model OR	Interaction model OR
Treatment conditions				
Mailing sent				
$0 (reference)				
$10	1.54**	1.11	1.35**	1.09
$20	1.76**	1.26	1.45**	1.13
Control variables				
Likelihood of moving before 2011:				
Probably or definitely	0.71**	0.57**	1.75**	1.03
None or uncertain (reference)				
Missing	0.63**	0.63**	1.01	1.00
Family income				
Lowest quintile	0.99	0.75	1.68**	1.78*
Highest 4 quintiles (reference)				
Number of ’09 calls to finalize the case:				
1–3 (reference)				
4 or more	0.48**	0.48**	1.17*	1.17*
Age of head of family:				
Less than 35	0.62**	0.50**	1.33**	1.31
35 or older (reference)				
Sample types:				
SRC (reference)				
SEO	0.53**	0.41**	1.39**	1.30
Original PSID sample (reference)				
Immigrant sample	0.42**	0.42**	1.22	1.22
Non split-off family (reference)				
Split-off family	0.63**	0.62**	0.86	0.86
Interaction terms				
Likely movers				
$10 incentive * Probably or definitely		1.24		1.61
$20 incentive * Probably or definitely		1.30		1.98**
Age of head of family:				
$10 incentive * Less than 35		1.26		1.08
$20 incentive * Less than 35		1.36		0.95
Family income				
$10 incentive * Lowest quintile		1.61*		0.88
$20 incentive * Lowest quintile		1.18		1.01
Sample type:				
$10 incentive * SEO sample		1.29		1.12
$20 incentive * SEO sample		1.38		1.01
Mean of dependent variable	0.685		0.135	

OR = odds ratio. **$p <= 0.01$, *$p <= 0.05$.

incentive conditions than higher income families, with those in the $10 group 61 percent more likely to return the postcard than those in higher income quintiles. The financial incentives did not lead to a higher rate of postcard return for families as a function of likelihood of moving, age of household head, or sample membership. Results from the interaction model for provision of a new telephone number found that likely movers who were sent $20 were particularly responsive to the financial incentives compared to non-movers, with the odds of new telephone number provision higher by nearly 100 percent for those in the $20 condition. The financial incentives did not generate higher rates of new telephone number provision for families as a function of age of household head, family income, or sample membership.

4.3. Effects of Treatment Conditions on Production Outcomes: Multivariate Results

Logistic regression was used to examine the effects of the treatment conditions on production outcomes in 2011, including "tracking required," "high calls," and "completed an interview." Poisson regression was used to estimate the effects of the total number of calls. Each model included the set of control variables included in Table 3. Because results from initial regression models found no discernible differences between the incentive amounts on production outcomes, the results presented in Table 4 are based on models that estimate the overall effect of the mailing, collapsed across all three mailing conditions ($0, $10, $20) compared to the no mailing condition, on each of the production outcomes. There was an overall significant effect of the mailing, regardless of incentive amount, compared to no mailing on three of the production outcomes, such that the mailing significantly reduced the odds of tracking, being in the high call group, and needing a high number of total calls to finalize the case.

The control variables had strong effects on each of the production outcomes. Having four or more calls to finalize the case in 2009 was a significant and positive predictor of each indicator of high data collection effort in 2011. Young age of household head and being a member of the original low-income oversample, the immigrant refresher sample, or a split-off family were respondent characteristics that each predicted increased odds of tracking, needing high calls to complete the case, and a higher average number of total calls. As would be expected, being a likely mover significantly predicted higher odds of tracking and higher total calls. Low family income had mixed effects on production outcomes, increasing the odds of tracking and nonresponse, yet predicting fewer overall total calls on average, and reducing the odds of needing a high number of calls to finalize the case.

Several models were estimated to examine the differential impact of the incentive amounts by likelihood of moving, being in the lowest quintile of income, age of household head, and membership in the low-income SEO oversample. In contrast to the findings for contact update outcomes, no evidence was found for differential effects of the incentive amounts on any of the production outcomes by these family characteristics.

5. Discussion

The goals of the current study were to examine the effects of an incentive on cooperation with the request for between-wave contact information, to determine the optimal incentive

Table 4. Effects of treatment conditions on production outcomes for all families: (n = 8,690)

	Production outcomes in 2011			
	Tracking required: OR	Total calls: Poisson *b*	High calls: OR	Completed interview: OR
Treatment conditions				
Mailing sent($0, $10, $20)	0.82*	− 0.06**	0.84*	1.29
No mailing sent (reference)				
Control variables				
Likelihood of moving before 2011:				
Probably or definitely	1.28**	0.02**	1.05	1.28*
None or uncertain (reference)				
Missing	1.25	− 0.01	1.08	0.78
Family income				
Lowest quintile	1.25**	− 0.24**	0.76**	0.71**
Highest 4 quintiles (reference)				
Number of ’09 calls to finalize the case:				
1–3 (reference)				
4 or more	2.48**	0.89**	4.66**	0.46**
Age of head of family:				
Less than 35	1.58**	0.26**	1.45**	1.16
35 or older (reference)				
Sample types:				
SRC (reference)				
SEO	1.61**	0.13**	1.27**	1.31**
Original PSID sample (reference)				
Immigrant sample	1.41**	0.16**	1.63**	0.99
Non split-off family (reference)				
Split-off family	1.44**	0.15**	1.36**	0.80
Mean of dependent variable	0.219	15.3	0.243	0.943

OR= odds ratio. **$p <= 0.01$, *$p <= 0.05$.

amount, and to document the effectiveness of the contact update mailing used in PSID since its inception for data collection outcomes. An important caveat to note at the outset is that the experiment evaluated in this study was conducted in the PSID, which has a two-year period between interviews, making the generalization of results to surveys that have different lengths of time between interviews, such as commonly used annual surveys, uncertain. Four findings from this study are discussed below.

First, we found positive effects of the experimental manipulations of assignment to the mailing conditions on the contact update outcomes. The conditions that included a financial incentive yielded greater cooperation with the request for updated contact information than the no incentive condition. Furthermore, there was a higher rate of postcard return and new telephone number provision for those families receiving $20 compared to those receiving $10, but the magnitude of this difference was fairly small, and does not justify the provision of double the incentive amount for return of the contact information. Moreover, multivariate analyses documented that some socio-economic characteristics of families – including being very low income and having a high likelihood of moving – were associated with elevated responsivity to the incentives.

The beneficial effects of the incentives on cooperation with the contact update request is consistent with other research identifying positive effects of incentives on survey response rates (e.g., Singer 1999; 2002). Nonetheless, more than half of the families in the no incentive group returned the postcard (59 percent). This high response rate among the no incentive group may reflect the loyalty and commitment of the families in the study, and possibly an expectation of receiving future incentives given the consistency of receiving them for past participation in the PSID. This hypothesis is consistent with other research demonstrating the persistence of the positive effects of incentives over waves of panel studies (Mack et al. 1998; Scherpenzeel et al. 2002). It is also worth noting that analyses examining the lagged effects on 2011 outcomes of the treatment conditions to which families were assigned in the first (2008) experiment found that those families assigned to the prepaid incentive condition had significantly higher rates of postcard returns in 2011 than those in the postpaid condition – despite the fact that postcard returns in 2009 were generally unaffected by whether the family was prepaid or postpaid in 2008 (see McGonagle et al. 2011). How long this effect persists will be a topic for additional research, as will the examination of whether in future waves there are lagged positive effects among families receiving $20 in the current experiment, as well as lagged *negative* effects among families who received no incentive.

Second, we demonstrated the overall utility of the between-wave contact update mailing, finding positive effects on key indicators of data collection effort. In multivariate models, families that were randomly assigned to be sent the mailing required less tracking and fewer telephone calls to finalize the case. We generated a basic estimate of the cost effectiveness of the mailing by calculating the approximate point at which it was beneficial in terms of its cost savings in number of calls. We first estimated the approximate cost of the mailing, the bulk of which was due to the incentive payments made to the families who provided updated contact information in the $10 and $20 incentive conditions (approximately $98,000). Across the families that were sent the mailing, the cost was $12.50 per family ($98,000/7,814 families). We then estimated the predicted reduction in calls as a consequence of the mailing compared to no mailing and found it was about 1 call. Thus, across the sample of families sent the mailing, there was a savings of approximately 7,814 calls (i.e., 1.0 calls x 7,814 families). While costs of calls vary, our estimate is that an average call during 2011 data collection costs less than $12.50, suggesting that this design was not cost effective in terms of reducing calls. However, it is likely that there are other potential cost savings that are more difficult to assess, including the resolution of cases earlier in the field period, reductions in tracking effort, the confidence instilled in interviewers who have updated information prior to making a call, and the goodwill generated by the mailing – all of which may be nontrivial factors that contribute to its net effectiveness. Future research devoted to developing cost modeling strategies to evaluate field effort would be very useful to survey practitioners who wish to evaluate interventions designed to improve data collection operations.

Third, there were no differential effects of the financial incentive conditions on production outcomes in multivariate models. Families who were sent the mailing had reduced field effort, regardless of the incentive amount, or whether or not they were promised any incentive. This suggests that it is the mailing per se – and not the particular incentive amount – that may be of most value in achieving positive effects on production outcomes, although it is not known at this time whether these incentive amounts will have lagged effects on future data collection outcomes. Worthy of future study is the identification of the exact mechanism through which the mailing affects production outcomes. It may simply be that the updated contact information makes it easier for interviewers to locate and contact families, resulting in lower rates of tracking and fewer calls. Additionally, the mailing itself may underscore the legitimacy and value of a survey, and evoke principles of social exchange and reciprocation – mechanisms that, it is speculated, also underlie the positive effects of other respondent materials, such as letters sent in advance of data collection (de Leeuw et al. 2007). Furthermore, for a panel survey with a lengthy time span between interviews, the mailing may provide a connection to the study from one wave to the next, and in doing so, heighten the effects of familiarity, which have well documented links to positive affect and other positive outcomes (Bornstein 1989; Lee 2001).

Fourth, certain socioeconomic and demographic characteristics of respondents were associated with an elevated responsivity to the incentives. The finding that low income families who received incentives had higher rates of postcard return compared to higher income families is consistent with other research identifying a greater impact of financial incentives on survey participation by those in poverty and with a low income (James 1997; Mack et al. 1998; Ryu et al. 2006). The current study also finds that incentives increase the odds that likely movers will provide a new telephone number. These results support the use of tailored protocols to avoid a potential source of nonresponse bias, such as determining the incentive amount based on a family's income level, or targeting special incentives at individuals who report in a prior wave that they are likely to relocate in return for maintaining contact information. However, there was no evidence that these differential effects carried over to lower rates of tracking, calls to complete the case, or a higher response rate. In line with an economic exchange model (see Singer 2002), it may be that these differential effects are fleeting and dissipate once the "transaction" of providing the updated information and receiving the financial incentive has been completed.

In conclusion, the results from the current study, along with those from the first experiment, support the ongoing use of a contact update mailing to keep track of families between waves. The most effective approach is one that includes a re-mail to nonresponders. With regards to amount and type of incentive, the two studies together suggest that a postpaid incentive works as well for most panel members as one that is prepaid. In the current study, the $10 and $20 postpaid incentives yielded greater benefits for contact update outcomes than the no incentive condition. While the $20 postpaid incentive produced significantly higher postcard returns and new telephone numbers than the $10 incentive, the difference in effect sizes was small. Moreover, there were no differential effects of the incentive amounts on production outcomes. The results tentatively suggest that the amount of the incentive is less important than sending a mailing between waves as a way to maintain updated contact information as well as to sustain a connection with panel members. However, it is possible that this result is specific to the study design of the PSID and should be replicated in other studies. Moreover, research will be conducted to examine possible lagged effects of the incentive conditions on contact update and production outcomes in the next wave, in order to provide additional information about optimal incentive amounts in panel studies.

6. References

Bornstein, R.F. (1989). Exposure and Affect: Overview and Metaanalysis of Research, 1968–1987. Psychological Bulletin, 106, 265–289.

Budowski, M. and Scherpenzeel, A. (2005). Encouraging and Maintaining Participation in Household Surveys: The Case of the Swiss Household Panel. ZUMA-Nachrichten, 56, 10–36.

Calderwood, L. (2010). Keeping in Touch with Mobile Families in the UK Millennium Cohort Study. Centre for Longitudinal Studies Working Paper Series 2010/7. London: Centre for Longitudinal Studies.

Couper, M.P. and Ofstedal, M.B. (2009). Keeping in Contact with Mobile Sample Members. Methodology of Longitudinal Surveys, P. Lynn (ed.). New York: Wiley, 188–203.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., and Lensvelt-Mulders, G. (2007). The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis. Public Opinion Quarterly, 71, 413–443.

Duncan, D.B. (1955). Multiple Range and Multiple F Tests. Biometrics, 11, 1–42.

Fumagalli, L., Laurie, H., and Lynn, P. (2010). Experiments with Methods to Reduce Attrition in Longitudinal Surveys. Institute for Social and Economic Research Working Paper 2010-04. Colchester: University of Essex.

James, T.L. (1997). Results of the Wave 1 Incentive Experiment in the 1996 Survey of Income and Program Participation. Proceedings of the Survey Research Methods Section of the American Statistical Association. Washington, DC: American Statistical Association, 834–839.

Laurie, H. (2007). The Effect Of Increasing Financial Incentives In A Panel Survey: An Experiment On The British Household Panel Survey, Wave 14. ISER Working Paper, No. 2007-05. Colchester: University of Essex. Available at: www.iser.essex.ac.uk/pubs/workpaps/pdf/2007-05.pdf (Accessed May 31, 2012).

Laurie, H., Smith, R., and Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey. Journal of Official Statistics, 15, 269–282.

Laurie, H. and Lynn, P. (2009). The Use of Respondent Incentives on Longitudinal Surveys. Methodology of Longitudinal Surveys, P. Lynn (ed.). New York: Wiley, 205–233.

Lee, A.Y. (2001). The Mere Exposure Effect: An Uncertainty Reduction Explanation Revisited. Personality and Social Psychology Bulletin, 27, 1255–1266.

Mack, S., Huggins, V., Keathley, D., and Sundukchi, M. (1998). Do Monetary Incentives Improve Response Rates in the Survey of Income And Program Participation? Proceedings of the American Statistical Association, Survey Research Methods Section. Washington, DC: American Statistical Association, 529–534.

Martin, E., Abreu, D., and Winters, F. (2001). Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation. Journal of Official Statistics, 17, 267–284.

McGonagle, K.A., Schoeni, R.F., Sastry, N., and Freedman, V.A. (2012). The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research. Longitudinal and Life Course Studies, 3, 268–284.

McGonagle, K.A., Couper, M.P., and Schoeni, R.F. (2011). Keeping Track of Panel Members: An Experimental Test of a Between-Wave Contact Strategy. Journal of Official Statistics, 27, 319–338.

Panel Study of Income Dynamics, public use dataset (2009). Produced and distributed by the Institute for Social Research, Survey Research Center. Ann Arbor, MI: University of Michigan.

Ribisl, K.M., Walton, M.A., Mowbray, C.T., Luke, D.A., Davidson, W.S., and Bootsmiller, B.J. (1996). Minimizing Participant Attrition in Panel Studies Through the Use of Effective Retention and Tracing Strategies: Review and Recommendations. Evaluation and Program Planning, 19, 1–25.

Rodgers, W. (2002). Size of Incentive Effects in a Longitudinal Study. Proceedings of the Survey Research Methods Section of the American Statistical Association. Washington, DC: American Statistical Association, 2930–2935.

Ryu, E., Couper, M.P., and Marans, R.W. (2006). Survey Incentives: Cash vs In-kind; Face-to-face vs Mail; Response Rate vs Nonresponse Error. International Journal of Public Opinion Research, 18, 89–106.

Scherpenzeel, A., Zimmermann, E., Budowski, M., Tillmann, R., Wernli, B., and Gabadinho, A. (2002). Experimental Pre-Test of the Biographical Questionnaire, Working Paper, No. 5-02. Neuchatel: Swiss Household Panel. Available at: http://aresoas.unil.ch/workingpapers/WP5_02.pdf. (Accessed May 31, 2012).

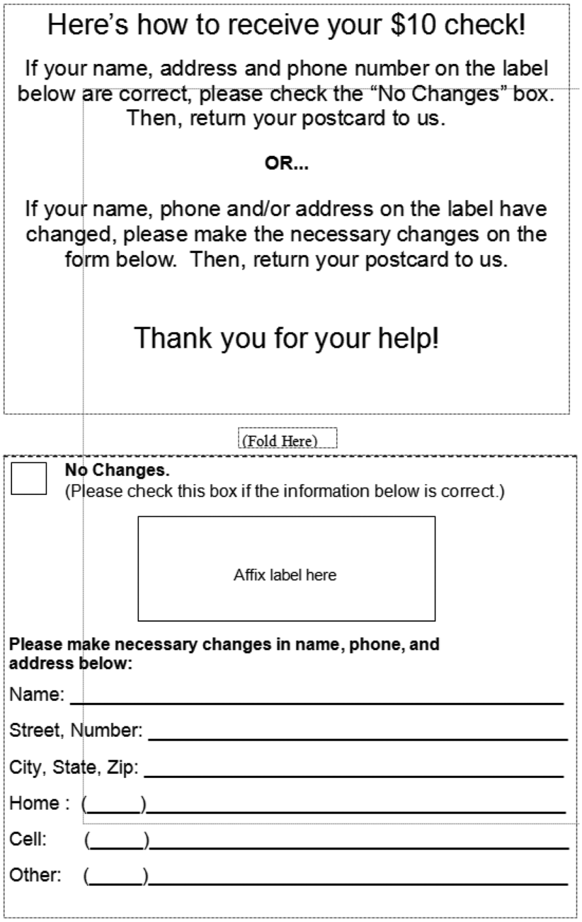
Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 163–177.

Singer, E., Gebler, N., Raghunathan, T., van Hoewyk, J., and McGonagle, K. (1999a). The Effect of Incentives in Telephone and Face-to-Face Surveys. Journal of Official Statistics, 15, 217–230.

Singer, E., van Hoewyk, J., and Gebler, N. (1999b). The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys. Journal of Official Statistics, 15, 217–230.

Received July 2012
Revised November 2012
Accepted November 2012

Appendix I



Here's how to receive your $10 check!

If your name, address and phone number on the label below are correct, please check the "No Changes" box. Then, return your postcard to us.

OR...

If your name, phone and/or address on the label have changed, please make the necessary changes on the form below. Then, return your postcard to us.

Thank you for your help!

(Fold Here)

□ **No Changes.**
(Please check this box if the information below is correct.)

Affix label here

Please make necessary changes in name, phone, and address below:

Name: ______________________________

Street, Number: ______________________________

City, State, Zip: ______________________________

Home : (_____)______________________________

Cell: (_____)______________________________

Other: (_____)______________________________

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 277–297

"Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?

Brady T. West[1], *Frauke Kreuter*[2], *and Ursula Jaenichen*[3]

Recent research has attempted to examine the proportion of interviewer variance that is due to interviewers systematically varying in their success in obtaining cooperation from respondents with varying characteristics (i.e., nonresponse error variance), rather than variance among interviewers in systematic measurement difficulties (i.e., measurement error variance) – that is, whether correlated responses within interviewers arise due to variance among interviewers in the pools of respondents recruited, or variance in interviewer-specific mean response biases. Unfortunately, work to date has only considered data from a CATI survey, and thus suffers from two limitations: Interviewer effects are commonly much smaller in CATI surveys, and, more importantly, sample units are often contacted by several CATI interviewers before a final outcome (response or final refusal) is achieved. The latter introduces difficulties in assigning nonrespondents to interviewers, and thus interviewer variance components are only estimable under strong assumptions. This study aims to replicate this initial work, analyzing data from a national CAPI survey in Germany where CAPI interviewers were responsible for working a fixed subset of cases.

Key words: Interviewer variance; nonresponse error variance; measurement error variance; face-to-face data collection; multilevel modeling; PASS study.

1. Introduction

Effects of interviewers on the measurement of survey variables have been clearly documented in surveys using face-to-face interviewing (Hansen et al. 1960; Kish 1962; Fellegi 1964; Freeman and Butler 1976; Collins and Butcher 1982; Mangione et al. 1992; O'Muircheartaigh and Campanelli 1998; Schnell and Kreuter 2005). In general, for certain survey items, responses collected by the same interviewer will tend to be more similar than responses collected by different interviewers. Assuming that random subsamples of the full sample have been assigned to the interviewers (interpenetrated assignment), one possible source of this between-interviewer variance is measurement error variance among the interviewers (e.g., Hansen et al. 1960; Biemer and Stokes 1991; Biemer and Trewin 1997; Groves 2004, ch. 8).

[1] Survey Methodology Program (SMP), Survey Research Center (SRC), Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, U.S.A. Email: bwest@umich.edu
[2] Joint Program in Survey Methodology (JPSM), Ludwig Maximilian University of Munich, Institute for Employment Research (IAB), Nuremberg, Germany. Email: fkreuter@survey.umd.edu
[3] Institute for Employment Research (IAB), Nuremberg, Germany. Email: Ursula.Jaenichen@iab.de

Acknowledgments: We are indebted to Gerrit Müller, Mark Trappmann and Daniela Hochfellner from the Institute for Employment Research (IAB) in Germany for their help in accessing and navigating the various data sets necessary to conduct this research. We are also grateful to Theresa Scholz from the RDC at IAB for providing the PSU-level data, and to the project "Nachbarschaftseffekte: Die Analyse individuell-rationalen Verhaltens im sozialen Kontext" conducted by the IAB, DIW Berlin and RWI Essen for granting us permission to use the data.

Various hypotheses have been proposed in the literature concerning the source of these intra-interviewer correlations (see Schaeffer et al. 2010 for a recent review), and many have been related to interactions between the interviewer and the respondents, such as differential interviewer probing (e.g., Mangione et al. 1992) or social desirability (Schnell 1997). Factual, non-sensitive survey items (e.g., age) would seemingly be immune to most of these issues, and often show lower intra-interviewer correlations than non-factual and sensitive items (Schnell and Kreuter 2005). However, the existing literature does show evidence of significant interviewer variance in responses to factual items despite interpenetrated designs (see West and Olson 2010 for a review). What has never been tested for face-to-face surveys is the alternative hypothesis that intra-interviewer correlations arise from variance among interviewers in the types of respondents successfully recruited (i.e., nonresponse error variance), which could provide an explanation for these unusual findings. We aim to test this hypothesis with this study.

Intra-interviewer correlations in survey responses have negative impacts on the quality of survey estimates, which makes the identification of mechanisms introducing the correlations important. Do the values on certain survey variables tend to be correlated within an interviewer due to sample assignment, the similarity of sample units recruited by an interviewer, or systematic measurement problems associated with an interviewer? Whatever its source, the intra-interviewer correlation has a multiplicative effect on the variance of an estimated mean in surveys that either achieve a 100% response rate, or have a plausible *missing completely at random* mechanism, where nonrespondents are a random subsample of the full sample. This effect is sometimes referred to as an *interviewer effect* (Davis and Scott 1995). This multiplicative interviewer effect on the variance is written as $1 + (\bar{m} - 1)\rho_{int}$, where $\bar{m}$ represents the average sample workload for an interviewer and ρ_{int} represents an item-specific intra-interviewer correlation (Kish 1962). This effect, which is similar in definition to (and often larger than) the *design effect* due to cluster sampling (Davis and Scott 1995; Schnell and Kreuter 2005), is an undesirable product of the data collection process that can reduce the precision of survey estimates, limit the power of statistical analyses based on survey data, and increase costs of data collection, due to the need for a larger sample size to offset the losses in precision.

Unfortunately, face-to-face surveys rarely (if ever) achieve a 100% response rate, and interviewers play a key role in the recruitment of respondents. Repeatedly, interviewers have been shown to have differential recruitment success in face-to-face surveys; that is, response rates have consistently been shown to vary across interviewers (Wiggins et al. 1992; Morton-Williams 1993; Snijkers et al. 1999; Campanelli and O'Muircheartaigh 1999; O'Muircheartaigh and Campanelli 1999; Pickery and Loosveldt 2002; Hox and de Leeuw 2002; Durrant et al. 2010). In light of these findings, West and Olson (2010) recently analyzed data from a telephone survey (the Wisconsin Divorce Study, or WDS) to test a hypothesis that variation across interviewers arose due to *nonresponse error variance*, or variance across interviewers of the nonresponse biases specific to each interviewer's subsample of cases when he or she achieves less than 100 percent response rates, rather than measurement error variance.

These authors found empirical support for the nonresponse error variance hypothesis for certain survey items. For example, in a CATI setting, with interpenetrated assignment of sample cases approximated (and demonstrated empirically) by performing separate analyses for different calling shifts, the mean ages at times of divorce for WDS respondents (known from official records) were found to vary significantly across interviewers, and nearly 81% of the total interviewer variance in the reported age at divorce values arose from this nonresponse error variance. For some other items (e.g., months since divorce), measurement error variance among interviewers made up the vast majority of the total interviewer variance. The findings of West and Olson therefore suggested that interviewers may actually recruit respondents with different features according to certain variables (e.g., ages), and then measure them appropriately on those variables.

More generally, the ability to decompose interviewer variance into its various sources due to differential nonresponse error and measurement error across interviewers, given the right data set, provides survey researchers with a natural opportunity to simultaneously investigate two important error sources that define the larger *Total Survey Error* (TSE) framework, and recent publications in TSE have called for more investigations of this type (Biemer 2010; Groves and Lyberg 2010, p. 871). To date, the nonresponse error variance hypothesis tested by West and Olson (2010) has not been tested using data from a face-to-face survey. With the present study, we aim to test such a hypothesis, and contribute another empirical examination of simultaneous error sources to this growing area of research in Total Survey Error.

The literature on interviewer effects has shown that interviewer effects are much smaller in telephone surveys than in face-to-face surveys (Groves and Magilavy 1986), motivating additional research of this phenomenon in a personal interview setting. In face-to-face surveys, the social distance between the interviewer and the respondent is much smaller than for telephone surveys, introducing more opportunities for complex respondent-interviewer interactions than would be possible over the telephone. Existing studies have shown that respondents to telephone surveys are more likely to satisfice, less engaged and cooperative, more suspicious about the interview process and confidentiality, and more likely to present socially desirable responses (Beland and St. Pierre 2008; Holbrook et al. 2003). Interviewers in face-to-face surveys can address these concerns in a more personal manner, and differential ability to address these concerns and/or issues could impact both decisions to participate and measurement errors in a differential manner across interviewers, resulting in greater interviewer variance (Brunton-Smith et al. 2012). Differences in nonresponse error variance across interviewers may be one of the primary contributors to the unexpected interviewer variance in more objective survey items (such as age) that has been reported previously for face-to-face surveys, given that the interviewer plays a larger and more personal role in securing cooperation and establishing rapport with the respondent in these surveys. However, no existing studies have demonstrated this empirically, and we aim to contribute to this gap in the literature with the present study.

Testing the nonresponse error variance hypothesis using data from a telephone survey is also difficult due to cases being worked by multiple interviewers. Face-to-face surveys offer an advantage in this respect, in that cases are typically only worked by one interviewer, and there are typically few or no refusal conversion activities. Unfortunately, the interviewer-related nonresponse error variance hypothesis cannot be tested in the absence of interpenetrated assignment of subsamples to interviewers. Interpenetrated assignment is much more difficult in personal interview surveys, primarily for cost reasons: Interviewers are often assigned to work in a single primary sampling unit, or PSU. The availability of records containing "true" values for key survey items for both the entire sample and the respondents, however, would enable estimation of interviewer variance in the:

(1) means of true values for assigned sample units;
(2) nonresponse errors, or differences between the means of true values for respondents and the means of assigned sample values for each interviewer; and
(3) mean response deviations, or mean differences between reported values and true values for each interviewer.

The additional contributions of the last two variance components to total interviewer variance, above and beyond any variance that may be attributed to sampling or PSU-level features, can still be examined even in the absence of truly interpenetrated assignment of subsamples. The objective of this study is to estimate these components of interviewer variance in a face-to-face survey conducted in Germany, where a sample was drawn from administrative records containing true values for selected survey variables.

2. Data and Methods

2.1. Overview of the Labour Market and Social Security (PASS) Study

The Labour Market and Social Security (PASS, or "Panel Arbeitsmarkt und Soziale Sicherung" in German) study is a panel survey conducted by the Institute for Employment Research (IAB) in Nuremberg, Germany, and uses both CAPI and CATI interviewing techniques to collect labor market, household income, and unemployment benefit receipt data from a nationally representative sample of the German population, covering more than 12,000 households annually. This study has a stated purpose of providing "a new database which will allow social processes and the nonintended side-effects of labour market reforms to be assessed empirically" (http://www.iab.de/en/befragungen.aspx). Two samples of roughly equal size (initially 6,000 households each) have data collected from them annually: a sample of households receiving unemployment benefits (as recorded in registers of the German Federal Employment Agency), hereafter referred to as the UB sample, and a sample of households from the general German population with low-income households oversampled, hereafter referred to as the GP sample. The UB sample is refreshed each year by a sample of new entries to the population, and the GP sample was refreshed in the fifth wave (2011) with a cluster (municipality) sample from the German population register. Due to the availability of administrative information for the UB samples, we focus exclusively on selected UB samples from PASS for this study.

To date, five waves of PASS data collection have been completed (2006–2011). Using the AAPOR RR1 calculation, response rates at the household level in the initial wave of each UB sample have varied between 26.3% for the refreshment UB sample in Wave 2 (Gebhardt et al. 2009) and 31.3% for the refreshment UB sample in Wave 3 (Büngeler et al. 2010). While these response rates could be seen as relatively low due to the difficulty of collecting data in a face-to-face setting from this particular population (recipients of unemployment benefits), we note that response rates for face-to-face surveys in Germany are generally lower than in the U.S. and most northern European countries. For example, response rates to the European Social Survey (ESS) have varied between 30% and 56% across waves in Germany (see http://ess.nsd.uib.no and Schnell 1997). In addition, strong refusal conversions are uncommon in Germany (Schnell 2012, p. 223).

The relatively low response rates in PASS and possible interviewer variance in the response rates have the potential to introduce large amounts of nonresponse error variance across interviewers, but recent studies have shown that nonresponse rates and nonresponse errors tend to have a very weak association (Groves and Peytcheva 2008), and the relative contributions of nonresponse error variance and measurement error variance to total interviewer variance have never been demonstrated empirically in a face-to-face survey, let alone in face-to-face surveys with different response rates. Replications of this study using face-to-face surveys with higher response rates will certainly be needed in the future to see whether the same patterns of results emerge. For additional details on the general design of the PASS study, readers can refer to Christoph et al. (2008) or Trappmann et al. (2010). Additional sampling details can be found in the German publication by Rudolph and Trappmann (2007).

In the PASS study, as in many large area probability sample surveys, CAPI interviewers are assigned to work in a single sampling point (or primary sampling unit). This introduces concerns regarding lack of interpenetration that we will address in our data analysis. To avoid interviewer effects that may arise due to repeated interviews conducted with the same household in the panel sample, we consider all sample cases attempted using CAPI in the original Wave 1 sample, and only refreshment sample households attempted using CAPI in Wave 2. At the time of this writing, the Wave 3 data lacked sufficient contact protocol data to be included in the analyses presented here, Wave 4 administrative data were not yet available, and Wave 5 data collection was still ongoing.

2.2. *Data Set Construction*

Before we describe the specific PASS data sets that we constructed for this investigation, it is worth describing the "ideal" data set that would be needed for this type of analysis. This ideal data set, collected in a face-to-face survey, would include:

- A variable containing interviewer ID codes, with a "large" random subsample of all sampled units assigned to each of a "large" number of interviewers (i.e., interpenetrated sampling).
- A response indicator, equal to 1 for sampled units cooperating with the survey request and 0 otherwise, with non-contacts preferably separated from refusals.
- A "large" number of respondents (say, at least 20; see Hox 1998), nested within each of a "large" number of interviewers (say, at least 50) and providing responses on a variety of objective and subjective variables.
- True values for selected objective (and possibly subjective) variables for the *full sample*, which are required for estimation of nonresponse error variance across interviewers, linked from administrative records or some other external source of record information, and

- Covariates describing characteristics of the geographic areas from which the respondents were sampled and where the interviewers are assigned to work.

We note that the availability of "true" values for selected survey variables for the full sample, from administrative records or some other external source, enables computation of nonresponse errors for each interviewer, in addition to the measurement errors associated with reported values on the survey variables. Given these data, one could then analyze the interviewer variance in the nonresponse errors and the measurement errors simultaneously.

At the time of this writing, we are not aware of the existence of any such "ideal" data sets for potential secondary analyses. The design of a study producing such a data set would generally be cost prohibitive, but this description represents a target for future studies. We therefore attempted to construct an approximation of this ideal data set using available information from the first two waves of the PASS data. For purposes of the analyses presented in this article, we constructed a pooled data set of households ever attempted using CAPI in the first two PASS waves. We focus specifically on the heads of households in the UB samples in these two waves, because the "true" values on several PASS study variables are only available for households that were sampled from the administrative records of unemployment benefit recipients. Additional persons responding to the survey in these households may have been recruited by the head of the household rather than the interviewer, so we do not include these individuals, who are also less likely to have administrative records, in the data sets.

The resulting data set contains 4,829 heads of households nested within 211 professional CAPI interviewers, amounting to nearly 23 households per interviewer, along with the following variables:

- The ID of the CAPI interviewer working the sampled household.
- A response indicator (1 = household responded; 0 = nonrespondent).
- "True" values representing official administrative data for the *heads of households* on eleven variables; these variables are extracted from the Integrated Employment Biographies (IEB) data, which are provided by the Research Data Center (FDZ) of the German Federal Employment Agency, and include:
  - Age (based on the date of birth in the IEB data and the date of the interview; for nonrespondents, the median interview date of the respondents in the same sample release is used).
  - Gender.
  - Foreign nationality status.
  - An indicator of whether the household was receiving unemployment benefits at the time of the interview.
  - Employment status at the time of the interview (used to create indicators of employment and unemployment), and
  - Five indicators for educational background and vocational status, including 1) no educational degree up to intermediate secondary school degree, without vocational training; 2) up to intermediate secondary school degree, together with completed vocational training; 3) completed secondary school (entry level for university); 4) technical college or other college for applied sciences; and 5) university degree.

- Aggregate information available from the administrative employment and unemployment records for the postal code areas in Germany, which are the PSUs for the PASS sample; following Bauer et al. (2011a, 2011b), these PSU-level covariates included:
    - Local Unemployment Rate (%).
    - Local Long-Term Unemployment Rate (%).
    - Size of workforce (employees subject to social security and the registered unemployed).
    - Percent of workers with a college or university degree.
    - Percent of workforce that is foreign workers.
    - Percent of workforce that is untrained workers.
    - Percent of workers age 20–30.
    - Percent of workers age 50–65.

Given the area probability sample selected for the PASS study, interpenetrated assignment of sampled households to interviewers was not cost efficient, and a single interviewer was typically assigned to work in a given primary sampling area. We therefore include fixed effects of the PSU-level covariates described above in the models used for the eleven PASS variables in our interviewer variance decomposition analyses, to account for any interviewer variance that may arise from variance among PSUs in these characteristics. While we feel that these PSU characteristics are relevant for the eleven specific survey variables analyzed in this study, there could certainly be additional PSU characteristics not available to us that introduced variance in the features of sample assignments across PSUs, and we acknowledge this as a limitation of our approach. Future research needs to consider more elegant techniques for estimating interviewer variance in non-interpenetrated designs (see Biemer 2010), and we expand on this suggestion in our Discussion.

We also constructed a second data set containing survey reports on the eleven variables mentioned above for the 1,472 heads of households cooperating with the PASS study request, and including interviewer ID codes and the PSU covariates. As noted earlier, response rates to face-to-face surveys in Germany tend to be lower than those in the United States and other northern European countries, and were likely lower in this case because we are working with the UB samples. This limited the power of our multilevel analyses of the respondents, as there were only about seven respondents per interviewer. However, we still find evidence of significant total interviewer variance components based on analyses of the respondents, which will be described in the upcoming Results section. This second data set did not include true values of the survey variables from the administrative data; this kind of linking required respondent consent, and only about 80 percent of the PASS respondents consented. As a consequence, we could not compute response deviations for the responding households without reducing the size of the respondent subset.

Item-missing data among the respondents were considered on a variable-by-variable basis, and heads of households were treated as nonrespondents if failing to provide data on a particular variable. Based on the item-missing data patterns, only interviewers with at least two survey respondents for a particular variable were considered in the eventual interviewer variance decomposition analyses, given that interviewer effects cannot be estimated based on a sample of size 1. As a result, the number of interviewers and the number of full sample and responding cases vary slightly depending on the variable being analyzed. We acknowledge that future research in this area should aim to use larger samples in face-to-face surveys with higher response rates to minimize concerns about statistical power when analyzing interviewer variance among the respondents.

All analyses of these data were conducted in the IAB Research Data Center (RDC) at the University of Michigan-Ann Arbor (Bender and Heining 2011).

2.3. Statistical Analysis

Initial descriptive analyses focused on computation of response rates (using the AAPOR RR1 calculation, and including noncontact as a form of nonresponse) for each of the 211 CAPI interviewers in the pooled PASS data set, and plotting the distribution of the response rates. The intra-interviewer correlation in the response indicators was estimated by using the xtmelogit command in Stata (Version 12.1) to fit a simple multilevel logistic regression model to the PASS response indicator, including a fixed intercept and random interviewer effects.

For each of the eleven analysis variables described in Section 2.2, the three-step estimation methodology described by West and Olson (2010) was then applied to decompose total interviewer variance into variance in means of true values for full subsamples, nonresponse error variance, and measurement error variance. Let y_{ij} be the true value of survey variable Y for sample unit j assigned to interviewer i, and let x_{ij} be the reported value for Y for that sample unit if the unit responds to the survey. Further, let R denote respondents, and NR denote nonrespondents. Assuming interpenetrated assignment of subsamples to interviewers, the expectation of the mean of respondent reports $\bar{x}_i$ for interviewer i can be written as

$$E(\bar{x}_i|i) = \bar{Y} + Bias_{NR,i} + Bias_{ME,i} = \bar{Y} + (\bar{y}_{R,i} - \bar{y}_i) + (\bar{x}_i - \bar{y}_{R,i}) \tag{1}$$

That is, the expected value of the respondent mean for interviewer i is the sum of 1) the population mean of the true values, $\bar{Y}$; 2) the difference between the mean of the true values for all *respondents* interviewed by interviewer i, $\bar{y}_{R,i}$, and the mean of the true values for all *units* assigned to interviewer i, $\bar{y}_i$ (nonresponse error); and 3) the difference between the mean of the *reported* values for all respondents interviewed by interviewer i, $\bar{x}_i$, and $\bar{y}_{R,i}$ (measurement error).

Assuming interpenetration and negligible covariance between the two bias terms, the variance of the expectation in (1) is defined by the sum of two variance components: $Var(Bias_{NR,i})$ and $Var(Bias_{ME,i})$. We were unable to compute the empirical correlation of these two error sources in the PASS data set, given that not all respondents consented to having their administrative data *linked* to their survey reports. West and Olson (2010) suggest that these correlations are small in a CATI survey, but the correlations could be different in a face-to-face survey; more work is certainly needed to examine this assumption. We sought unbiased estimates of these two variance components along with the relative proportions of the total interviewer variance contributed by these two sources of variance among interviewers. Estimation of these variance components using closed-form estimators is possible for very simple design and response scenarios (e.g., equal assignment sizes and response rates across interviewers) typically not experienced in practice. Given the unequal assignment sizes and respondent counts for each CAPI interviewer in the PASS survey, we estimated these components for each variable using three distinct steps. The three steps below apply to the continuous variable *age*, and outline the general analytic approach; subsequent remarks address the estimation steps for the ten binary indicators. Item nonresponse on each variable introduced slight differences in terms of the respondent sample sizes for the variables.

Estimation Step 1: Estimate the interviewer variance in the means of the true values for the assigned sample cases. First, we estimated the variance among interviewers in the means of the true values for all CAPI sample cases assigned to each interviewer. Assuming interpenetrated assignment of cases to interviewers, this variance component should be negligible, as all interviewers should have a full sample mean of true values equal, on average, to the population mean. We estimated this component using a one-way random effects model, assuming that the interviewers were a random subsample from a larger hypothetical population of interviewers:

$$y_{ij} = \bar{Y} + b_i + e_{ij} \tag{2}$$

In this notation, y_{ij} is the true value of variable Y for sample unit j assigned to interviewer i, b_i is the random deviation of interviewer i's assignment mean from the population mean of the true values, $\bar{Y}$, and e_{ij} is a normally distributed random error with mean 0 and constant variance (the element variance within each assignment). We estimated the variance of the b_i, or $Var(b_i) = \sigma^2_{\text{int},full}$, using restricted maximum likelihood (REML) estimation to obtain an unbiased estimate of this variance component given unequal interviewer workloads (Patterson and Thompson 1971).

We tested a null hypothesis that the interviewer variance component is equal to zero, $H_0 : \sigma^2_{\text{int},full} = 0$, versus the alternative that assignment means vary across interviewers, $H_A : \sigma^2_{\text{int},full} > 0$; see West and Olson (2010) for a discussion of appropriate methods for testing this null hypothesis. We performed asymptotic likelihood ratio tests of this null hypothesis, referring the standard likelihood ratio chi-square statistic to a mixture of chi-square distributions, with degrees of freedom 0 and 1 and equal weight 0.5 (Zhang and Lin 2008). We report p-values for the test statistics under this null distribution. The xtmixed and xtmelogit commands in Stata facilitate these tests.

Estimation Step 2: Estimate the interviewer variance in the means of the true values for the responding sample cases. Second, when assignments are interpenetrated, each interviewer has the same $\bar{y}_i$ in expectation, and the variance of the nonresponse biases simplifies to $Var(Bias_{NR,i}) = Var(\bar{y}_{R,i} - \bar{y}_i) = Var(\bar{y}_{R,i})$. We thus estimated the nonresponse error variance component by estimating the variance across interviewers in the means of the true values for *respondents*. We again used a one-way random effects model for the true values of *respondents* to the survey request, $y_{R,ij}$:

$$y_{R,ij} = \bar{Y}_R + b'_i + e'_{ij} \tag{3}$$

Here, b'_i captures the random deviation of each interviewer's mean for their recruited respondents' true values from the expected value of the mean of the true values for respondents over all possible sample assignments to interviewers ($\bar{Y}_R$). We note that in the absence of overall nonresponse bias, $\bar{Y}_R$ will be equal to the population mean of the true values, $\bar{Y}$. We estimated the variance of these random effects, $Var(b_i^{l}) = \sigma^2_{\text{int},resp}$, using REML to obtain an unbiased estimate of $Var(Bias_{NR,i})$. We tested this component of variance against zero using the likelihood ratio test described above.

Estimation Step 3: Estimate the interviewer variance in the means of the reported values for the responding sample cases (the total interviewer variance). Third, under an assumption of interpenetrated assignment of subsamples to interviewers, Equation 1 can be rewritten as:

$$E(\bar{x}_i|i) = \bar{Y} + Bias_{NR,i} + Bias_{ME,i} = \bar{Y} + (\bar{y}_{R,i} - \bar{Y}) + (\bar{x}_i - \bar{y}_{R,i}) = \bar{x}_i \tag{4}$$

Using the sample mean of the respondent reports for interviewer i as an estimate of this expectation, we then computed an unbiased estimate of the variance in the means of the reported values across interviewers (Equation 4) using a one-way random effects model and REML estimation:

$$x_{ij} = \bar{X}_R + b_i^{ll} + e_{ij}^{ll} \tag{5}$$

This is the interviewer variance model that is often estimated in practice using respondent data only. The variance of the random interviewer deviations (b_i^{ll}) around the expected value of the mean of the respondent reports over all possible sample assignments to interviewers ($\bar{X}_R$), or $Var(b_i^{ll}) = \sigma^2_{\text{int},resp,obs}$, captures measurement error variance and nonresponse error variance introduced by the interviewers, along with sampling variance in the absence of interpenetration, and is thus the "total variance" due to interviewers. We tested this variance component for significance using appropriate likelihood ratio tests. We then subtracted the estimate of the nonresponse error variance from Estimation Step 2 to get an estimate of the measurement error variance across interviewers. The estimated proportion of variance introduced by interviewers due to nonresponse error variance was then computed as:

$$\frac{\hat{\sigma}^2_{\text{int},resp} - \hat{\sigma}^2_{\text{int},full}}{\hat{\sigma}^2_{\text{int},resp,obs} - \hat{\sigma}^2_{\text{int},full}} \tag{6}$$

Evidence of successful interpenetration from Estimation Step 1 implies that $\sigma^2_{\text{int},full} = 0$ (i.e., interviewer-level means of true values for their full assignments do not vary). We subtracted estimates of $\sigma^2_{\text{int},full}$ from the numerator and denominator to remove components of variance that were not due to the interviewer from this calculation.

Given that ten of the eleven PASS variables were binary in nature ($1 = $ yes, $0 = $ no), multilevel logistic regression models with random interviewer effects were fitted in each estimation step to estimate the components of variance due to interviewers in the *log-odds* of each binary variable being equal to 1, rather than variance in the means of outcomes in the standard one-way random effects models for the age variable, which followed a roughly normal distribution. These models were fitted using approximate maximum likelihood methods, based on the adaptive Gauss-Hermite approximation to the log-likelihood of the model, as implemented in the xtmelogit command in the Stata software.

Finally, to address the confounding of interviewer ID and PSU (or sampling area) in the PASS data set, we repeated the analyses described above after including fixed effects of the PSU-level covariates in the models used for each analysis step. Estimated components of variance at each analysis step may arise due to variance in PSU-level features rather than interviewers; for example, nonresponse error variance among interviewers for a specific variable may be due to between-PSU variance in features that are correlated with both the variable of interest and response propensity. These analyses were performed to estimate components of variance due to the interviewers after removing any sources of interviewer variance (at any step) due to between-PSU variance in the values of the available PSU-level covariates.

3. Results

Figure 1 below presents a histogram showing the distribution of unweighted response rates across the 211 CAPI interviewers.

Figure 1 illustrates the substantial variance among these interviewers in terms of their unweighted response rates, which is consistent with the existing literature (Wiggins et al. 1992; Morton-Williams 1993; Snijkers et al. 1999; Campanelli and O'Muircheartaigh 1999; O'Muircheartaigh and Campanelli 1999; Pickery and Loosveldt 2002; Hox and de Leeuw 2002; Durrant et al. 2010) and provides motivation for the hypothesis being examined in this study. The estimated intra-interviewer correlation in the binary response indicators for the sampled households assigned to the 211 interviewers was computed as follows:

$$\hat{\rho}_{\text{int},R} = \frac{\hat{\sigma}^2_{\text{int},R}}{\hat{\sigma}^2_{\text{int},R} + \pi^2/3} = \frac{0.385}{0.385 + \pi^2/3} = 0.105$$

The estimated between-interviewer variance in the log-odds of responding (0.385) was found to be significantly greater than zero based on an asymptotic likelihood ratio test

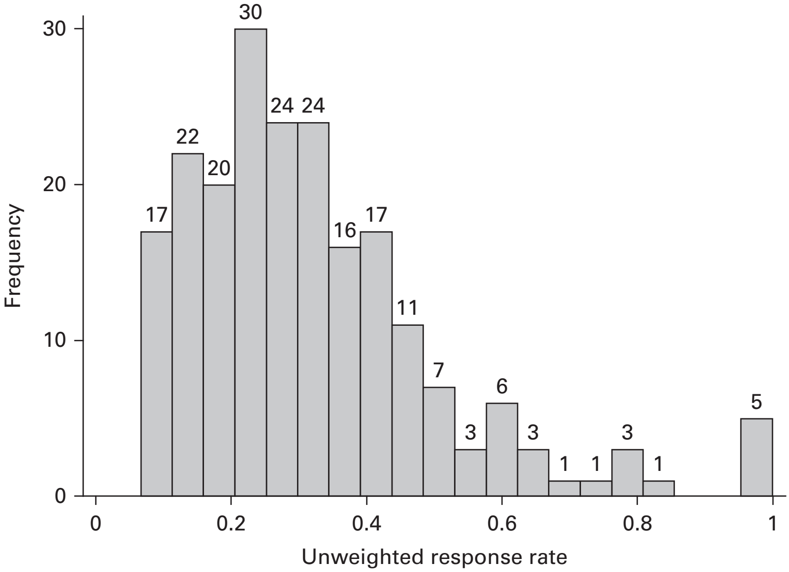


Fig. 1. Distribution of unweighted response rates across the 211 CAPI interviewers in PASS Waves 1 and 2

using a mixture of chi-square distributions ($p < .001$), providing strong support for the conclusion that the within-interviewer correlation in response indicators was significant. The existence of significant between-interviewer variance in response rates introduces the possibility of varying nonresponse error between interviewers, which we examine for the eleven PASS variables below. The estimated interviewer variance component for the log-odds of responding dropped to 0.343 after including the fixed effects of all of the PSU-level covariates in the logistic regression model, meaning that the PSU-level features accounted for about 10.9% of the between-interviewer variance in the response rates, but this component of variance remained significantly greater than zero. Small amounts of item-missing data on each of these eleven survey variables did not substantially alter the interviewer variance components for the response indicators when repeating the same analysis on a variable-by-variable basis.

Previous studies of interviewer effects on survey variables (Schnell and Kreuter 2005) have noted that interviewer variance components tend to be larger than variance components due to sampling areas. The findings above, which show significant interviewer variance in the response rates, suggest that while the same may be true in terms of response indicators, it would still be important to adjust for the relationships of area-level features with response before inferring how much of the variability in cooperation rates is truly due to interviewers. We find that there is still variance among interviewers in response probabilities, even after accounting for the relationships of the PSU-level covariates with the probability of response.

We now turn to our interviewer variance decomposition analyses. Table 1 below provides estimates of the variance components of interest at each step of the estimation process. Table 1 also provides estimates of intra-interviewer correlations ($\hat{\rho}_{\text{int}}$) based on the estimated total interviewer variance components, computed using the respondent reports.

Examining the estimates of the total interviewer variance components in row 8 of Table 1, we find evidence that five of the total interviewer variance components are greater than zero (using $p < .10$ for marginal significance, and $p < .05$ for significance) after adjusting for fixed effects of the PSU covariates: age of household head, household receipt of unemployment benefits, employment of household head, having up to an intermediate secondary degree together with vocational training, and having a degree from a technical college or from another college for applied sciences. Notably, given the fairly low response rates for the households in the UB samples in Waves 1 and 2 of the PASS survey, we have limited power to detect significant components of variance due to the interviewers based on the *respondent* data. We have close to 23 sampled households per CAPI interviewer in the full sample (row 3 of Table 1), and only about seven responding households per interviewer in the respondent subset (row 6 of Table 1). We therefore have adequate power for detecting moderate interviewer variance components based on the full sample, having more than 50 interviewers with more than 20 households per interviewer (see Hox 1998), but limited power based on the respondent subset. Despite this limitation, we still find five significant total interviewer variance components, and our research interest lies in the sources of these significant variance components. We focus on the estimates and tests of significance in Table 1 *after* adjusting for the effects of the PSU-level covariates, to account for any variance among interviewers arising from the areas of Germany in which they are working.

Table 1. Decompositions of interviewer variance components for eleven PASS survey variables with available validation data, with estimated interviewer variance components presented before and after adjustment for selected PSU-level covariates

								Education and Vocational Training				
		Age	Benefit Receipt	Gender (Female)	Foreign Status	EMP	UNEMP	Up to Inter-mediate Sec Degree, no VOC	Up to Inter-mediate Sec Degree, VOC	SEC School, Complete	TECH College	UNIV
1	Interviewers	210	209	211	210	211	211	211	211	211	211	211
Step 1: Full Sample True Values												
2	Full sample *n*	4790	4787	4803	4801	4803	4803	4803	4803	4803	4803	4803
3	Full *n*/Int.	22.8	22.9	22.8	22.9	22.8	22.8	22.8	22.8	22.8	22.8	22.8
4	$\hat{\sigma}^2_{\text{int},full}$	3.03***/ 2.12***	0.07***/ 0.04*	0.04**/ 0.02	0.49***/ 0.02	0.06***/ 0.04**	0.19***/ 0.07***	0.26***/ 0.11***	0.20***/ 0.12***	0.18***/ 0.10**	<0.01/ <0.01	0.61***/ 0.22***
Step 2: Respondent True Values												
5	Respondent *n*	1377	1351	1383	1378	1383	1383	1368	1368	1368	1368	1368
6	Resp. *n*/Int.	6.6	6.5	6.6	6.6	6.6	6.6	6.5	6.5	6.5	6.5	6.5
7	$\hat{\sigma}^2_{\text{int},resp}$	4.62**/ 3.54*	0.31***/ 0.15	0.06/ 0.01	0.48***/ 0.05	<0.01/ <0.01	0.79***/ 0.44***	0.42***/ 0.17**	0.33***/ 0.19***	0.08/ <0.01	<0.01/ <0.01	1.53***/ 1.00***
Step 3: Respondent Reported Values												
8	$\hat{\sigma}^2_{\text{int},resp,obs}$	5.37**/ 3.58*	0.44***/ 0.33***	0.08*/ 0.03	0.64***/ 0.09	0.51***/ 0.32**	0.12**/ <0.01	0.31***/ 0.08	0.30***/ 0.13**	0.35**/ 0.13	1.56***/ 1.59**	0.30/0.09
9	$\hat{\rho}_{\text{int}}$	0.03	0.09	0.01	0.03	0.09	<0.01	0.02	0.04	0.04	0.33	0.03

NOTES: Cells presenting estimated variance components (rows 4, 7, and 8) include the estimates and tests of significance before/after adjustment for fixed effects of PSU-level covariates. Likelihood Ratio Test results: *** $p < .01$, ** $p < .05$, * $p < .10$. Variable Abbreviations: EMP = Employed; UNEMP = Unemployed; VOC = Vocational Training; SEC = Secondary; TECH = Technical; UNIV = University.

First considering the age variable, we see evidence of between-interviewer variance in the average ages of the heads of households for the full sample assignments, both before and after controlling for fixed effects of the available PSU-level covariates in the multilevel model. This apparent lack of interpenetrated assignment of households to interviewers based on the ages of the heads of households precludes use of the aforementioned three-step estimation methodology to decompose the interviewer variance. Nevertheless, we note a 67% increase in the between-interviewer variance of the true age values when considering respondents only, from an estimate of 2.12 for the full sample to an estimate of 3.54 for the respondent subset (controlling for the PSU-level covariates), followed by a very slight 1% increase in the between-interviewer variance when considering the *reported* age values, rather than the true age values, from an estimate of 3.54 to an estimate of 3.58. While we cannot definitively decompose the total interviewer variance component for age based on these estimates, due to the lack of interpenetrated assignments, these findings suggest that the marginally significant ($p < .10$) total interviewer variance for the ages of the heads of households arises from a combination of variance in the full sample assignments *and* variance in the ages of the household heads for the recruited respondents, rather than variance in the measurement errors. This is to be expected for a very objective survey item such as age, where extensive cognitive processing or interviewer clarification is generally not needed.

We see different results for the survey variable measuring current household receipt of unemployment benefits (yes/no). Analyzing the true values of this indicator variable for the sampled households assigned to each interviewer, we find marginal evidence of variance among interviewers ($p < .10$) after controlling for the fixed effects of the PSU-level covariates, suggesting that interpenetrated assignment of households is at least a plausible assumption after accounting for the PSU-level covariates. For this variable, the estimate of the total interviewer variance component is 0.33 ($p < .001$). Further assuming negligible correlations of the interviewer-specific nonresponse errors and measurement errors, which we cannot estimate using this data set, we can subtract the small portion of the variance due to the assigned samples (0.04) from the estimated interviewer variance in true values among respondents (0.15, resulting in 0.11) and from the estimated total variance (0.33, resulting in 0.29). We can then estimate that $0.11/0.29 = 0.38 = 38\%$ of the additional variance in unemployment benefit receipt indicator values introduced by the interviewers is due to nonresponse error variance, with the remaining portion (62%) due to measurement error variance. These results suggest that interviewers are tending to recruit households with different current benefit receipt status, and then (potentially) having differential difficulty accurately measuring this variable in the interview (to a greater extent than their respondent pools are differing). After adjusting for the PSU-level covariates, the total interviewer variance for this variable corresponds to an estimated intra-interviewer correlation of $\hat{\rho}_{\text{int}} = 0.33/(0.33 + \pi^2/3) = 0.09$.

We find yet another pattern of results when considering the employment indicator for the head of the household. We only find weak evidence of interpenetrated assignment of sampled households to interviewers, with the variance in the proportions of households with the heads employed among the full interviewer samples significant at the 0.05 level after accounting for the fixed effects of the PSU-level covariates. While this once again precludes the use of our decomposition approach in theory, we note that the estimate of between-interviewer variance when considering true values of the employment indicator for respondents only is extremely small. This finding suggests that little of the original variance in the full sample assignments remains, and that the sets of respondents recruited by the interviewers now vary in a way that eliminated the between-interviewer variance in the original assignments. In other words, the proportions of household heads that are employed are similar across interviewers in their recruited respondents. We then find evidence of significant between-interviewer variance in terms of the *reported* employment values for the respondents; the adjusted estimate of the total interviewer variance component (0.32) is actually 700% larger than the original variance component arising from the full sample assignments (0.04). Although the interviewers did appear to be working different samples to begin with, these differences became much more pronounced when considering the reported values among respondents, and these differences did not appear to arise due to differential recruitment. After adjusting for the PSU-level covariates, the total interviewer variance for this variable corresponds to an estimated intra-interviewer correlation of $\hat{\rho}_{\text{int}} = 0.32/(0.32 + \pi^2/3) = 0.09$.

Finally, we find evidence of significant total interviewer variance for two of the indicators of educational level and vocational training of the head of the household. For the variable indicating up to an intermediate secondary degree together with completed vocational training, which is the modal category in this population, we once again find evidence of a lack of interpenetrated assignment of households. After adjusting for the fixed effects of the PSU-level covariates, the estimated interviewer variance in the indicator values for this category increased somewhat when considering true values for respondents only, and then fell back to the level of the original full sample when considering *reported* values for respondents only. Unusual patterns of results like this, where the total interviewer variance is smaller than the variance in the true values for respondents among interviewers, could arise from a *negative covariance* between nonresponse errors and measurement errors among interviewers. For example, interviewers may recruit sets of respondents with different educational profiles (relative to their assigned samples), with certain interviewers recruiting higher than expected proportions of individuals with up to an intermediate secondary degree and completed vocational training (i.e., a positive nonresponse error). However, these interviewers may then collect reported values of education that are lower than the truth (i.e., a negative measurement error), resulting in proportions that are lower than those based on the true values for their subset of recruited respondents. Unfortunately, we are unable to estimate the covariance of these two error sources for all interviewers using the PASS data sets, as the measurement errors require linking of the survey data with the administrative data, and this is only possible if the respondents consent to this process (and not all PASS respondents consented).

When considering the education/vocation indicator for technical college or other colleges for applied sciences, there was evidence of interpenetrated assignment, no evidence of a change in the interviewer variance given the features of the recruited samples, and then a substantial increase in the interviewer variance when considering the reported values on this indicator. These results point to possible differential misunderstanding among the interviewers, the respondents, or both when considering what exactly this category corresponds to.

In general, we note that in nearly all cases in Table 1, the estimated interviewer variance components in the various multilevel models are substantially decreased when controlling for the fixed effects of the PSU-level covariates. This was also noted earlier when examining interviewer variance in the response rates for each individual variable. Schnell and Kreuter (2005) showed that interviewer effects generally tend to be larger than area effects, which would suggest that it is not always necessary to disentangle area effects and interviewer effects when studying interviewer variance. The results in this study indicate a general need to adjust for the effects of area-level covariates when using multilevel models to quantify interviewer variance in a face-to-face survey, where interviewers are not assigned random subsets of the full sample; this is often the case in practice when interviewers only work in a single sampled area. While area effects may be smaller than interviewer effects, they can cause interviewer variance components to seem larger than they really are, and that was noted in this study. In this case, variance in the features of sampled areas could be introducing what seems like variance among interviewers in both the features of the recruited respondents and the collected survey responses.

4. Discussion

This is the first study to examine the contributions of nonresponse error variance and measurement error variance among interviewers to total interviewer variance in a face-to-face survey setting. The presence of administrative records in the PASS study, which collects data annually from a national sample of households receiving unemployment benefits in Germany, enabled the decomposition of total interviewer variance estimates for selected variables into nonresponse error variance and measurement error variance components. We first found significant variance among CAPI interviewers in terms of response rates for newly recruited cases, even after adjusting for the relationships of several features of the sampling areas with the response indicators. This finding is consistent with the prior literature in this area, and provides further evidence that interviewers do tend to vary in terms of their response rates. This widely reported result introduces the possibility of variance among the interviewers in nonresponse errors.

Second, we found significant (or marginally significant) total interviewer variance components for five of eleven PASS variables with available administrative data. For two of these variables, there was strong evidence ($p < .01$) of a lack of interpenetrated assignment of sampled households to the interviewers, even after adjusting for the relationships of PSU-level covariates with these variables, which prevented the application of our decomposition methodology. Nevertheless, we did find evidence that significant interviewer variance in the ages of the heads of the assigned households was substantially increased when considering the responding heads of households, suggesting that the demographic features of recruited respondents may vary among interviewers to an even greater extent than their assigned samples vary. This finding was consistent with the findings of West and Olson (2010), who analyzed interpenetrated assignments in a telephone survey.

For the remaining three variables exhibiting evidence of significant ($p < .05$) interviewer variance in the respondent reports in combination with at least marginal evidence of originally interpenetrated assignments, in terms of the true values on these three variables, we found that the bulk of the total interviewer variance appeared to be driven by measurement error variance. Notably, for the indicator of whether a sampled household is currently receiving unemployment benefits, there appeared to be a mixture of nonresponse error variance and measurement error variance driving the total interviewer variance, with the majority of the variance stemming from measurement error variance. For the remaining two variables, an employment indicator and an indicator of completing technical college, measurement error variance appeared to be the primary source of the interviewer variance. These results therefore suggest that variance among interviewers in measurement difficulties tends to be the primary driver of total interviewer variance, at least for these PASS variables.

This study was certainly not without limitations. First, as indicated above, we were limited in this investigation by a lack of interpenetrated assignment of sampled households to the CAPI interviewers. Even after controlling for the relationships of PSU-level covariates with the variables in question, we still found significant variability among the interviewers in the true values of the assigned sample households for seven of the eleven PASS variables. While this could have been due to measurement error in the administrative records, especially for the education indicators (given the population of interest), this is a general problem with studying interviewer variance in face-to-face surveys. We can monitor changes in interviewer variance components from the full assigned samples to the respondents in a descriptive sense, but the estimation methodology outlined in this article relies on interpenetrated assignments for decomposing the interviewer variance. While the design of a future face-to-face survey of a large population with random subsamples of the full population sample assigned to interviewers would be ideal for future studies of this problem, such a design is generally not an economic reality for most survey organizations. Future research needs to focus on the development of methods for both estimating and decomposing interviewer variance in face-to-face surveys where interviewers were not originally assigned random subsamples of the full sample. For example, subsets of interviewers might be matched in terms of similarities of the sampled areas in which they are working, and then interviewer variance could be estimated based on the random samples assigned to the interviewers in these matched areas. However, interviewer sample sizes could continue to be an issue here.

Second, in the estimation steps presented in this article, we examined components of variance assuming that sampling errors, nonresponse errors, and measurement errors associated with the interviewers were independent. Non-zero covariances between these error sources may be important components of total interviewer variance, but we did not have the type of data needed to estimate these covariances; the survey data could only be linked to the administrative data with consent from the PASS respondents. More empirical work is certainly needed to examine assumptions of negligible covariance between these three error sources, and this work will require data sets where the true values for selected survey variables are available for the entire sample and can be linked to the respondent reports on those survey variables. West and Olson (2010) did examine such correlations among these interviewer-specific error sources and found that they were generally small, but this work needs to be repeated using data from a face-to-face survey.

Third, we had limited statistical power to detect components of variance due to interviewers among respondents, given the relatively low response rates found in the PASS samples of German households receiving unemployment benefits. This tends to be a difficult population to survey, and response rates for face-to-face surveys in Germany also tend to be lower than those found in the U.S. and other northern European countries. The ideal study of this problem would require CAPI interviewing, interpenetrated assignment of large subsamples to a large number of interviewers, a relatively high response rate, and available record values of good quality on respondents and nonrespondents for variables measured in the survey. Although potentially expensive, studies with these features will greatly enhance our understanding of these issues.

There are promising extensions of this work that we leave to future research. At present, estimates of intra-interviewer correlations are based on respondent reports only, and do not recognize possible contributions of nonresponse error variance to interviewer variance. The same is true for estimates of intra-cluster correlations, which are often used to estimate design effects in complex area probability samples. Collectively, the empirical work in this study and the study by West and Olson (2010) provide motivation for the analytical development of estimators of intra-interviewer (and intra-cluster) correlations that recognize contributions of sampling variance, refusal error variance, noncontact error variance, and measurement error variance to the total variance among interviewers (or clusters). Existing work by Biemer (1980) and Groves and Magilavy (1984) provides some possible avenues for this analytical development.

Alternative applications of multilevel modeling may also prove useful for studying this problem in the future. For example, given a data set with true values available for a full sample, a multilevel modeling framework incorporating multiple imputations of measurement errors for survey nonrespondents would enable estimation of interviewer variance components due to nonresponse error variance, via random interviewer coefficients for a response indicator with possible values $1 =$ respondent, $-1 =$ nonrespondent (predicting actual and imputed respondent reports), and measurement error variance, via random interviewer intercepts, in addition to the *covariance* of these two errors sources at the interviewer level. Future applications of these types would also enhance our understanding of these issues.

Finally, if administrative records are available for selected items, field supervisors could monitor empirical best linear unbiased predictors (EBLUPs) of the random interviewer effects in the models that we propose, and identify interviewers with extremely unusual random effects in the models for either nonresponse error variance or total variance. For example, the EBLUPs computed from the nonresponse error variance estimation step for the active interviewers in a given survey could be ordered from highest to lowest, where the highest values for a particular survey variable would indicate interviewers with mean values for respondents that are substantially higher than the mean values for their assigned sample; this would mean that they only tend to recruit respondents with higher values on the particular variable. Interviewers having one of the five highest or one of the five lowest EBLUPs could have their recruited cases examined in more detail to see whether the supervisors would need to intervene and alter the recruiting strategies being used by these interviewers. Future studies could then evaluate the ability of this type of intervention to reduce total interviewer variance.

5. References

Bauer, T.K., Fertig, M., and Vorrell, M. (2011a). Neighborhood Effects and Individual Unemployment. SOEP paper 409, DIW Berlin.

Bauer, T.K., Flake, R., and Sinning, M.G. (2011b). Labor Market Effects of Immigration: Evidence from Neighborhood Data. Ruhr Economic Papers #257, Bochum, Dortmund, Duisburg, Essen.

Beland, Y. and St-Pierre, M. (2008). Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI. Advances in Telephone Survey Methodology, Chapter 14, J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, R.L. Sangsler (eds). New York: Wiley.

Bender, S. and Heining, J. (2011). The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing. FDZ Methodenreport, 07/2011, Nürnberg.

Biemer, P.P. (1980). A Survey Error Model which Includes Edit and Imputation Error. Proceedings of the American Statistical Association, Section on Survey Research Methods, 616–621.

Biemer, P.P. (2010). Total Survey Error: Design, Implementation, and Evaluation. Public Opinion Quarterly, 74, 817–848.

Biemer, P.P. and Stokes, S.L. (1991). Approaches to the Modeling of Measurement Error. Measurement Errors in Surveys, P.P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley.

Biemer, P.P. and Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data. In Survey Measurement and Process Quality, L. Lyberg, P.P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley-Interscience.

Brunton-Smith, I., Sturgis, P., and Williams, J. (2012). Is Success in Obtaining Contact and Cooperation Correlated with the Magnitude of Interviewer Variance? Public Opinion Quarterly, 76, 265–286.

Büngeler, K., Gensicke, M., Hartmann, J., Jäckle, R., and Tschersich, N. (2010). IAB-Haushaltspanel im Niedrigeinkommensbereich Welle 3 (2008/2009). Methoden- und Feldbericht. FDZ Methodenreport, 10/2010, Nürnberg.

Campanelli, P. and O'Muircheartaigh, C. (1999). Interviewers, Interviewer Continuity, and Panel Survey Nonresponse. Journal of Official Statistics, 2, 303–314.

Christoph, B., Müller, G., Gebhardt, D., Wenzig, C., Trappmann, M., Achatz, J., Tisch, A., and Gayer, C. (2008). Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Volume 1: Introduction and Overview, Wave 1 (2006/2007). FDZ Datenreport, 05/2008, Nürnberg.

Collins, M. and Butcher, B. (1982). Interviewer and Clustering Effects in an Attitude Survey. Journal of the Market Research Society, 25, 39–58.

Davis, P. and Scott, A. (1995). The Effect of Interviewer Variance on Domain Comparisons. Survey Methodology, 21, 99–106.

Durrant, G.B., Groves, R.M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. Public Opinion Quarterly, 74, 1–36.

Fellegi, I.P. (1964). Response Variance and its Estimation. Journal of the American Statistical Association, 59, 1016–1041.

Freeman, J. and Butler, E.W. (1976). Some Sources of Interviewer Variance in Surveys. Public Opinion Quarterly, 40, 79–91.

Gebhardt, D., Müller, G., Bethmann, A., Trappmann, M., Christoph, B., Gayer, C., Müller, B., Tisch, A., Siflinger, B., Kiesl, H., Huyer-May, B., Achatz, J., Wenzig, C., Rudolph, H., Graf, T., and Biedermann, A. (2009). Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Volume 1: Introduction and Overview, Wave 2 (2007/2008). FDZ Datenreport, 06/2009, Nürnberg.

Groves, R.M. (2004). The Interviewer as a Source of Survey Measurement Error. Survey Errors and Survey Costs, (Second Edition). New York: Wiley-Interscience.

Groves, R.M. and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74, 849–879.

Groves, R.M. and Magilavy, L.J. (1984). An Experimental Measurement of Total Survey Error. Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Survey Research Methods, 698–703.

Groves, R.M. and Magilavy, L.J. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50, 251–266.

Groves, R.M. and Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. Public Opinion Quarterly, 72, 167–189.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1960). Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 32nd Session, Vol. 38, Part 2, 359–374.

Holbrook, A.L., Green, M.C., and Krosnick, J.A. (2003). Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Bias. Public Opinion Quarterly, 67, 79–125.

Hox, J.J. (1998). Multilevel Modeling: When and Why. In Classification, Data Analysis, and Data Highways, I. Balderjahn, R. Mathar, and M. Schader (eds). New York: Springer.

Hox, J.J. and de Leeuw, E.D. (2002). The Influence of Interviewers' Attitude and Behavior on Household Survey Nonresponse: An International Comparison. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, 92–115.

Mangione, T.W., Fowler, F.J., and Louis, T.A. (1992). Question Characteristics and Interviewer Effects. Journal of Official Statistics, 8, 293–307.

Morton-Williams, J. (1993). Interviewer Approaches. Aldershot: Dartmouth Publishing Company Limited.

O'Muircheartaigh, C. and Campanelli, P. (1998). The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. Journal of the Royal Statistical Society, Series A, 161, 63–77.

O'Muircheartaigh, C. and Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. Journal of the Royal Statistical Society, Series A, 162(Part 3), 437–446.

Patterson, H.D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika, 58, 545–554.

Pickery, J. and Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse. Quality and Quantity, 36, 427–437.

Rudolph, H. and Trappmann, M. (2007). Design und Stichprobe des Panels "Arbeitsmarkt und Soziale Sicherung" (PASS). In Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung", (ed.) M. Promberger, IAB-Forschungsbericht, 12/2007, Nürnberg.

Schaeffer, N.C., Dykema, J., and Maynard, D.W. (2010). Interviewers and Interviewing. In Handbook of Survey Research, J.D. Wright and P.V. Marsden (eds). (Second Edition). Bingley, U.K. Emerald Group Publishing Limited.

Schnell, R. (1997). Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen. Opladen: Leske + Budrich.

Schnell, R. (2012). Survey-Interviews. Standardisierte Befragungen in den Sozialwissenschaften. Wiesbaden: VS-Verlag.

Schnell, R. and Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects. Journal of Official Statistics, 21, 389–410.

Snijkers, G., Hox, J.J., and de Leeuw, E.D. (1999). Interviewers' Tactics for Fighting Survey Nonresponse. Journal of Official Statistics, 15, 185–198.

Trappmann, M., Gundert, S., Wenzig, C., and Gebhardt, D. (2010). PASS: a Household Panel Survey for Research on Unemployment and Poverty. Proceedings of the American Statistical Association, Section on Survey Research Methods, 130, 609–622.

West, B.T. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? Public Opinion Quarterly, 74, 1004–1026.

Wiggins, R.D., Longford, N.T., and O'Muircheartaigh, C.A. (1992). A Variance Components Approach to Interviewer Effects. In Survey and Statistical Computing, A. Westlake, R. Banks, C. Payne, and T. Orchard (eds). Amsterdam: North-Holland.

Zhang, D. and Lin, X. (2008). Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics. Random Effect and Latent Variable Model Selection, D.B. Dunson (ed.)., Springer Lecture Notes in Statistics, 192.

Received November 2012
Revised February 2013
Accepted March 2013

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 299–321

Calibrated Hot-Deck Donor Imputation Subject to Edit Restrictions

Wieger Coutinho[1], *Ton de Waal*[2], *and Natalie Shlomo*[3]

A major challenge faced by basically all institutes that collect statistical data on persons, households or enterprises is that data may be missing in the observed data sets. The most common solution for handling missing data is imputation. Imputation is complicated owing to the existence of constraints in the form of edit restrictions that have to be satisfied by the data. Examples of such edit restrictions are that someone who is less than 16 years old cannot be married in the Netherlands, and that someone whose marital status is unmarried cannot be the spouse of the head of household. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. A further complication when imputing categorical data is that the frequencies of certain categories are sometimes known from other sources or have previously been estimated. In this article we develop imputation methods for imputing missing values in categorical data that take both the edit restrictions and known frequencies into account.

Key words: Categorical data; edit rules; imputation; population frequencies.

1. Introduction

National statistical institutes (NSIs) publish figures on many aspects of society. To this end, NSIs collect and process data on persons, households, enterprises, public bodies, and so on. A major challenge faced by NSIs is that data may be missing from the collected data sets. Some units that are selected for data collection cannot be contacted or may refuse to respond altogether. This is called unit nonresponse. For many individual units, data on some of the items may be missing. Persons may, for instance, refuse to provide information on their income or on their sexual habits, while at the same time giving answers to other, less sensitive questions on the questionnaire. Enterprises may not provide answers to certain questions, because they may consider it too complicated or too time consuming to answer these specific questions. Missing items of otherwise responding units is called item nonresponse. Whenever we refer to missing data in this article we will mean item nonresponse, rather than unit nonresponse.

In the statistical literature, ample attention is paid to missing data. The most common solution for handling missing data in data sets is imputation, where missing values are

[1] Loket Aangepast-Lezen, PO Box 84010, 2508 AA The Hague, The Netherlands
[2] Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands Emails: t.dewaal@cbs.nl and twal@cbs.nl
[3] School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester, M13 9PL, United Kingdom Email: natalie.shlomo@manchester.ac.uk

estimated and filled in. An important problem of imputation is to preserve the statistical distribution of the data set. This is a complicated problem, especially for high-dimensional data. For more on this aspect of imputation and on imputation in general we refer to several articles and books on imputation, such as Kalton and Kasprzyk (1986), Rubin (1987), Schafer (1997), Little and Rubin (2002), Longford (2005), and De Waal et al. (2011). Imputation methods can be divided into two broad classes: methods for categorical data and methods for numerical data. In the present article we focus on imputation of missing categorical data.

At NSIs the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions, or edits for short, that have to be satisfied by the data. Examples of such edits are that someone who is less than 16 years old cannot be married in the Netherlands, and that someone whose marital status is unmarried cannot be the spouse of the head of household. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. The problem of missing categorical data having to satisfy edits is examined by Winkler (2003) and De Waal et al. (2011).

A further complication for categorical data is that the frequencies of certain categories are sometimes known from other sources or have previously been estimated. Such frequencies will also be referred to as totals in this article. A population frequency of a category may, for instance, be known from an available related register. Alternatively, previously estimated frequencies may be known, and assumed fixed. In the Dutch Social Statistical Database estimated frequencies are fixed and later used to calibrate estimates of other quantities (see Houbiers 2004, and Knottnerus and Van Duin 2006). In fact, this strategy of fixing frequencies and later using these fixed frequencies to calibrate other quantities to be estimated forms the basis of the so-called repeated weighting method: a weighting method designed to obtain unified estimates when combining data from different sources.

In the present article we develop imputation methods for categorical data that take edits and known frequencies into account. The problem of imputation of missing categorical data having to satisfy edits and to preserve totals is also discussed in Favre et al. (2005). In contrast to the methods proposed here, the imputation is not used as an estimation technique, rather as a way to obtain consistency with edits and previously estimated totals. Another difference is that in Favre et al. (2005) only one variable to be imputed is considered. Their method does not guarantee that edits involving several variables to be imputed will be satisfied. The related problem of imputation of missing numerical data having to satisfy edits and to preserve totals is examined in Pannekoek et al. (2008). Liu and Rancourt (1999) discuss imputation of missing categorical data having to preserve totals. They do not consider edits, however.

The imputation methods developed in this article are intended to be used in the situation where one wants to impute all units of the (sub-)population under consideration. By imputing, we pursue three goals:

- To preserve the statistical distribution of the true, but unknown, data as well as possible.
- To facilitate further processing, for example producing statistical tables after imputation is simply a matter of counting, without having to worry about inconsistencies between various tables or logical inconsistencies.

- To integrate data from different sources; for example, microdata from one source are calibrated to totals from another source. In this sense the imputation methods we propose in this article can be seen as data integration techniques (also see ESSnet on Data Integration 2011).

A word of caution is in place here: an imputed data set should not be seen as a restored complete data set. In particular, in an imputed data set variances may be underestimated and correlations may be disturbed, despite attempts to preserve them as well as possible. Estimating variances and correlations taking into account the imputations is quite complex and will not be considered in this article. For an overview of methods for estimating variances with data that have undergone imputations, we refer to Chapter 10 by Haziza in Pfefferman and Rao (2009).

Rubin (1976) introduced a classification of missing data mechanisms. He distinguishes between Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). Roughly speaking, in the case of MCAR there is no relation between the missing data pattern, that is, which data are missing, and the values of the data, either observed or missing. In the case of MAR there is a relation between the missing data pattern and the values of the observed data, but not between the missing data pattern and the values of the missing data. Using the values of the observed data one can then correct for the relation between the missing data pattern and the values of the observed data, since within classes of the observed data the missing data mechanism is MCAR again. In the case of NMAR there is a relation between the missing data pattern and the values of the missing data. Such a relation cannot be corrected for without positing a model. Given that the missing data mechanism is either MCAR or MAR, we can test whether the data are MCAR or MAR. However, there are no statistical tests to differentiate between MCAR/MAR and NMAR. In practice, the only way to distinguish MCAR/MAR from NMAR is by logical reasoning. For more on missing data mechanisms we refer to Little and Rubin (2002), McKnight et al. (2007) and Schafer (1997).

In this article we assume that the missing data mechanism is MCAR. Our imputation methods can, however, easily be extended to the case of MAR, by constructing imputation classes within which the missing data mechanism is MCAR.

The remainder of this article is organized as follows. Section 2 introduces the edit restrictions we consider in this article. Section 3 describes the imputation algorithms we have developed for our problem. An evaluation study using real data is described in Section 4. Finally, Section 5 ends the article with a brief discussion.

2. Edits and Frequencies for Categorical Data

2.1. Edits for Categorical Data

We denote the number of variables by n. Furthermore, we denote the domain, that is the set of all allowed values of a variable i, by Dom_i. All domains are assumed to be non-empty. In the case of categorical data, an edit j is usually written in so-called *normal form*, that is as a Cartesian product of non-empty sets $F_i^j\,(i = 1, 2, \ldots, n)$:

$$F_1^j \times F_2^j \times \ldots \times F_n^j,$$

meaning that if for a record with values $(v_1, v_2, \ldots, v_n)$ we have $v_i \in F_i^j$ for all $i = 1, 2, \ldots, n$, then the record fails edit j, otherwise the record satisfies edit j. One generally demands that at least one of the $F_i^j\,(i = 1, 2, \ldots, n)$ should be a proper subset of the domain Dom_i, that is, should be strictly contained in Dom_i, as the "edit" with all $F_i^j\,(i = 1, 2, \ldots, n)$ equal to Dom_i cannot be failed by any record.

Example: Suppose we have three variables: *Marital Status*, *Age* and *Relation to Head of Household*. The possible values *of Marital Status* are "Married", "Unmarried", "Divorced" and "Widowed", of *Age* "< 16 years" and "≥ 16 years", and of *Relation to Head of Household* "Spouse", "Child", and "Other". Suppose we have two edits, the first edit saying that someone who is less than 16 years cannot be married, and the second one that someone who is not married cannot be the spouse of the head of household. In normal form the first edit can be written as

$$(\{\text{Married}\},\ \{<16\,\text{years}\},\ \{\text{Spouse, Child, Other}\}), \tag{1}$$

and the second one as

$$(\{\text{Unmarried, Divorced, Widowed}\},\ \{<16\,\text{years}, \geq\ 16\,\text{years}\},\ \{\text{Spouse}\}). \tag{2}$$

2.2. *Frequencies for Categorical Data*

When a frequency for categorical data is known, for instance because it has already been estimated in another source, this simply means that one knows how many units in the data set should have a specific value for a certain variable. For instance, one may know how many people in the data set have a certain age and how many people in the data set are married, even though some values of the variable *Age* and the variable *Marital Status* are missing in an observed, but incomplete data set. In this article we assume that for several categories such frequencies are known, and our aim is to obtain a fully imputed data set that preserves these frequencies.

Note that if the known frequencies are available from administrative data, then our imputation methods will duplicate the distribution of administrative marginal totals in the completed data. Our imputation methods do not necessarily preserve the distributions in the reported data. In our evaluation study in Section 4 we will examine how well the distributions in the reported data are preserved.

In practice, it may happen that a variable is fully observed in the data set while at the same time a different total is known from another source. In that case either (at least) one of the sources contains errors, or the differences are caused by different concepts, different definitions, different moments of observation and so on. We recommend using statistical data editing and data integration techniques to correct these errors and other differences before proceeding with the imputation process (see De Waal et al. 2011, and ESSnet on Data Integration 2011, for an overview of statistical data editing and data integration techniques, respectively).

3. The Imputation Methods

3.1. The Basic Idea

The imputation methods we apply in this article are all based on a hot-deck donor approach. When hot-deck donor imputation is used, for each record containing missing values, the so-called recipient record, one uses the values of one or more other records, the so-called donor record(s), to impute these missing values.

Usually, hot-deck donor imputation is applied multivariately, that is several missing values in a record are imputed simultaneously, using the same donor record. For our problem this approach is less suited. If an imputed record fails the edits, all one can do is reject the donor record and use another donor record. For a relatively complicated set of edits, one may have to test many different potential donor records until a donor record is found that leads to an imputed record satisfying all edits. Moreover, for a relatively complicated set of edits one may not even be able to find a donor record for a certain recipient record such that the resulting imputed record satisfies all edits.

Even if we were able to find single donor records for all records requiring imputation, this would then solve only part of our problem, as the totals would only be preserved in very rare cases.

We therefore apply sequential univariate hot-deck donor imputation, where for each missing value in a record requiring imputation a different donor record may be selected. The variables with missing values are imputed sequentially. For each variable, the records for which the value of this variable is missing are imputed one by one. Once all records for this variable have been imputed, the next variable with missing values is considered. The univariate hot-deck imputation methods we apply are described in Subsection 3.2. These univariate hot-deck imputation methods are used to construct a list of possible donor values for a certain missing field. Whether a value is actually used to impute the missing field depends on whether the edits can be satisfied and the totals can be preserved.

While imputing a missing value, care is taken to ensure that the record can satisfy all edits. Only values of donor records that can result in a consistent record, that is a record that satisfies all edits, are eligible to be used. In Subsection 3.3 we explain how we determine whether a value is eligible to be used for imputation. For each record we make a list of values eligible for imputation for the variable under consideration.

An eligible value may only be used for actual imputation if the total can be preserved. Before an eligible value is actually used to impute a value, we first check whether the corresponding total can be preserved. If so, we use the value for imputation. If the total cannot be preserved, the value is rejected and the next value on the list of eligible values is selected. This process goes on until we find an eligible value such that the corresponding total can be preserved.

3.2. Univariate Hot-Deck Imputation Methods

In this article we apply two univariate hot-deck donor imputation methods: a nearest-neighbour approach and a random hot-deck approach.

3.2.1. Nearest-Neighbour Hot-Deck Imputation

Suppose we want to impute a certain variable v in a record r_0 using a pool of donors where the variable v is not missing. In the nearest-neighbour approach we calculate for each other record r in the pool of donors for which the value of v is not missing a distance given by

$$Dist\,(r_0, r) = \sum_{i \neq v} w_i\left(x_i^0, x_i^r\right), \tag{3}$$

where the sum is taken over all variables except variable v, x_i^0 denotes the value of the i-th variable in record r_0, x_i^r the value of the i-th variable in record r, and $0 \leq w_i\left(x_i^0, x_i^r\right) \leq 1$ a user-specified weight expressing how serious one considers a difference between x_i^0 and x_i^r to be. The weight $w_i\left(x_i^0, x_i^r\right)$ equals zero if $x_i^0 = x_i^r$. The weight $w_i\left(x_i^0, x_i^r\right)$ is large if one considers the difference between x_i^0 and x_i^r to be important, and small if one considers the difference to be unimportant. The value of the i-th variable in record r_0, x_i^0, or the value of the i-th variable in record r, x_i^r, may be missing. If x_i^0 or x_i^r is missing, we set $w_i\left(x_i^0, x_i^r\right)$ to 1.

To impute a missing value, we first select the potential donor value from the record with the smallest distance. If that value is allowed according to the edits (see Section 3.3), we put this value on an ordered list of potential donor values: the list of eligible values. If that value is not allowed according to the edits, we try the category corresponding to the record with the second smallest distance, and so on until we find a donor value that is allowed according to the edits. After all potential donor records have been checked for eligible values, we try all values not observed in the donor records (if any). Generally all possible values are observed in the donor records. However, in principle, some values may not be observed in the donor records and may be needed to satisfy the edits and preserve totals.

Note that, once a potential donor record has been checked, all subsequent records with the same value for v will give the same result for the check, and hence do not have to be checked.

As a remark, if we used the subset of variables that are observed for all records in (3) instead of the set of all variables, the potential donor records for a certain recipient record would be ordered in the same way for each variable with missing values. In that case, if possible, multivariate imputation, using several values from the first potential donor record on this list, would be used. Only if a value of the first potential donor record could not be used because this would lead to failed edits or nonpreserved totals, a value from another potential donor record would be used.

3.2.2. Random Hot-Deck Imputation

When random hot-deck imputation is applied, a random donor record is selected, often within certain subgroups defined by auxiliary data. In our case we use random hot-deck to construct a list of possible donor values for the missing field. Let K denote the number of categories of the variable to be imputed, and let R be the total number of records with an observed value for this variable. For each category c_k ($k = 1, \ldots, K$) we determine the ratio p_k defined by the number of records for which the observed value for the variable to be imputed is equal to c_k divided by R. We then draw categories c_k ($k = 1, \ldots, K$) without replacement with probabilities p_k ($k = 1, \ldots, K$) in the donor population.

To impute a missing value, we first select the potential donor value that was drawn first. If that value is allowed according to the edits (see Subsection 3.3), we put this value on a list of potential donor values. If that value is not allowed, we try the potential donor value that was drawn second, and so on until we find a donor value that is allowed according to the edits. After all potential donor records have been checked for eligible values, we try all values not occurring in the donor records (if any) in a random order. Again, once a potential donor has been checked, all subsequent records with the same potential donor value will give the same result for the check and do not have to be checked anymore. As for nearest-neighbour imputation, we thus construct a list of potential donor values. For random hot-deck imputation, the exact order of the potential donor values is less important than for nearest-neighbour imputation. The important point here is that a list with all potential donor values is constructed.

3.3. Satisfying Edit Restrictions

In order to ensure that the set of edits can be satisfied, we derive so-called implied edits. These implied edits are necessary to guarantee that whenever we impute the current variable, the remaining variables can indeed be imputed in a manner consistent with the edits.

To determine the set of edits for the remaining variables to be imputed while imputing the current variable, we use the method proposed by Fellegi and Holt (1976) to eliminate a variable.

To eliminate a variable v_t, we start by determining all index sets S such that

$$\bigcup_{j \in S} F_t^j = Dom_t \tag{4}$$

and

$$\bigcap_{j \in S} F_i^j \neq \emptyset \quad \text{for } i \neq t. \tag{5}$$

From these index sets we select the *minimal* ones, that is the index sets S that obey (4) and (5), but none of whose proper subsets obey (4). Given such a minimal index set S we construct the implied edit

$$\bigcap_{j \in S} F_1^j \times \ldots \times \bigcap_{j \in S} F_{t-1}^j \times Dom_t \times \bigcap_{j \in S} F_{t+1}^j \times \ldots \times \bigcap_{j \in S} F_n^j.$$

By adding the implied edits resulting from all minimal sets S to the current set of edits and then removing all edits involving the eliminated variable, one obtains a set of edits for the remaining variables. It can be shown that if, and only if, this set of edits for the remaining variables can be satisfied, a value for the eliminated variable exists such that the original set of edits can be satisfied. We call this the lifting property, namely that the set of edits can be satisfied when a certain number of variables is "lifted" to a higher number of variables. The idea of the proof of the lifting property is that if a value does not exist for the eliminated variable such that the original set of edits can be satisfied, then one would be able to construct a violated implied edit, which would be a contradiction (see Fellegi and Holt 1976, and De Waal and Quere 2003, for details of the proof).

For records where multiple values are missing, we now order these variables in some order that we will describe in Subsection 3.5. Next, we eliminate the variables according to this order. Let us assume that, say, the values of variables v_1 to v_m are missing. We first substitute the values of the other variables into the original set of edits. This gives a set of edits E_0 that have to be satisfied by variables v_1 to v_m. We then eliminate variable v_1 from E_0 and obtain a set of edits E_1 that have to be satisfied by variables v_2 to v_m. Next, we eliminate variable v_2 from E_1 and obtain a set of edits E_2 that have to be satisfied by variables v_3 to v_m. We continue this process until we eliminate v_{m-1} from E_{m-2}, and obtain a set of edits E_{m-1} for variable v_m. For a single variable, edits simply define a set of allowed values for that variable. So, for variable v_m we now know which values are eligible for imputation. By a repeated application of the lifting property it can be shown that the original set of edits can be satisfied if and only if v_m satisfies E_{m-1}.

Once we have determined the edit sets $E_k\,(k = 0, \ldots, m - 1)$, we can impute the variables in reverse order. That is, we try to impute v_m by means of one of our hot-deck imputation methods (see Subsection 3.2) until we have selected an eligible value that can also preserve the total for this variable (see Section 3.4). We fill in this value for v_m into the edits in E_{m-2}. This gives us a set of eligible values for variable v_{m-1}. We continue this procedure until we have imputed all variables. What is important here is that whenever we want to impute a certain variable in a certain record, we know the set of eligible values for that variable in this record. We will use this property to preserve totals (see Subsection 3.4).

Implied edits are often used to automatically identify erroneous fields in a data set (see Fellegi and Holt 1976). It is well known that the number of implied edits may be very large. In order to identify erroneous fields automatically, one basically has to generate implied edits for every possible subset of the variables. In our case, however, the number of implied edits is much less since we only have to consider a limited number of possible subsets as the variables are eliminated in a fixed order. For instance, if there are five variables, we would have to consider 32 subsets (ranging from eliminating no variables to eliminating all five variables) for identifying errors automatically in the Fellegi and Holt approach. For our method, we only need to examine six subsets (ranging from eliminating no variable, eliminating variable 1, eliminating variables 1 and 2, etc., to eliminating variables 1, 2, 3, 4 and 5).

Example: To illustrate the use of implied edits, we assume that we have a data set with the three variables *Marital Status*, *Age* and *Relation to Head of Household* and their categories defined in Subsection 2.1. We also assume that these variables have to satisfy edits (1) and (2). Now suppose that both *Marital Status* and *Age* in a certain record are missing, and that the value of *Relation to Head of Household* equals "Spouse". Suppose that we first impute *Age* and subsequently *Marital Status*. In this case we cannot simply ignore the edits involving the variable to be imputed later, *Marital Status*, while imputing *Age*, since it would be possible to impute the value "<16 years" for the missing value of *Age*, leading to no value for *Marital Status* such that all edits are satisfied.

The edits (1) and (2) imply the edit

$$(\{\text{Married, Unmarried, Divorced, Widowed}\}, \{< 16\,\text{years}\}, \{\text{Spouse}\}), \tag{6}$$

which expresses that someone who is less than 16 years of age cannot be the spouse of the head of household. This follows from (4) and (5) by taking $S = \{1, 2\}$ and eliminating variable *Marital Status* as explained here: The sets F_i^j $(i = 1, 2, 3; j = 1, 2)$ are given by $F_1^1 = \{\text{Married}\}$, $F_2^1 = \{< 16\,\text{years}\}$, $F_3^1 = \{\text{Spouse, Child, Other}\}$, $F_1^2 = \{\text{Unmarried, Divorced, Widowed}\}$, $F_2^2 = \{< 16\,\text{years}, \geq 16\,\text{years}\}$ and $F_3^2 = \{\text{Spouse}\}$. In order to eliminate variable *Marital Status*, we take the union of F_1^1 and F_1^2, and the intersections of F_i^j $(i = 2, 3; j = 1, 2)$.

If we take the implied edit in (6) into account while imputing the missing value for *Age*, we find that we cannot impute the value "$<$16 years" and that only "$\geq$16 years" is allowed. When "$\geq$16 years" is imputed, *Marital Status* can indeed be imputed in a consistent manner.

Now that we have explained why implied edits are needed, we illustrate how we use them in our approach. Suppose we order the variables as follows: *Marital Status* and then *Age*. We substitute the value *of Relation to Head of Household* ("Spouse") into the edits (1) and (2), and obtain the edits

$$(\{\text{Married}\}, \{< 16\,\text{years}\}) \tag{7}$$

and

$$(\{\text{Unmarried, Divorced, Widowed}\}, \{< 16\,\text{years}, \geq 16\,\text{years}\}) \tag{8}$$

for *Marital Status* and *Age*. In this very simple case we now only have to eliminate one variable, *Marital Status*, and obtain the edit

$$(\{< 16\,\text{years}\}) \tag{9}$$

that has to be satisfied by *Age*. Edit (9) defines the set of eligible values for *Age:* in this case only the value "$\geq$16 years" is allowed. If we impute "$\geq$16 years" for the missing value of *Age*, we can be sure that a value for *Marital Status* exists such that all edits are satisfied. Imputing the value "$\geq$16 years" for *Age* and substituting this value into edits (7) and (8) gives the edit

$$(\{\text{Unmarried, Divorced, Widowed}\})$$

for *Marital Status*. The set of allowed values for *Marital Status* hence consists of the value "Married" only.

3.4. Preserving Totals

In the previous subsection we have explained that whenever we want to impute a certain variable in a record we know the set of eligible values. For every record we now construct such a set of eligible values for the variable to be imputed. Suppose the variable to be imputed has K categories c_1 to c_{k}. We can then summarise the problem in a table as shown in Table 1 where N_{rec} is the number of records, a 0 denotes that the category is not eligible for imputation, a "*" that the category is eligible for imputation and a 1 that this value is observed (not missing) in the corresponding record. The t_k $(k = 1, \ldots, K)$ denote the known totals.

Table 1. Illustration of the sets of eligible values

	Cat. c_1	Cat. c_2	. . .	Cat. c_K
Record 1	*	0	. . .	*
Record 2	1	0	. . .	0
Record 3	0	*	. . .	*
. . .	. . .	. . .	. . .	. . .
Record N_{rec}	*	*	. . .	0
	t_1	t_2		t_k

Now, we impute the variable under consideration record by record. We select a value from the set of eligible values for the variable to be imputed for record 1. As explained in Subsection 3.2, the list of eligible values has been constructed using one of our hot-deck approaches. After a category c_x has been selected from the list of eligible values, we perform the following two checks:

1. Is the number of records that have been assigned to the selected category c_x less than the total t_x? If so, we perform the second check. If not, we reject the selected category c_x and select a new one.
2. Will it be possible to preserve the totals involving this variable if we accept the selected category c_x? If so, we accept this value, and go to the next record to be imputed. If not, we reject the selected category c_x and select a new one, which is again subjected to the same checks.

Checking whether the total can be preserved if we accept the selected category c_x is a well-known problem of combinatorial mathematics. It is called the "Harem problem" (see Anderson 1989). The "Harem problem" is a generalization of the "Marriage problem" (see e.g., Anderson 1989, and Van Lint and Wilson 2001). In the "Harem problem", several men (the categories in our case) can select a specified number (the t_k in our case) of wives (the records in our case) they are willing to marry (assign a record to a category in our case) and add to their "harem". For each category we make a list of records that can be assigned to this category (using the $*$'s and the 0's in Table 1). The 1's in Table 1 correspond to records in which categories have been observed, and hence have already been assigned to these categories.

A condition and a constructive algorithm for solving the "Harem problem" are given in Anderson (1989). The condition given by Anderson (1989) is: $t_k\,(k = 1, \ldots, K)$ records can be assigned to categories $c_k\,(k = 1, \ldots, K)$ if, and only if, for every subset $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, K\}$ the lists of categories $c_{i_1}, \ldots, c_{i_m}$ contain in their union at least $t_{i_1} + \cdots + t_{i_m}$ records. This condition is hard to check directly. Fortunately, the constructive algorithm for solving the "Harem problem" described by Anderson (1989) provides a relatively simple way to check the condition and construct a solution at the same time. The underlying idea of this algorithm is to assign records to categories in a simple manner until one gets "stuck". Once that happens, a reshuffling algorithm (see the Appendix for a brief description of this algorithm, or Anderson 1989, for more details) is applied with the aim to assign one more record to the categories. This algorithm is repeatedly applied until either all records are assigned to categories, or until one again gets “stuck”. In the first case we have constructed a solution to this instance of the “Harem problem”, and we have shown that it is possible to preserve the totals if we accept the selected category c_x. In the second case we have demonstrated that a solution to this instance of the “Harem problem” is not possible.

Note that if, for a certain variable to be imputed, the first record with a missing value has a solution to the “Harem problem”, by construction all subsequent records to be imputed for that variable also have solutions to the “Harem problem”.

Example: We illustrate the “Harem problem” and our approach to the imputation problem by means of a simple example. Suppose that for a certain variable to be imputed, we have summarised the problem in Table 2.

Now if we select category c_3 for the first record, the “Harem problem” for the remaining records turns out to be infeasible. This is easy to see: The remaining total of four records must be assigned to categories c_1 and c_2 in some way. However, record 3 cannot be assigned to either of these categories since a 0 denotes an ineligible category. This means that category c_3 must be rejected for record 1, and we have to impute category c_2 for this record. The “Harem problem” for the remaining records is then feasible. In fact, there is only one solution: Assign record 1 to category c_2, record 3 to category c_3, and records 2, 4 and 5 to category c_1.

3.5. Order of Imputing Variables and Records

In our evaluation study in Section 4 we have imputed the variables in increasing order of missing values. That is, we impute the variable(s) with the least number of missing values first, and end with the variable(s) with the most missing values. Possibly better orders for the variables to be imputed can be developed (see, e.g., Di Zio et al. 2004).

Obviously, for a given variable, for the first record it is generally easier to find solutions to the “Harem problem” than for later records. That is, for later records, one generally needs to try more potential donor values on average before one finds a value that satisfies edits and preserves the total (although one can be sure that such a value exists if the “Harem problem” has a solution for the first record). Since it may be difficult to find suitable imputation values for different variables of the same record, we randomize the records each time before we start imputing a new variable.

As noted in the previous subsection, for each new variable, it is only for the first record to be imputed that it may be impossible to find an imputation value that satisfies all edits and preserves the total. If we cannot find a suitable imputation value for that record, we

Table 2. An example of the “Harem problem”

	Cat. c_1	Cat. c_2	Cat. c_3
Record 1	0	*	*
Record 2	*	*	*
Record 3	0	0	*
Record 4	*	*	*
Record 5	*	0	*
	3	1	1

would have to backtrack. That is, we would have to return to a previously imputed variable, and impute one or more missing values for that variable in another way. This would lead to an extremely complicated and time consuming process.

By imputing the variable(s) with the least number of missing values first and the variable(s) with the most missing values last, we try to avoid having to backtrack. The later in the imputation process, the more difficult it is to satisfy all edits and preserve all totals. Therefore, by imputing the variables with the most missing values last, we try to make finding solutions for those variables a bit easier as the more values are missing, the more “freedom” one has to satisfy edits and to preserve the totals.

In addition, in order to avoid having to backtrack, we can also try to fill in values that deactivate edits at the start of the imputation process for variables to be imputed later, even if this leads to a slightly higher distance in (3) for the nearest-neighbour approach. For instance, edit (1) could be deactivated for *Relation to Head of Household* by filling in the value “Unmarried” for *Marital Status*. Instead of backtracking or deactivating edits one could also relax the problem by removing edits or by tolerating edits or totals to not be strictly satisfied. In our evaluation study described in Section 4, we did not have to backtrack or relax the problem. We did deactivate edits while imputing the first variable. For later variables we applied the usual approach described in Sections 3.1 to 3.4.

4. Evaluation Study

In this section we describe a study on a real data set to evaluate our imputation approaches. However, as the results may be influenced by the nonresponse mechanism, we ensure MCAR by artificially creating missingness.

4.1. Evaluation Data

The evaluation data set consists of observed data from the 2001 UK Census. The data set included 1,000 randomly selected households from one area. In the data set we have one record per person in the selected households. In total the data set contained 2,447 records. Each record contained six variables (the numbers of categories are in parenthesis): *Age* (4), *Ethnicity* (12), *Employment Status* (4), *Sex* (2), *Marital Status* (6) and *Relation to Head of Household* (10). In our evaluation study we assume that totals are known for all six variables.

For this data set three explicit categorical edits were defined:

- Someone whose age is less than 16 years cannot be employed.
- Someone whose age is less than 16 years cannot be married.
- Someone whose relation to the head of household is husband or wife has to be married.

The original data set for the 2001 UK Census did not contain any missing values. In this data set we randomly introduced fixed percentages of missing values using an MCAR mechanism where for each variable we created exactly the same percentages of missing values. We created ten replications of six data sets, each data set having a fixed percentage of missing values per variable: 1%, 2%, 5%, 10%, 20% and 90%. These data sets were imputed, using the imputation methods described in Section 3. The resulting imputed data sets were subsequently compared to the original data. The evaluation measures used for this comparison are discussed in Subsection 4.3 and are calculated by averaging the evaluation measures calculated for each replication of the six data sets according to the percentage of missing values. As the nearest-neighbour imputation is deterministic in our implementation, all ten replications gave the same imputations. The evaluation measures for the random hot-deck method were relatively stable across the ten replicates.

Note that although we carried out ten replications of the imputation methods on each of the six data sets, our methods are in essence single imputation methods, rather than multiple imputation methods (see Rubin 1987). In practice, single imputation methods are preferred at NSIs rather than multiple imputation methods. In principle, our imputation methods can be adapted to multiple imputation to account for the extra variation arising from imputation.

4.2. *The Imputation Methods*

We evaluated two different imputation methods: one based on random hot-deck donor imputation and one based on nearest-neighbour hot-deck imputation. For the imputation method based on nearest-neighbour hot-deck imputation we have examined two versions. For both versions based on nearest-neighbour imputation, $w_i\left(x_i^0, x_i^r\right) = 0$ if $x_i^0 = x_i^r$ and $w_i\left(x_i^0, x_i^r\right) = 1$ if $x_i^0 \neq x_i^r$ for all variables except *Age* in the distance function (3). The two versions based on nearest-neighbour hot-deck imputation differ with respect to the weights used in the distance function (3) for variable *Age*.

In the distance function the values of *Age* are subdivided into four age groups. In one version of the method based on nearest-neighbour hot-deck imputation, $w_i\left(x_i^0, x_i^r\right) = 0$ if x_i^0 is in the same age group as x_i^r and $w_i\left(x_i^0, x_i^r\right) = 1$ if x_i^0 is in a different age group than x_i^r. This imputation method is referred to as the “equal nearest neighbour method”. In the other version of the method based on nearest-neighbour hot-deck imputation, if, $w_i\left(x_i^0, x_i^r\right) = 0$ if x_i^0 is in the same age group as x_i^r, $w_i\left(x_i^0, x_i^r\right) = 0.25$ if x_i^0 and x_i^r differ by only one age group, $w_i\left(x_i^0, x_i^r\right) = 0.5$ if x_i^0 and x_i^r differ by two age groups, and $w_i\left(x_i^0, x_i^r\right) = 0.75$ if x_i^0 and x_i^r differ by three age groups. This imputation method is referred to as the “unequal nearest neighbour method”.

4.3. *Evaluation Results*

The imputation methods are compared using the quality measures described as follows. Note that the measures are used as indicators where the smaller the value, the more the method is preferred.

Let *T* represent a frequency distribution for a two-way table produced from the data and let *T*(*r*,*c*) be the frequency in the cell in row *r* and column *c*.[1] (In this section *r* and *c* refer to “row”, respectively “column”, instead of to “record” and “category” as in earlier sections.)

Distance metric: We use the Hellinger’s Distance defined as:

$$HD\left(T_{orig}, T_{imp}\right) = \left\{0.5 \sum_{r,c} \left(\sqrt{T_{orig}(r,c)} - \sqrt{T_{imp}(r,c)}\right)^2\right\}^{1/2}$$

with *orig* and *imp* referring to the original and imputed tables respectively. The *HD* provides a measure of similarity between two probability distributions typically used for positive or zero counts.

Impact on measure of association: The first measure of association is defined as the per cent difference in the Cramer's V statistic as:

$$RCV\,(T_{orig}, T_{imp}) = \frac{100 \times \{CV\,(T_{imp}) - CV\,(T_{orig})\}}{CV\,(T_{orig})}$$

where

$$CV\,(T) = \sqrt{\frac{\chi^2}{\min\,(N_R - 1, N_C - 1)}}$$

is the Cramer's V measure of association defined in terms of χ^2, the usual Pearson chi-squared statistic for testing independence in a two-way table, N_R is the number of rows and N_C is the number of columns. The *RCV* provides a measure of attenuation of the association in the table.

The second measure of association is defined as the per cent difference in the variance of the cell counts:

$$RV\,(T_{orig}, T_{imp}) = \frac{100 \times \{V\,(T_{imp}) - V\,(T_{orig})\}}{V\,(T_{orig})}$$

where

$$V\,(T) = \frac{\sum_{r,c}(T(r,c) - \bar{T})^2}{N_R N_C - 1}.$$

The *RV* provides a measure of attenuation of the counts in the table indicating whether the cell counts are "flattening" as a result of the imputation.

Impact on an ANOVA analysis: Another form of bivariate analysis consists of comparing proportions in a category of a column (outcome) variable between categories of a row (explanatory) variable. Let

$$P^c(r) = \frac{T(r,c)}{\sum_c T(r,c)}$$

be the proportion in column c for row r and define the between-row variance of this proportion by:

$$BV(P^c) = \frac{\sum_r (P^c(r) - P^c)^2}{N_R - 1} \text{ where } P^c = \frac{\sum_r T(r,c)}{\sum_{r,c} T(r,c)}.$$

The measure is defined as:

$$BVR\left(P^c_{orig}, P^c_{imp}\right) = \frac{100 \times \left\{BV\left(P^c_{imp}\right) - BV\left(P^c_{orig}\right)\right\}}{BV\left(P^c_{orig}\right)}$$

The BVR provides a measure of attenuation of between group differences in an ANOVA analysis and indicates the undesirable result that the group proportions are “flattening” towards the overall proportion.

Figures 1 through 4 present graphs of the average quality measures across the ten replicates for some main distributions in the data set. The unequal nearest neighbour method provided similar results to the equal nearest neighbour method and hence we compare the random hot-deck method (denoted by “random”) with the equal nearest neighbour method (denoted by “equal_nn”) in the figures.

Figure 1a presents the Hellinger’s Distance (*HD*) on a table of counts spanned by *Age Group* and *Employment Status* (16 cells). For all imputation rates, the equal nearest

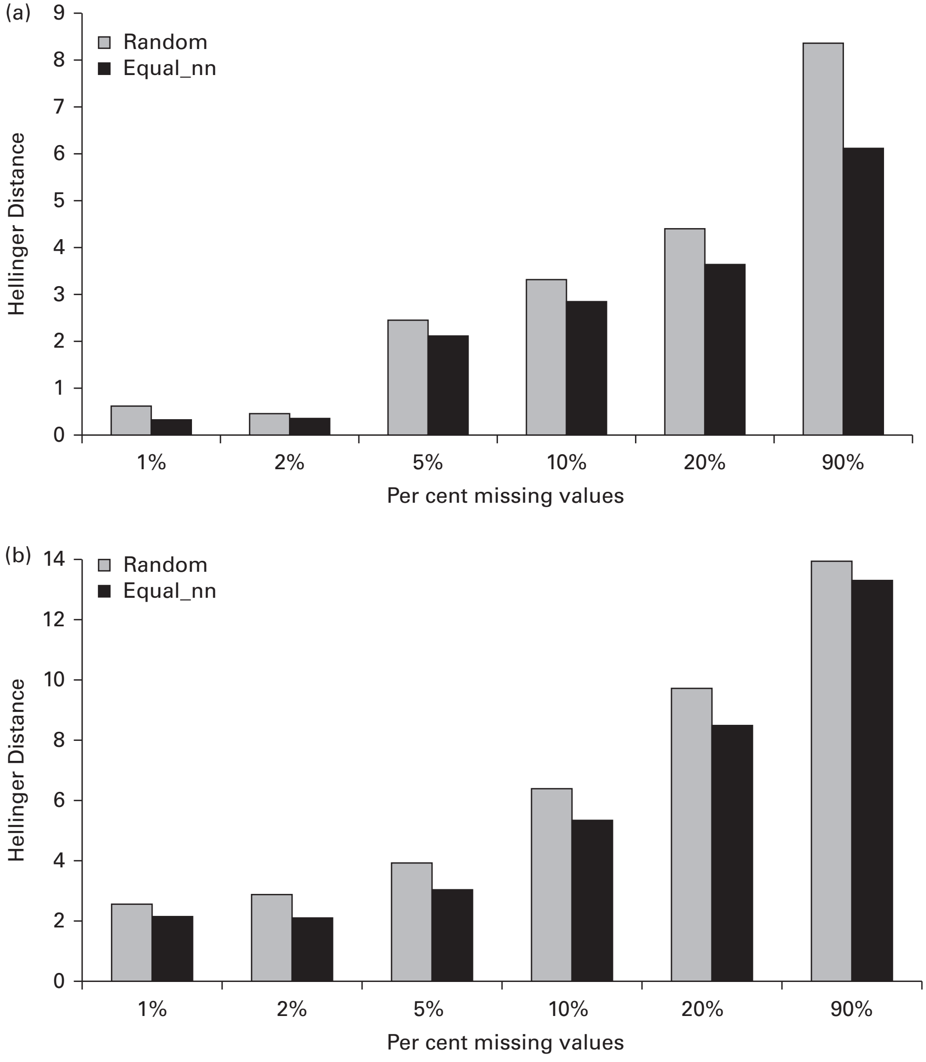


Fig. 1. (a) Average Hellinger’s Distance (HD) across replicates on the table Age Group and Employment Status. (b) Average Hellinger’s Distance (HD) across replicates on the table Age Group and Relation to Head of Household

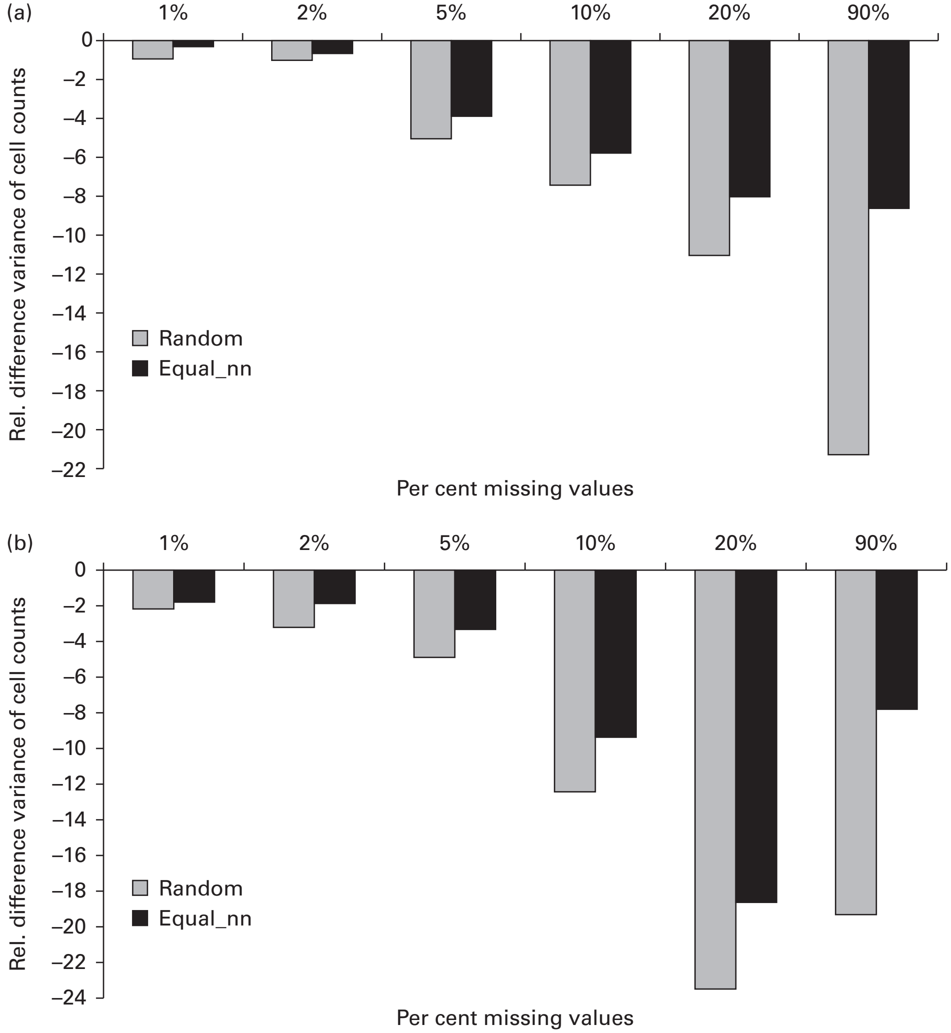


Fig. 2. (a) Average per cent relative difference in variance of cell counts (RV) across replicates on the table Age Group and Employment Status. (b) Average per cent relative difference in variance of cell counts (RV) across replicates on the table Age Group and Relation to Head of Household

neighbour method has lower Hellinger's Distance compared to the random method. Figure 1b presents the Hellinger's Distance for the table spanned by *Age Group* and *Relation to Head of Household* (40 cells) showing similar results.

Figure 2a presents the per cent relative difference in the variance of the cell counts for the table spanned by *Age Group* and *Employment Status*. The negative values of the *RV* measure means that the variance of counts with imputed values is less than the original variance of counts. The cell counts are "flattened" as a result of the imputation, leading to a smaller variance of the counts. The equal nearest neighbour method (as well as the unequal nearest neighbour method) has less change in the variance of the cell counts compared to the random method. Figure 2b presents the *RV* measure for the table spanned by *Age Group* and the *Relation to the Head of Household* with similar results.

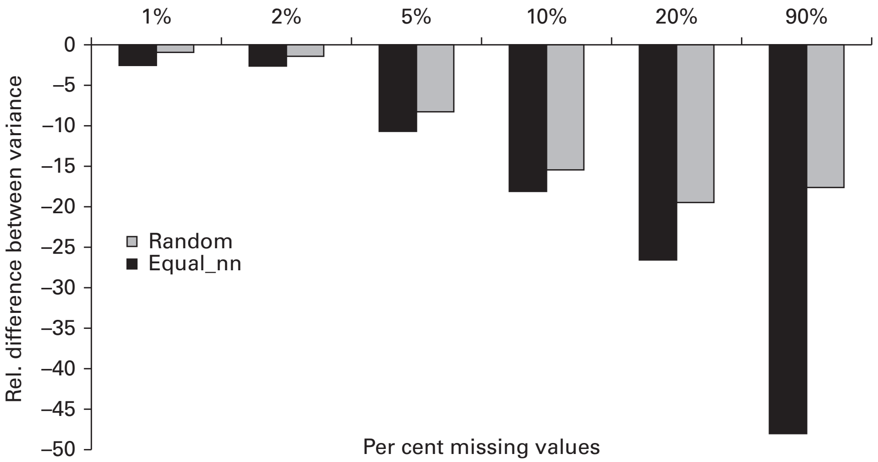


Fig. 3. Average per cent relative difference in between variance (BVR) across replicates of proportion of Employed across Sex and Age Groups

Figure 3 presents the per cent relative difference in the between variance of the proportion of employed persons in groups defined by *Sex* and *Age* groups (*BVR*). The negative values of the *BVR* measure means that the between variance of the group proportions of employed persons with imputed values is less than the original between variance. The group proportions are attenuating to the overall proportion as a result of the imputation. Again, equal nearest neighbour method (and the unequal nearest neighbour method) has less change in the *BVR* compared to the random method.

Figure 4a presents the per cent relative difference in the Cramer's V statistic of the table spanned by *Age Groups* and *Employment Status* (*RCV*). The negative values of the *RCV* measure means that the Cramer's V statistic on the table with imputed values is less than the original Cramer's V statistic. The table of counts is attenuating towards assumptions of independence compared to the original table. For all imputation rates, the equal nearest neighbour method has less change in the Cramer's V statistic than the random method and similarly for the unequal nearest neighbour method. Figure 4b presents the *RCV* measure for the table spanned by *Age Groups* and *Relation to Head of Household* with similar results.

In Figure 5, we present box plots of the proportion of values that were *not* imputed back to their original value in the data set according to the percentage missing and imputation method. Each box plot includes a total of 38 proportions which is the number of categories of the six variables in the data set. The proportions were calculated as the average across the replications. The proportion is very small for the data sets, with 1% and 2% missing values. Based on the data sets with 5% missing values and over, we can see a slight advantage to the equal nearest neighbour approach with less outlying proportions, a smaller interquartile range of the proportions and a slightly smaller median proportion.

5. Discussion

In this article we have developed two imputation methods for categorical data that take edits and known totals into account while imputing a record. One of the imputation

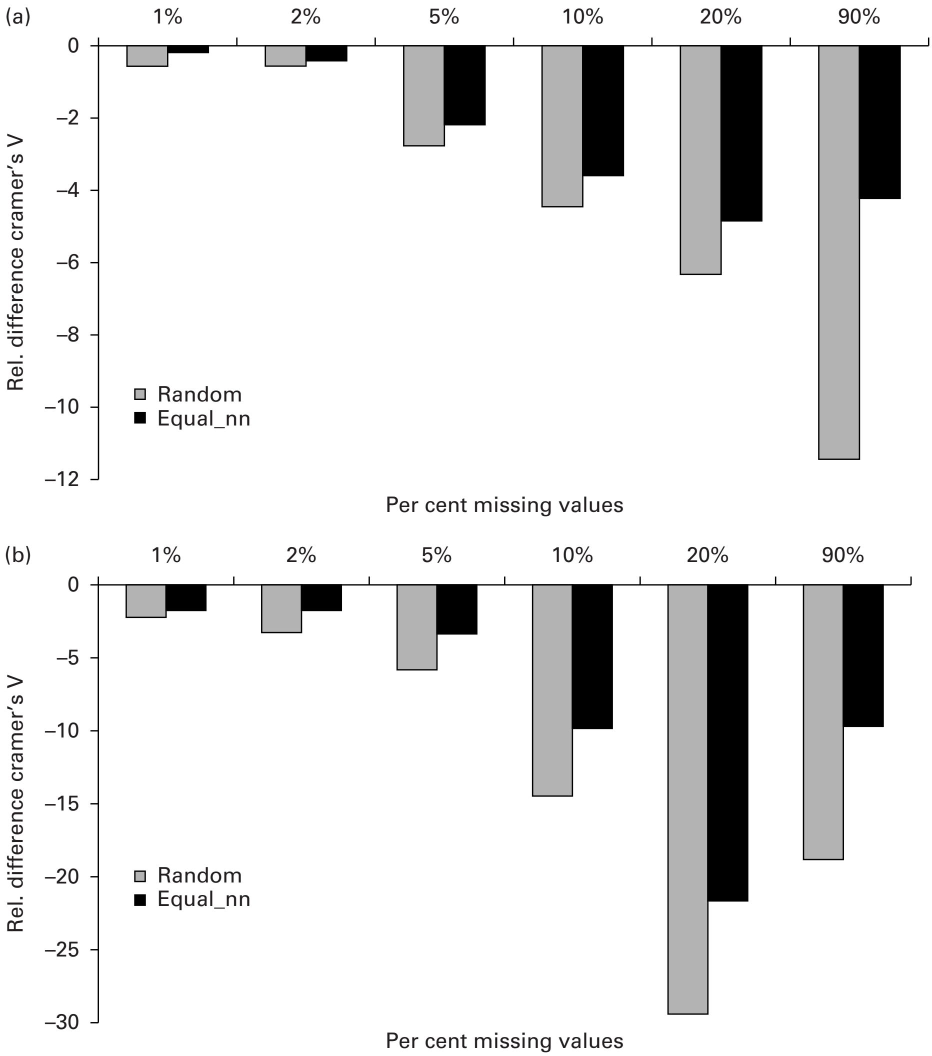


Fig. 4. (a) Average per cent relative difference in Cramer's V across replicates on the table of Age Groups and Employment Status. (b) Average percent relative difference in Cramer's V across replicates on the table of Age Groups and Relation to the Head of Household

methods proposed in this article is based on random hot-deck donor imputation and the other on nearest-neighbour donor imputation. Our evaluation study shows that the method based on nearest-neighbour imputation performs slightly better than the method based on random imputation. In our evaluation study, changing the weights in the distance function of the method based on nearest neighbour imputation had little or no effect on the outcome of the results. All imputation methods provide exactly the totals to those used in the benchmarking. For non-benchmarked subdomain totals, one can assess the potential bias as shown by the Hellinger's Distance in Figures 1a and 1b. To ensure totals for subdomains of interest, the imputation methods can be carried out separately in each subdomain assuming that the totals are known.

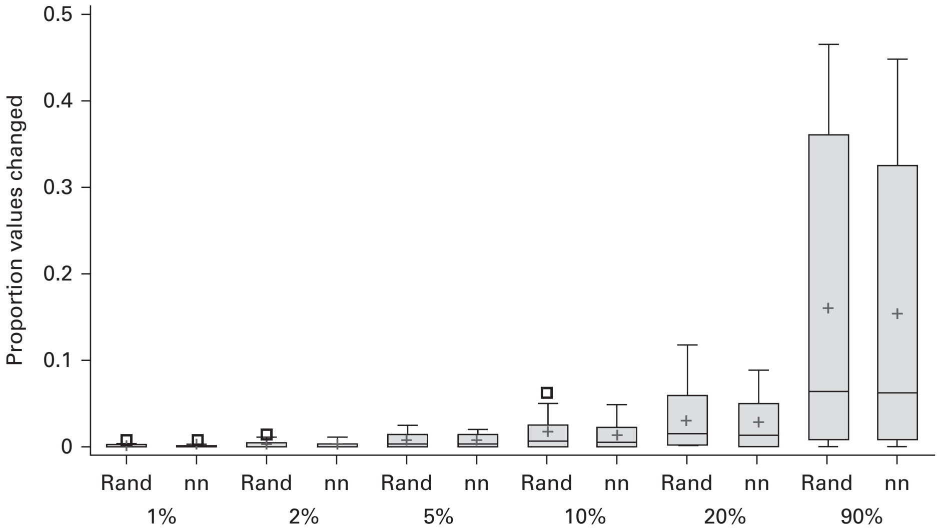


Fig. 5. Proportion of values changed in the data set for the Random (rand) and Equal Nearest Neighbour (nn) approaches according to per cent missing values (average across replicates)

The problem of imputing missing data while satisfying edits and preserving totals has hardly been studied in the literature. Our methods are among the first for this kind of problem. Many aspects of the developed methods can undoubtedly be extended and improved upon.

A possible extension is to develop similar methods for the situation where one wants to impute a sample data set, instead of all units in the population as in the current article. In order to impute a sample data set so that population totals are preserved, one would have to extend our methods to deal with sampling weights. If all sampling weights are integers, a first idea would be to simply make w copies of a record with sampling weight w, and then apply the methods described in this article. When translating this back to the sample, fractions of categories would then be "imputed" in each record. If sampling weights are not integers, the situation is more complicated, and one would have to do some rounding. It is very likely that more efficient and better approaches can be developed for extending our methods to sample data sets.

Another interesting extension is to develop similar imputation methods for the case where bivariate marginal distributions with overlapping variables, say of the pair of variables (X,Y) and the pair of variables (X,Z), are known instead of only univariate marginal distributions. In principle, this could be solved by constructing the crossings of (X,Y) and of (X,Z), and adding these crossings to the set of variables. In order to avoid any inconsistencies between the marginals of these crossings and the marginals of variables X, Y and Z, one would then need to add edits, for example: "if ($X = x, Y = y$) then ($X = x$)" and "if ($X = x$,$Y = y$) then ($Y = y$)" for the crossing of X and Y, and similar edits for the crossing of Y and Z.

Although this is, in principle, a possible approach, it is likely to be time consuming with more chances of getting "stuck" in the "Harem problem" and having to backtrack. A more efficient approach for this situation remains to be developed.

Alternatively, knowing the marginal of (*X*,*Y*) and (*X*,*Z*), one could estimate (*X*,*Y*,*Z*) using log-linear modelling and carry out the imputation separately in each subdomain of this cross-classification. Again, it is likely that better approaches can be developed.

It is unclear whether the use of known totals in the imputation process preserves correlations better between variables compared to when totals are not used in the imputation process. We hope to explore this in future research.

Our imputation methods consist of two different parts: a statistical part (drawing potential donor values) and a combinatorial part (satisfying edits and preserving totals). The final aim of research in this area should be to develop a statistical framework that organically incorporates the combinatorial part as well.

Appendix: The Reshuffling Algorithm for the "Harem Problem"

Assume that (some) records have already been assigned to categories by means of a simple algorithm, for example by a "greedy" algorithm where first as many records as possible are assigned to the first category without exceeding the total for this category, then as many records as possible out of the remaining records are assigned to the second category without exceeding the total for that category, and so on, until either all records have been assigned to categories or one gets stuck. In the first case, the "Harem problem" has been solved. In the second case, we apply the reshuffling algorithm sketched below, which aims to assign one extra record to the categories.

As in Subsection 3.4, we denote the number of records by N_{rec}. Define $L(r_i)$ as the set of categories that are eligible for imputation of record $r_i\,(i = 1, \ldots, N_{rec})$. With $r_{[j]}$ we denote the *j*-th record that is selected in the procedure sketched below. For example, if the first record selected is r_3, then $r_{[1]} = r_3$ and $L(r_{[1]}) = L(r_3)$. The same record may be selected several times, so some of the $r_{[j]}$ may refer to the same record. Likewise, we use $c_{[j]}$ to denote the *j*-th category that is selected in the procedure, for example if the first category selected is c_3 then $c_{[1]} = c_3$. Again, the same category may be selected several times, so some of the $c_{[j]}$ may refer to the same category.

1. Select a record $r_{[1]}$ that has not yet been assigned to a category.
2. Select a category $c_{[1]}$ from $L(r_{[1]})$.
   - ○ If $r_{[1]}$ may be assigned to $c_{[1]}$ without exceeding the total for this category, we are obviously done.
   - ○ If $r_{[1]}$ may not be assigned to $c_{[1]}$, we set $L(r_{[1]}) := L(r_{[1]}) - \{c_{[1]}\}$, i.e., $c_{[1]}$ is dropped from $L(r_{[1]})$. Go to Step 3.
3. Select a record $r_{[2]}$ that has been assigned to $c_{[1]}$, and set $L(r_{[2]}) := L(r_{[2]}) - \{c_{[1]}\}$.
4. Select a category $c_{[2]}$ from $L(r_{[2]})$.
   - ○ If $r_{[2]}$ may be assigned to $c_{[2]}$ without exceeding the total for this category, we are done (see below).
   - ○ If $r_{[2]}$ may not be assigned to $c_{[2]}$, we set $L(r_{[2]}) := L(r_{[2]}) - \{c_{[2]}\}$ and go to Step 5.
5. Select a record $r_{[3]}$ that has been assigned to $c_{[2]}$, and set $L(r_{[3]}) := L(r_{[3]}) - \{c_{[2]}\}$.
6. And so on.

This reshuffling algorithm will eventually terminate. It can terminate in two possible ways:

Table A.1. Preliminary assignment of records to categories

	Cat. c_1	Cat. c_2	Cat. c_3
Record 1	0	$\underline{1}$	$\underline{0}$
Record 2	$\underline{0}$	$\underline{0}$	$\underline{1}$
Record 3	0	0	$\underline{0}$
Record 4	$\underline{1}$	$\underline{0}$	$\underline{0}$
Record 5	$\underline{1}$	0	$\underline{0}$
	3	1	1

a. We can assign some $r_{[k]}$ to a category $c_{[k]}$. In this case we can assign an extra record to a category. Namely, we can assign $r_{[k]}$ to $c_{[k]}$. Previously, $r_{[k]}$ had been assigned to a category $c_{[m]}$ $(m \leq k - 1)$. To this $c_{[m]}$ we can assign a record $r_{[p]}$ $(p \leq m)$. We can continue in this way until we can assign record $r_{[1]}$ to category $c_{[1]}$.
   At this moment we have assigned an extra record to a category, and we are ready to restart the algorithm with another record that has not yet been assigned to a category. When there are no more records that need to be assigned to a category, this instance of the "Harem problem" has been solved.

b. We try to select a category from an empty set $L(r_{[j]})$. In this case we can conclude that this instance of the "Harem problem" is unsolvable.

We illustrate the above algorithm on the "Harem problem" given in Table 2. We assume that some records have already been assigned to categories by means of a simple "greedy" algorithm. The preliminary assignment of records to categories after application of this "greedy" algorithm is summarised in Table A.1, where categories that are eligible for imputation are underlined.

$L(r_1) = \{c_2, c_3\}$, $L(r_2) = \{c_1, c_2, c_3\}$, $L(r_3) = \{c_3\}$, $L(r_4) = \{c_1, c_2, c_3\}$ and $L(r_5) = \{c_1, c_3\}$. Only r_3 has not yet been assigned to a category, so we select $r_{[1]} = r_3$. We select $c_{[1]} = c_3$ from $L(r_3)$, and update $L(r_3) := \emptyset$. We select a record $r_{[2]}$ that has been assigned to c_3. In this case there is only one option, namely record r_2, so, $r_{[2]} = r_2$ and we update $L(r_2) := \{c_1, c_2\}$. We select a category, say $c_{[2]} = c_2$, from $L(r_2)$, and update $L(r_2)\!: = \{c_1\}$. We select a record $r_{[3]}$ that has been assigned to c_2. In this case there is again only one option, namely record r_1, so, $r_{[3]} = r_1$, and we update $L(r_1)\!: = \{c_3\}$. We select $c_{[3]} = c_3$ from $L(r_1)$, and update $L(r_1) := \emptyset$. We select a record $r_{[4]}$ that has been assigned

Table A.2. Assignment of records to categories after the reshuffling algorithm

	Cat. c_1	Cat. c_2	Cat. c_3
Record 1	0	$\underline{1}$	$\underline{0}$
Record 2	$\underline{1}$	$\underline{0}$	$\underline{0}$
Record 3	0	0	$\underline{1}$
Record 4	$\underline{1}$	$\underline{0}$	$\underline{0}$
Record 5	$\underline{1}$	0	$\underline{0}$
	3	1	1

to c_3. In this case there is again only one option, namely record r_2, so, $r_{[4]} = r_2$. Updating $L(r_2)$ has no effect: $L(r_2)$: $= \{c_1\}$. We select $c_{[4]} = c_1$, from $L(r_2)$.

Record $r_{[4]} = r_2$ can be assigned to $c_{[4]} = c_1$. Previously, r_2 had been assigned to category $c_{[1]} = c_3$. In turn, we can assign record $r_{[1]} = r_3$ to category $c_{[1]} = c_3$. The assignment of records to categories after the reshuffling algorithm is summarised in Table A.2.

In this case, the "Harem problem" has been solved. In general one needs to apply the reshuffling algorithm several times before the "Harem problem" is solved, or before one can conclude that this instance of the problem is unsolvable.

6. References

Anderson, I. (1989). A First Course in Combinatorial Mathematics, (second edition). Oxford: Oxford University Press.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: John Wiley & Sons.

De Waal, T. and Quere, R. (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. Journal of Official Statistics, 19, 383–402.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004). Bayesian Networks for Imputation. Journal of the Royal Statistical Society: Series A, 167, 309–322.

ESSnet on Data Integration (2011). Report on WP 2: Methodological Developments. Available at: http://www.essnet-portal.eu/sites/default/files/131/WP2.pdf.

Favre, A.-C., Matei, A., and Tillé, Y. (2005). Calibrated Random Imputation for Qualitative Data. Journal of Statistical Planning and Inference, 128, 411–425.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Houbiers, M. (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics, 20, 55–75.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Knottnerus, P. and Van Duin, C. (2006). Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. Journal of Official Statistics, 22, 565–584.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data (second edition). New York: John Wiley & Sons.

Liu, T.-P. and Rancourt, E. (1999). Categorical Constraints Guided Imputation for Nonresponse in Survey. Report, Statistics Canada.

Longford, N.T. (2005). Missing Data and Small-Area Estimation. New York: Springer.

McKnight, P.E., McKnight, K.M., Sidani, S., and Figueredo, A.J. (2007). Missing Data – A Gentle Introduction. New York: The Guilford Press.

Pannekoek, J., Shlomo, N., and De Waal, T. (2008). Calibrated Imputation of Numerical Data under Linear Edit Restriction. UN/ECE Work Session on Statistical Data Editing, Vienna.

Pfefferman, D. and Rao, C.R. (2009). Handbook of Statistics 29, Volume 29A. Amsterdam: Elsevier.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Van Lint, J.H. and Wilson, R.M. (2001). A Course in Combinatorics (second edition). Cambridge: Cambridge University Press.

Winkler, W.E. (2003). Contingency-Table Model for Imputing Data Satisfying Analytic Constraints. U.S. Bureau of the Census, Washington, D.C.

Received April 2012
Revised February 2013
Accepted February 2013

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 323–327

Book Reviews

Books for review are to be sent to the Book Review Editor Jaki S. McCarthy, USDA/NASS, Research and Development Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.
Email: jaki_mccarthy@nass.usda.gov

Gerald J. Hahn and Necip Doganaksoy. *A Career in Statistics: Beyond the Numbers*. Hoboken, NJ: John Wiley & Sons, Inc., 2011. ISBN 978-0-470-40441-6, 340 pp, $69.95.

Hahn and Doganaksoy provide a valuable service to the statistics community of academics and practitioners through writing and publishing *A Career in Statistics: Beyond the Numbers*. Their contributions for readers are many, as they provide both an overview of the field of statistics as well as specific advice for students and practitioners.

The first four chapters provide background for careers in statistics (Chapter 1) and review what statisticians do in business and industry (Chapter 2), in official government roles (Chapter 3) and in other areas of application (Chapter 4).

Chapter 5 – The Work Environment and On-the-job Challenges – is a "must-read" for any student of statistics or for any professional who needs to either learn or be reminded that most professionals that they will work with are not themselves statisticians, nor do they have training in statistics. For example, many co-workers may not understand your role as a statistician, the appropriate use of statistics for different applied problems or how statistical analysis could add value to their tasks and projects. A statistician may spend a significant amount of his or her time "marketing" his or her skills within their own organization. Subsection 5.4 addresses the issue of role delineation: Is a statistician a consultant or a member of a team? When do the roles fuse and/or change? All of this information is extremely relevant for statisticians who want to be useful and effective within their organizations and beyond.

Chapter 6 focuses on traits and behaviors of successful statisticians, with a focus on "soft" or "people" skills and an assumption that readers possess the necessary technical skills (which are however insufficient on their own) to do their jobs well. These topics may typically receive little attention in a graduate or professional skills training situation focused on statistics and numerical analysis; however, having these skills is important to the success of any professional and, the authors argue, especially professional statisticians.

Professional training and advanced degrees are the topic of Chapter 7, with the authors providing insight into graduate programs as well as the value of informal educational experiences, such as internships and participating in consulting arrangements. Chapter 14 gives attention to lifelong learning for statisticians.

Chapter 8 focuses on the job search and the recruiting process specific to statisticians while Chapters 12 and 13 discuss different career paths, including academia.

Best considered as an on-the-job primer for statisticians, Chapters 9, 10 and 11 discuss many practical topics that can be encountered by one new to the field, or one seasoned in the field. These include: project selection, estimating project costs, and successfully executing projects. Subsection 9.4 includes key advice that seasoned practitioners would share with aspiring statisticians planning on or embarking on a career. Much of the advice reminds readers to be relevant, keep it simple and find ways to be of value to their organizations.

This book is extremely well done. The sidebars and "major takeaways" offered in the text are very useful and present quick and easy summaries for the reader. I would recommend this book to any person considering an analytical support or analytical leadership position in statistics (and even related fields). Portions of this book, if not the entire text, would be appropriate required reading for professional training for graduate students in statistics.

Heather H. Boyd, Ph.D., C.P.M.
University of Notre Dame,
Research Development Program Director
Office of the Vice President for Research
940 Grace Hall
Notre Dame, Indiana 46556
(phone) 574 631 4104
Email: hboyd@nd.edu

Ton de Waal, Jeroen Pannekoek, Sander Scholtus. *Handbook of Statistical Data Editing and Imputation*. New York: Wiley, 2011, ISBN 978-0-470-54280-4, Hardcover $149.95.

The handbook compiled by Ton de Waal, Jeroen Pannekoek, and Sander Scholtus of Statistics Netherlands is an enjoyable, informative, instructive, and comprehensive compendium of known methods for the editing and imputation of major surveys. Expert technical knowledge is expressed clearly on topics of statistical science, mathematics, and linear programming, with separate discussions for automated processing and interactive processing of edits. Methods for automation of editing and imputation are a focus. Tools supporting the Fellegi-Holt paradigm are emphasized (Fellegi and Holt 1976). Tools for interactive edit and manual imputation such as Blaise (Statistics Netherlands) are discussed briefly in the context of selective editing. Discussion of donor imputation using nearest neighbor imputation methodology (NIM, Bankier 1999) includes a detailed comparison with Fellegi-Holt methodology. Discussion of methods for variance estimation compares the bootstrap and jackknife methods with multiple imputation and fractional imputation. The handbook develops handling of edits through a series of mathematical theorems with proofs and clear examples of their application. The theoretical development leads incrementally to sophisticated tools for automated edit and imputation of categorical variables as well as continuous variables. Computational methods such as *branch-and-bound* algorithms are thoroughly discussed throughout the book. Each chapter ends with a generous listing of international references. The subject index is thorough and reliable.

Automation of editing and imputation emphasizes the need for classification of edits. The authors address this early in the handbook, favoring a distinction between *hard edits* which are logically necessary for a record to be consistent (e.g., NET PAY = GROSS PAY – DEDUCTIONS) and *soft edits* which serve as guidelines to flag potential errors in the record (e.g., AGE $\leq$ 110 years). The authors define subclasses by mathematical form of the edit. Classification of edits is conceptually important for any survey group that is contemplating the use of a centralized edit repository (i.e., database) which is maintained independently of other production systems (e.g., automated edit and imputation systems). The classification of an edit might be used to determine the scope for its application. For example, the use of an edit as a prescriptive relationship amongst survey variables (e.g., adjustment procedures in Chapter 10) may be suitable for hard edits and less desirable for soft edits. The text offers best practices (e.g., strategy to mitigate overediting) and characterizes the intended use for automated procedures. The philosophy for editing and imputation adopted by the authors might not generalize to survey organizations lacking a national registry or centralized data collection. However, the technical understanding of procedures portrayed by the authors would inform any usage of the procedures.

The handbook serves as a guide suggesting options for handling of edit constraints in tandem with automated imputation of survey reports (records). The authors discuss procedures to meet three steps: 1) the edit; 2) the imputation; and 3) making the imputed values consistent with the edits. For step (2), deterministic imputation procedures and stochastic imputation procedures are described. For step (3), incorporation of edit rules into the automated construction of imputed values is discussed in Chapter 9; and adjustment of imputed values to meet edit rules is discussed in Chapter 10. Nearest neighbor imputation methodology (NIM) is discussed in the context of detection and correction of errors (e.g., Johanson 2012).

At the Washington Statistical Society's 2011 Morris Hansen Lecture hosted by the National Agricultural Statistics Service (NASS), Roderick Little spoke of the lack of congruence between the theory for modeling and the theory for sampling as a "schizophrenia." The handbook includes a useful theoretical development of the relationship between model-based imputation and sample-based weights in Subsection 7.3.4 *Connection between Imputation and Weighting*. In particular, there is the question of how to manage design weights and adjustment factors in the context of model building for imputation and estimation. The handbook partially addresses the issue with examples and theoretical proofs suggesting appropriate procedures for incorporating design weights into a model-based imputation. In particular, the authors have proven conditions (Theorem 7.1) under which estimators are equivalent across 1) weighting the respondent data by applying the *regression estimator*; 2) imputing the nonrespondents using regression imputation and then weighting the entire sample by applying the regression estimator; and 3) imputing nonrespondents and nonsampled elements using regression imputation. The authors refer to these three cases as the *weighting approach*, the *combined approach*, and the *mass imputation approach* respectively. Further discussion of the design/model compromise (DMC) is provided by Roderick Little (Little 2012).

Being somewhat new to surveys, I considered the book in terms of my current projects: implementing a new system for significance editing called SignEdit (Kosler 2012); utilizing procedures from the Banff System commercialized by Statistics Canada (Johanson 2012); implementing iterative sequential regression (ISR) with edit constraints for the *Agricultural Resource Management Survey* (Robbins et al. 2012); and construction of a centralized edit repository. In a supportive manner, the handbook provided useful and in-depth technical background on most editing and imputation topics pertinent to these projects (e.g., donor and ISR imputation procedures). Several topics were treated with a history of the development of known methods (e.g., error localization) and a comparison of approaches (e.g., adjustment of imputed values to meet edit constraints).

The handbook's authors synthesized a broad range of material for the practicing survey statistician, gathering topics as diverse as *group random hot deck imputation*, *Gibbs sampling*, and *Fourier-Motzkin elimination*. For useful methods, the reader would find a convenient combination of textbook level descriptions of methodology, examples commonly published in hard-to-find technical reports (e.g., selective editing for the Dutch Agricultural Census), and theoretical proofs commonly published in major journals (e.g., EM algorithm for a Dirichlet distribution). It was helpful to see the thought process behind researchers and developers at Statistics Netherlands, given their leadership in the theory and practice of editing and imputation of survey data. One might notice that the handbook does not directly address the application of editing and imputation methodology in the context of non-probability sampling. In any event, the handbook would be a valuable resource for implementation of new editing and imputation programs in any agency. The thorough integration of theoretical, computational, and practical information covered the bases for management of edits or business rules.

References

Bankier, M. (1999). Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses. Working Paper No. 24. Rome: UN/ECE Work Session on Statistical Data Editing.

Fellegi, I. and Holt, D.T. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Johanson, J.M. (2012). Banff Automated Edit and Imputation on a Hog Survey. Proceedings of the Fourth International Conference of Establishment Surveys, June 11–14, 2012. Montréal, Canada [CD-ROM]: American Statistical Association.

Kosler, J.S. (2012). Survey Process Control with Significance Editing: Foundations, Perspectives, and Plans for Development. Proceedings of the Fourth International Conference of Establishment Surveys, June 11–14, 2012. Montréal, Canada [CD–ROM]: American Statistical Association.

Little, R.J. (2012). Calibrated Bayes, an Inferential Paradigm for Official Statistics. Journal of Official Statistics, 28, 309–334.

Robbins, M.W., Ghosh, S.K., and Habiger, J.D. (2012). Imputation in High Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. Journal of the American Statistical Association (recently accepted for publication).

Joseph S Kosler, PhD
United States Department of Agriculture
National Agricultural Statistics Service (NASS)
Research and Development Division
3251 Old Lee Highway
Fairfax, VA 22030
Email: Joseph.Kosler@nass.usda.gov