



Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study

Brooke Helppie-McFall and Amanda Sonnega



Feasibility and Reliability of Automated Coding of Occupation and Industry in the Health and Retirement Study

Brooke Helppie-McFall
University of Michigan

Amanda Sonnega
University of Michigan

December 2018

Michigan Retirement Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48104
www.mrrc.isr.umich.edu
(734) 615-0422

Acknowledgements

The research reported herein was performed pursuant to a grant from the U.S. Social Security Administration (SSA) funded as part of the Retirement Research Consortium through the University of Michigan Retirement Research Center Award RRC08098401-10. The opinions and conclusions expressed are solely those of the author(s) and do not represent the opinions or policy of SSA or any agency of the federal government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation or favoring by the United States government or any agency thereof.

Regents of the University of Michigan

Michael J. Behm, Grand Blanc; Mark J. Bernstein, Ann Arbor; Shauna Ryder Diggs, Grosse Pointe; Denise Ilitch, Bingham Farms; Andrea Fischer Newman, Ann Arbor; Andrew C. Richner, Grosse Pointe Park; Ron Weiser, Ann Arbor; Katherine E. White, Ann Arbor; Mark S. Schlissel, *ex officio*

Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study

Abstract

Due to advances in computing power and the increase in coverage of longitudinal datasets in the Health and Retirement Study (HRS) that provide information about detailed occupations, demand has increased among researchers for improved occupation and industry data. The detailed data are currently hard to use because they were coded at different times, and the codeframes are, therefore, not consistent over time. Additionally, the HRS gathers new occupation and industry information from respondents every two years, and coding of new data at each wave is costly and time-consuming. In this project, we tested the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) to see if it could improve processes for coding data from the HRS. We tested results from NIOCCS against results from a human coder for multiple datasets. NIOCCS does reasonably well compared to coding results from a highly-trained, professional occupation and industry coder, with kappa inter-rater reliability on detailed codes of just under 70 percent and agreement rates on broader codes of around 80 percent; however, code rates for NIOCCS for the datasets tested ranged from 60 percent to 72 percent, as compared to a professional coder's ability to code those same datasets that ranged from 95 percent to 100 percent. In its current form, we find that NIOCCS is a tool that might be best used to reduce the number of cases human coders must code, either in coding historical data to a consistent codeframe or in coding data from future HRS waves. However, it is not yet ready to fully replace human coders.

Citation

Helppie-McFall, Brooke and Amanda Sonnega. 2018. "Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study." Ann Arbor MI: University of Michigan Retirement Research Center (MRRC) Working Paper, WP 2018-392.
<https://mrdrc.isr.umich.edu/publications/papers/pdf/wp392.pdf>

Authors' acknowledgements

The authors are grateful to Stanley Hasper for his hard work and sharing his expertise on occupation and industry coding; to the NIOCCS team at NIOSH/CDC for all of their help, advice and support; as well as Taylor Slayton for excellent research assistance.

Introduction

Increasing numbers of researchers are interested in using occupational history data to study a variety of topics, including, for example, the effect of computerization on retirement (Willis 2013), effects of lifetime work characteristics on cognitive functioning in retirement (Fisher et al. 2014), and effects of work experiences on later life work transitions (Angrisani et al. 2015; Angrisani et al. 2017; Helppie-McFall et al. 2014; Sonnega et al. 2015). A challenge they all face, however, is that the occupation and industry codes created for waves of many longitudinal studies, including the HRS, are intertemporally inconsistent. For example, data from the Health and Retirement Study (HRS) collected between 1992 and 2010 were coded according to separate census classifications (1980 and 2000) that have since (with the 2010 Census) been revised to reflect the changes in the United States occupational and industry structure that have occurred during the last two decades. The earlier occupation codes provided great detail about manufacturing occupations, but little detail about occupations in the service sector and, most notably, provide no detail on the many new occupations that have appeared as a consequence of computerization and automation. For researchers interested in tracing occupational characteristics or exposures over time, or understanding job ladders and transitions, the current approach generally begins with creation of crosswalks to harmonize codes over time. Due to the differences over time in (1) the existence of certain types of jobs and (2) granularity of occupational categorizations in some areas, large numbers of detailed occupations may end up combined in broad categories that do not allow one to apply different characteristics to different detailed occupations, or to clearly indicate occupational changes over time.

A potential solution to this problem is to recode all occupation and industry data using the current 2010 Census coding schemes. All historical and current data would then be

compatible and more useful to researchers and policymakers. Hand-coding of occupation and industry, as has been done historically and currently at the HRS, is time-consuming, expensive, and has fairly low inter-rater reliability. This report presents evidence of the feasibility of using the National Institute for Occupational Safety and Health (NIOSH) Industry and Occupation Computerized Coding System (NIOCCS) to code historical occupation and industry data in the HRS. We provide evidence on the reliability of the machine coded data. The availability of consistent occupation and industry codes for all or much of the lifetime careers of HRS respondents would open up many important research questions concerning the determinants and consequences of the changing U.S. occupational structure. This may be especially useful for studying changes to the cognitive and physical demands of work on the economic security and physical and brain health of workers as they enter retirement.

Background

The HRS is a longitudinal biennial observational study of a national sample of adults older than age 50 in the U.S. As the primary successor to the Retirement History Study (Ireland 1972), the HRS has a primary objective to provide information on individuals' well-being before, during, and after retirement transitions, with a particular focus on the role of Social Security in retirement security. To do so, two birth cohorts were originally enrolled: in 1992, the HRS cohort (born 1931-1941) and in 1993, the Asset and Health Dynamics Among the Oldest Old (AHEAD) cohort (born before 1924). Then in 1998 the two cohorts were combined and two more cohorts were added: Children of the Depression (CODA, born 1924 to 1930) and War Babies (born 1942 to 1947). Since then, a new six-year birth cohort is added every six years for a steady-state design. Detailed information is obtained across a number of domains reflective of a multidisciplinary holistic view of well-being. As noted, a main focus has been employment and

employment transitions. Therefore, one area of detailed measurement is work history and current employment, including occupation and industry of all jobs reported. HRS respondents describe their occupations and industries verbally, and these descriptions are recorded verbatim by interviewers and serve as the basis for manual coding using detailed census codes. The detailed occupational data are available to users only under restricted conditions of use, while highly aggregated masked codes are available as part of the public data. Finally, an additional source of occupational information comes from a new HRS supplemental study, the Life History Mail Survey (LHMS), which obtains a retrospective employment history for jobs held early in the lives of HRS respondents.

While HRS asks a limited set of questions about past and current job characteristics (e.g., extent to which the work is physically demanding, involves intense concentration, involves working with computers, etc.), researchers increasingly are interested in linking the HRS data on occupations to the rich information on work available in the Occupational Information Network (O*NET) database. Sponsored by the U.S. Department of Labor's Employment and Training

Administration, the O*NET database provides detailed occupational information that was compiled from a combination of surveys among workers, expert assessments, and tests (Peterson et al. 1999). This linkage is accomplished by matching occupations in the HRS to occupations in O*NET using detailed occupational codes.

Over the years, multiple code frames have been used to classify occupation and industry in the HRS. Table 1 summarizes the three different coding regimes that have been used.

Table 1. Codeframe information for unmasked occupation and industry data in HRS, 1992 to 2016

HRS Core Interview Wave	Census Occupation Codeframe	Census Industry Codeframe
1992-2004	1980 Census	1980 Census
AHEAD (1993 and 1995)	SRC Masked only	not asked
2004-2010	2000 Census	2002 Census
2010-2016	2010 Census	2007 Census

As can be seen from Table 1, researchers who wish to use the data longitudinally must construct a crosswalk that links these different coding schemes. This is not straightforward because many new occupations have arisen in recent years, and others have declined and, therefore, been subsumed into broader categories under the newer codeframes. Researchers wishing to construct crosswalks have been required to spend a lot of time attempting to sensibly link these varying codes over time. It should be noted that this process relies on judgement of individual researchers and may result in inconsistencies between researcher-created crosswalks and the resulting research findings.

For example, in our work¹, we have used and augmented one developed by Peter Hudomiet (available [here](#)), which works reasonably well and in the HRS data (1992 to 2010) yields 192 separate occupational categories (down from 535 total in the 2010 Census framework). Importantly, 22 of these categories are groups of “other” occupations, e.g., *other financial specialists*, *other managers*, *other health and therapy occupations*. These tend to be relatively large (holding many codes), heterogeneous groupings that are less closely-related and less consistent than the categories that can be more closely linked over the three regimes. For example, the “other” categories contain some occupations that were part of the 1980 coding scheme but are not included the 2000 scheme, while others first emerged in 2000. Also, the

¹ Note that others have reported on alternative crosswalks; see for example, Frederick (2010).

“other” categories contain occupations with smaller cell counts that are otherwise hard to group. These issues mean that the O*Net variables for these “other” occupational categories may be subject to significant measurement error. These problems led the HRS team to identify recoding historical occupation data up to the 2010 regime as a means to improve intertemporal consistency. If this were feasible, in theory, such a variable could be added to the HRS restricted industry and occupation code data.

It should be noted that intertemporal inconsistency due to crosswalking historical data is only one source of error in occupational data; other important sources of error include miscoding by coders which speaks more directly to the initial validity of the coded results (Speer 2016).

As Conrad, Couper, and Sakshaug (2017) note, there are three basic occupation coding methods: manual, computer-assisted, and automated coding. These researchers and other groups are increasingly documenting the process of computer-assisted and automated occupational coding (e.g., Bethmann et al. 2014; Cheeseman Day 2014), generally reporting on comparisons to manually coded occupational data (Belloni et al. 2016; Patel et al. 2012). Such is the growing interest and popularity of automated coding, one recent report simply compares three different methods of automated occupational coding (Gweon et al. 2017). Interestingly, studies differ on whether they hold the human or the machine as the “gold standard.” Some researchers have suggested that the most valid coding may come from a combination of human and machine efforts (Belloni et al. 2016; Ossiander and Milham 2006).

Some research has addressed the impact on substantive findings of different occupational coding methods. In the field of industrial hygiene, for example, some research examines the reliability of occupational coding for determining occupational exposures as risks for disease (Koeman et al., 2012). Here there is high interest in developing automated coding methods given

the large numbers of potential interesting records, such as occupations coded in vital records. Ossiander and Milham (2006) created a simple, word-matching computer program to code occupation information provided on death certificates in Washington state. Code rates were 96 to 97 percent and matched the manual code 89 percent of the time. Similarly, another study reported on manual versus automated coding of occupation data and assignment of occupational exposure in a large governmental database, finding that automated coding of occupations results in assignment of exposures in reasonable agreement with results from manual coding (Burstyn et al. 2014).

The present research tested the feasibility, match rates, and reliability of NIOCCS occupation and industry codes relative to manual coding in the HRS. Specifically, we first used retrospective employment information provided by respondents in the Life History Mail Survey (LHMS), an off-year mail survey of HRS participants, to examine the usefulness of the NIOCCS tool for coding retrospective occupation and industry data. Second, we ran a pilot using the NIOCCS tool on a subset of HRS data previously coded using 1980 three-digit census occupation and industry codes.

Data and Methodology

Three main HRS data sources were used in this project: 1) 1992 restricted occupation and industry data and raw text from 1992 “gray cards,” 2) the raw occupation and industry 2017 Life History Mail Survey, and 3) 2010 raw occupation and industry data.

1992 data and gray cards

The first HRS survey interviews fielded in 1992 were all paper and pencil. While none of the original paper copy of interview data is available, nor the original scans, HRS did retain the

gray cards. The gray cards were employed to gather contact information that would later be used to ask employers for respondents' pension details. Gray cards contain hand-written employer name and address, as well as job title for each respondent's current, last (asked if there is no current job), and previous job (generally only asked if there is a current job). Additionally, detailed industry and occupation codes are available from 1992. These relied on data provided in the main questionnaire.

We located 8,193 gray cards, representing 65.3 percent of the original HRS participants interviewed in 1992. All cards were scanned and subsequently transcribed by a team of trained research assistants. We then sent these raw data to a human coder, who coded data from a sample of 934 gray cards containing information about the "current" job. We focus on these cases in our analysis.

The Life History Mail Survey (LHMS)

To fill in the gaps in early- and midlife information on educational, residential, employment, partnership, and health history not obtained in the HRS core survey, an off-year mail survey was fielded between 2015 and 2017 to all HRS participants, except the Late Baby Boomer cohort impaneled in 2016. Relevant to the present study, the LHMS contains an employment grid that participants completed for up to 10 jobs held "for one year or more after you finished full-time education." The first column asks for the "type of employer or business" and the second column asks for "job title." These two fields were transcribed by trained research assistants who corrected obvious misspellings and expanded abbreviations. The data used in this project were from the spring 2017 fielding of the LHMS, which gathered data on a subset of 3,844 HRS participants.

2010 raw occupation and industry data

The 2010 occupation and industry data contained detailed descriptions of “what does this business do or make,” as well as job title information for HRS respondents who were currently working in a new job in that year. We selected a random sample of 2,000 observations from this dataset to test with NIOCCS.

Automated coding

NIOCCS, an automated coding engine, translates text to standardized codes. The system was developed by NIOSH and released for public use in December 2012. NIOCCS has been continuously updated, and we used NIOCCS v. 3.0 in this project. NIOCCS employs a machine-learning algorithm that uses a large dataset of manually coded occupations as training data for automatic classification. Its stated purpose is to provide a tool that reduces the high cost of manually coding occupation and industry information while improving uniformity of the codes.²

We were in close consultation at every step of this project with NIOSH staff. They instructed us on the exact file structure and format needed for the coding engine. We also obtained a data use agreement from the University of Michigan to ensure protection of respondent confidentiality. The data were then uploaded to their system through a secure file transfer protocol. The output of the automated coding was returned to us via the same transfer protocol.

² More information about NIOCCS:

<https://www.cdc.gov/niosh/topics/coding/how.html>

<https://www.cdc.gov/niosh/topics/coding/NIOCCSUserDocumentation.html>

Analysis plan

We prepared three pilot-test data sets that included 1,000 randomly selected job entries for which either industry information or occupation information, or both, were available in the raw data. Two of these datasets used different sources of data from the 1992 interviews, while one used the data from the LHMS. (The 2010 data were not yet available for our use at this time.) Based on the results of these pilot tests, we then uploaded a complete dataset for all of the 2017 LHMS observations (N=14,947 jobs) and for the 1992 gray card set (N=15,995 jobs). Based on the two tests of 1992 data, we determined that using the “job title” information from the gray cards, plus the label from the 1980 Census industry code assigned during the original coding of the 1992 data, yielded the best information for both the manual and NIOCC coding. Specifically, the gray cards contained employer name and address, but that information was often inadequate for coding; the 1992 questionnaires (since destroyed) contained the descriptions of what the businesses did that were used in the original industry coding.

Appendix A provides an overview of the various data sources and the tests we conducted prior to uploading the full datasets for coding. The first two rows describe information the manual coder used. This information was quite simple in the LHMS, but slightly more involved for the 1992 data. A key difference between the LHMS and 1992 gray card industry data was that, in 1992, the gray card recorded the name of the business rather than type of employer (as in the LHMS). The manual coder reported that type of employer was more useful than name of business. For example, business name might be Acme Inc., which is quite generic. Even an internet search on this firm and its location does not guarantee good information about what this business did in 1992. In addition to these 1992 data fields, we also obtained the original codes from the HRS restricted data to which we applied the census code labels. While these were

available for both industry and occupation, we asked the manual coder to use only the industry code and label to make it more comparable to what the automated coding engine would use.

The middle section of the table describes the data submitted to NIOCCS for the pilot tests. For the 1992 gray card data, Test 1 provided business name and job title. Test 2 provided job title from the gray card and industry label from the 1992 restricted industry codes. From the LHMS, Test 3 provided type of business and job title. For the 2010 HRS, Test 4 (conducted later) provided job title and interviewer transcription of responses to the question “what does this business do or make?” Based on the results of Tests 1 through 3, our final comparisons between NIOCCS and the manual coder used results from 1992/gray card coding using job title and industry label from the 1992 manual coding as inputs, while the LHMS data simply used the data provided, namely, job title and type of employer.

We examined the relative frequency of entries auto-coded versus those entries that failed the auto-code criteria (code rates). Next, we compared the HRS manual coder’s results to the NIOCCS automated output for each set of comparisons described above. We calculated the Kappa coefficient to indicate the level of agreement across these two coding methods.

Results

Table 2 displays the code rates for both the manual and the automated coding for the LHMS data and the 1992 gray card data. Not surprisingly, the manual coder was able to code at a higher rate. The automated engine was able to code at a reasonably high rate, between approximately 60 and 72 percent.

Table 2. Code rates

	LHMS (14,947 jobs)		1992 data (15,995 jobs; 934 hand-coded)	
	Industry	Occupation	Industry	Occupation
Manual (SRO)	99.0%	95.8%	>100%	97.4%
Automated (NIOCCS)	70.9%	70.1%	71.7%	60.0%

Note: In some cases, the human coder was able to infer industry based on occupation even when the raw industry input was not available, making the industry code rate better than 100 percent.

Table 3 presents the Kappa coefficients for agreement between the manual coder and the NIOCCS automated coding in the LHMS samples for both unmasked (detailed) and masked (broader) codes. As an indicator of agreement between the two sources, Kappa coefficients in the ranges obtained for the masked data are considered quite good (78 to 83 percent) but are on the lower end of acceptable for the unmasked data.

Table 3. LHMS: Kappa coefficients for manual and automated coding (for cases in which both were able to code)

	Agreement	Expected agreement	Kappa	Std. Err.	Z	p-value	Obs.
Unmasked industry	67.46%	2.15%	0.6674	0.0014	479.07	0.0000	10,527
Unmasked occupation	68.08%	1.18%	0.6770	0.0011	639.50	0.0000	10,299
Masked industry	78.55%	8.78%	0.7649	0.003	258.77	0.0000	10,454
Masked occupation	83.29%	8.91%	0.8165	0.003	276.42	0.0000	9,956

Note: Masking categories provided by HRS are not complete. Thus, some observations are lost upon masking and the sample sizes for “masked” data are smaller.

Table 4 also presents the Kappa coefficients for agreement between the manual coder and the NIOCCS automated coding in the 1992/gray card samples for both unmasked (detailed) and masked codes. The results here are substantively similar to the findings for the LHMS comparisons.

Table 4. Gray card “current” jobs: Kappa coefficients for manual and automated coding

	Agreement	Expected agreement	Kappa	Std. Err.	Z	p-value	Obs.
Unmasked industry	63.17%	2.60%	0.62	0.006	100.15	0.0000	630
Unmasked occupation	69.32%	0.02%	0.69	0.006	122.87	0.0000	590
Masked industry	81.39%	11.54%	0.79	0.015	53.62	0.0000	548
Masked occupation	84.35%	9.62%	0.83	0.013	63.6	0.0000	575

Note: Masking categories provided by HRS are not complete. Thus, some observations are lost upon masking and the sample sizes for “masked” data are smaller.

After we uploaded the 2010 data for Pilot Test 4, we learned from NIOSH staff that the record length for any given field that NIOCCS will read is capped at 250 characters, and in fact NIOCCS performs optimally with just one to three words in the text field. The LHMS and 1992/gray card data therefore worked reasonably well within these constraints. However, in the 2010 to 2016 waves of HRS, the raw industry data were incompatible with NIOCCs due to the long descriptions recorded by interviewers for the question, “What does this business do or make?” Therefore, we did not conduct this set of comparisons. Given that the occupations field would work, a possible solution is to substitute the industry codes for the incompatible industry field. This would allow us to conduct our comparisons but would not be especially useful for reducing coding costs in the future without making changes to the way the HRS asks about “what this business does or makes.”³

³ However, according to research by Conrad, Couper, and Sakshaug (2017), while machine-code rates may be higher with shorter industry descriptions, reliability of codes using shorter descriptions is lower. Thus, a decision to change the HRS in this manner would likely represent a trade of lower quality for lower cost.

Conclusion

In summary, results of automated reading and coding of HRS data are encouraging. We found that NIOCCS works well only with short descriptions, one to three words each, of job title or job description and “what a business does or makes” as inputs, a finding that is in accord with recent research (Ellison 2014; Conrad, Couper, and Sakshaug 2017). NIOCCS performs reasonably well compared to coding results from a highly-trained, professional occupation and industry coder, with Kappa inter-rater reliability on detailed codes of just less than 70 percent and agreement rates on broader codes of around 80 percent. The main weakness of NIOCCS appears to be its failure to produce codes in many cases. Code rates for NIOCCS for the datasets tested ranged from 60 percent to 72 percent, as compared to a professional coder’s ability to code those same datasets that ranged from 95 percent to 100 percent.

Our findings line up reasonably well with other studies employing a similar methodology. For example, using raw data on occupations from various archival records in the Atherosclerosis Risk in Communities (ARIC) study, Patel et al. (2012) found automated code rates (using the Standardized Occupation and Industry Coding software program) of 71 percent for occupations and 76 percent for industry. Agreement between automated and manual coding was 73 percent for occupational codes and 69 percent for industry codes. Agreement was higher for the major (masked) occupational groups (89 percent). Belloni et al. (2016) used data from the Dutch sample of the Survey of Health, Ageing and Retirement in Europe (SHARE) and compared hand-coded, verbatim responses on current and last occupation with the Computer Assisted Structured Coding Tool (CASCOT) occupation coding software. They found that use of ancillary information dramatically increased more fine-grained coding in this combination. In accord with what we conclude, both of these studies suggest that a certain amount of manual

coding will always be required. As Belloni et al. (2016) note: “We believe that the combination of a high-quality software program (which automatically coded a high proportion of cases at the four-digit level...), an expert coder, the use of ancillary information, and the use of an extensive external job titles list ensured a high level of coding and provided better coding than manual SHARE coding.” This level of investment seems likely to yield not only reliable but the best chance at valid coding of occupational data.

Future directions

NIOCCS may be a useful tool for reducing the human-coder hours needed for coding industry and occupation data for the HRS and other studies and datasets. In its current form, it would be most useful as a way to reduce the number of cases a human coder must code, the amount of time a human coder must spend on each case, or as a first cut for coding historical data that do not crosswalk cleanly to a newer codeframe. For example, with respect to the historical data, which are currently in three different codeframes, we suggest that it would be most cost effective and reliable to do the following: 1) use a crosswalk to link only those entries that cleanly map into the newest codeframe for detailed occupations; (2) for the remaining entries, use job title and the label for the industry code that was assigned in the historical data to run through NIOCCS; and (3) manually code those cases NIOCCS is unable to autocode; then review all autcoded results and crosswalks.

The HRS is currently developing and testing a semiautomatic coding application using algorithms based on the manual coder’s choices from the past three waves of HRS occupation coding to automatically provide likely codes for new cases. The main objective with this project is to provide a scaffold of sorts to new coders, which would potentially allow expansion of the pool of coders, thereby expediting the coding process. When a coder opens a new case to code,

the new system will show the five highest probability occupations and industries for that case. Preliminary results suggest that, on the five highest codes, there is 76 percent agreement between “man” and “machine.” The first machine-provided code agrees 50 percent of the time with a highly-trained coder, the second 9 percent, and then descending percentages adding up to 76 percent. According to the head human coder at HRS, “the new system allows us to put in notes about what is wrong with the case. It also allows us to separate bad cases, and to copy text to put into a translator. Those make data quality a lot better, and are important features on the ‘nuts and bolts’ level.” All efforts to improve data quality are of paramount importance given research suggesting that “miscoding” rates — that is, the code assigned (regardless of method) is not correct — may be as high as 28 to 30 percent. Belloni et al. (2016) suggest that use of semiautomatic software, such as that being developed in-house at HRS, hold much promise for improving data quality. Finally, another exciting new direction that holds promise is to code job titles during CAPI interviews using a look-up table or coding index, such as Belloni and others (2016) have described in SHARE.

References

- Angrisani, M., Kapetyn, A. and Meijer, E. (2015). *Nonmonetary job characteristics and employment transitions at older ages* (Working Paper 2015-326). Retrieved from University of Michigan Retirement Research Center website:
<http://www.mrrc.isr.umich.edu/publications/papers/pdf/wp326.pdf>
- Angrisani, M., Hurd, M., Meijer, E., Parker, A. M., and Rohwedder, S. (2017). Personality and employment transitions at older ages: Direct and indirect effects through non-monetary job characteristics. *Labour*. Advance online publication. doi: 10.1111/labr.12090
- Belloni, M., Brugiavini, A., Meschi, E., and Tjidsens, K. (2016). Measuring and detecting errors in occupational coding: An analysis of SHARE data. *Journal of Official Statistics*, 32(4), 917–945. <https://doi.org/10.1515/jos-2016-0049>
- Bethmann, A., M. Schierholz, K. Wenzig, and M. Zielonka. (2014). Automatic coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys. In: Statistics Canada (Ed.), *Beyond traditional survey taking. Adapting to a changing world. Proceedings of Statistics Canada Symposium 2014, Quebec*. Available at: <http://fdz.iab.de/342/section.aspx/Publikation/k151124301>
- Burstyn, I., Slutsky, A., Lee, D. G., Singer, A. B., An, Y., and Michael, Y. L. (2014). Beyond Crosswalks: Reliability of Exposure Assessment Following Automated Coding of Free-Text Job Descriptions for Occupational Epidemiology. *The Annals of Occupational Hygiene*, 58(4), 482–492. <https://doi.org/10.1093/annhyg/meu006>
- Cheeseman Day, J. (2014). Using an Autocoder to Code Industry and Occupation in the American Community Survey. Presentation held at the Federal Economic Statistics Advisory Committee Meeting, 13 June 2014. Available at: http://www2.census.gov/adrm/fesac/2014-06-13_day.pdf.
- Conrad, F. G., Couper, M. P., and Sakshaug, J. W. (2016). Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes. *Journal of Official Statistics*, 32(1), 75–92. <https://doi.org/10.1515/jos-2016-0003>

- Ellison, R. 2014. Demonstration of Performance of CASCOT 5.0. Presentation held at the CASCOT: Occupational Coding in Multi-National Surveys Workshop, 10–11 April 2014, Venice. Available at: <http://dasish.eu/dasishevents/cascotworkshop/programmepres/>
- Fisher, G. G., Infurna, F. J., Grosch, J., Stachowski, A., Faul, J. D., and Tetrick, L. E. (2014). Mental work demands, retirement, and longitudinal trajectories of cognitive functioning. *Journal of Occupational Health Psychology*. 19(2), 231-242. doi:10.1037/a0035724
- Frederick, C. (2010) A crosswalk for using pre-2000 occupational status and prestige codes with post-2000 occupation codes. Center for Demography and Ecology Working Paper No. 2010-03.
- Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, S. (2017). Three Methods for Occupation Coding Based on Statistical Learning. *Journal of Official Statistics*, 33(1), 101–122. <https://doi.org/10.1515/jos-2017-0006>
- Irelan, L. (1972). *Retirement History Study: Introduction* (Social Security Bulletin). Retrieved from <https://www.ssa.gov/policy/docs/ssb/v35n11/v35n11p3.pdf>
- Koeman, T., Offermans, N. S. M., Christopher-de Vries, Y., Slottje, P., Van Den Brandt, P. A., Goldbohm, R. Alexandra, ... Vermeulen, R. (2012). JEMs and Incompatible Occupational Coding Systems: Effect of Manual and Automatic Recoding of Job Codes on Exposure Assignment. *The Annals of Occupational Hygiene*. <https://doi.org/10.1093/annhyg/mes046>
- McFall, B.H., Sonnega, A., Willis, R.J., and Hudomiet, P. (2015). Occupations and work characteristics: Effects on retirement expectations and timing. Michigan Retirement Research Center Working Paper Series WP 2015-331.
- Ossiander, E.M. and Milham, S. (2006). A computer system for coding occupation. *American Journal of Industrial Medicine*, 49, 854-857.
- Patel, M. D., Rose, K. M., Owens, C. R., Bang, H., and Kaufman, J. S. (2012). Performance of automated and manual coding systems for occupational data: A case study of historical records. *American Journal of Industrial Medicine*, 55(3), 228–231. <https://doi.org/10.1002/ajim.22005>

Peterson, N. G., and American Psychological Association (Eds.). (1999). *An occupational information system for the 21st century: the development of O*NET* (1st ed). Washington, D.C: American Psychological Association.

Sonnega, A., McFall, B.H., and Willis, R.J. (2016). Occupational changes at older ages: What moves are people making? Michigan Retirement Research Center Working Paper Series WP 2015-352.

Willis, R.J. (2013). The cognitive demands of works and the length of working life: the case of computerization. Working Longer and Retirement Conference Stanford Institute for Economic Policy Research October 10-11, 2013.

Appendix A. Overview of Data Sources and Coding Efforts

<u>HRS manually coded data sources</u>	Industry coding input	Occupation coding Input	Additional data for industry coding: codes	Additional data for industry coding: description
1992 Gray Cards	Business name from gray card, for example "Acme Inc"	Job title from gray card, for example, "janitor"	1992 restricted data industry code (1980 Census code); crosswalked industry code to 2007 Census frame	1992 restricted data industry label; label from crosswalked industry code to 2010 Census frame
LHMS	"Type of employer or business"	"Job Title "		
<u>NIOCCS Pilot Tests</u>				
Gray Cards				
Test 1	Business name from gray card, for example "Acme Inc"	Job title from gray card, for example, "janitor"		
Test 2		Job title from gray card, for example, "janitor"		1992 restricted data industry label
LHMS: Test 3	"Type of employer or business"	"Job Title "		
2010 HRS Core: Test 4	"What does this business do or make?"	"Job Title"		
<u>Final NIOCCS Coding</u>				
Gray Cards Final (like test 2)		Job title from gray card, for example, "janitor"		1992 restricted data industry label
LHMS Final (test 3)	"Type of employer or business"	"Job Title "		