DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 75–92, http://dx.doi.org/10.1515/JOS-2016-0003

Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes

Frederick G. Conrad[1], *Mick P. Couper*[1], *and Joseph W. Sakshaug*[2]

A source of survey processing error that has received insufficient study to date is the misclassification of open-ended responses. We report on efforts to understand the misclassification of open occupation descriptions in the Current Population Survey (CPS). We analyzed double-coded CPS descriptions to identify which features vary with intercoder reliability. One factor strongly related to reliability was the length of the occupation description: longer descriptions were less reliably coded than shorter ones. This effect was stronger for particular occupation terms. We then carried out an experiment to examine the joint effects of description length and classification "difficulty" of particular occupation terms. For easy occupation terms longer descriptions were less reliably coded, but for difficult occupation terms longer descriptions were slightly more reliably coded than short descriptions. Finally, we observed as coders provided verbal reports on their decision making. One practice, evident in coders' verbal reports, is their use of informal coding rules based on superficial features of the description. Such rules are likely to promote reliability, though not necessarily validity, of coding. To the extent that coders use informal rules for long descriptions involving difficult terms, this could help explain the observed relationship between description length and difficulty of coding particular terms.

Key words: Survey processing error; coding error; occupational classification.

1. Introduction

Survey responses are imperfect measures. The origins and implications of response error are increasingly well understood (e.g., Sudman et al. 1996, ch. 2; Tourangeau et al. 2000, ch. 1) but the vast majority of this knowledge concerns closed-form questions, i.e., questions that present response options which respondents select to report their answers. However, essential information for official statistics is derived from open responses, i.e., answers reported in respondents' own words. These open responses are coded – assigned to categories – in order to be quantified, and the coding process can introduce

[1] Survey Research Center, Institute for Social Research, University of Michigan, PO Box 1248, Ann Arbor, MI 48106, U.S.A., and Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, MD 20742, U.S.A. Emails: fconrad@umich.edu and mcouper@umich.edu
[2] Cathie Marsh Institute for Social Research, University of Manchester, Humanities Bridgeford Street Building, Manchester M13 9PL, U.K., and Institute for Employment Research, Regensburger Str. 104, Nuremberg 90478, Germany. Email: joe.sakshaug@manchester.ac.uk
Acknowledgments: We wish to acknowledge the assistance and advice of Adelle Belinger, Laverne Bowman, Cathy Dippo, Janet Harmon, Fred McKee, George Mitcham, Ted Sands, Ron Tucker, John Van Hoewyk, and Greg Weyland. We are particularly indebted to the coding staff at the Census Bureau's Data Processing Center for its participation in the study reported here. We thank Lars Lyberg for making helpful suggestions about an earlier version of this article. Finally, we are grateful to four anonymous reviewers and the Associate Editor for their valuable comments that improved the article.

error to the survey data. Coding error does not necessarily mean that a code is "inaccurate." Responses can vary in the degree to which they belong in particular coding categories, i.e., category membership is graded (e.g., Barsalou 1985), so simply assigning a code to an open response can involve error because the fit is not perfect, even if there is no better fitting code.

One domain in which open responses are essential is the measurement of occupation. The coding of occupation has many unique features compared to coding other types of open responses: coders must assign a description to one of hundreds of possible codes rather than just a handful, the open answers are short and factual rather than long and attitudinal, and occupation coders who make this their profession become very skilled. The data about occupation that are produced through the coding process are commonly used to study various phenomena occurring in the labor force, including sex segregation (Anker 1998), work-related injuries (Cawley and Homce 2003; Layne 2004; Reichard and Jackson 2010), health-related exposures (Kromhout et al. 1993; Hammond et al. 1995; Kauppinen et al. 2000), wage inequality (Lettau 2003; Heywood and O'Halloran 2005; Bjerk 2007), mobility (Shniper 2005; Moscarini and Thomsson 2007), and occupational projections (Rosenthal 1992).

Despite differences in the details of coding open responses in different domains, there seem to be certain commonalities in the ways coders contribute to overall error. Coder error is a type of processing error (Biemer and Lyberg 2003, 234–241) that is introduced by coders interpreting respondents' verbalization of their thinking. By some measures, it can substantially inflate other sources of error. For example, in a study of time use, Sturgis (2004) demonstrated that correlated coder error – analogous to interviewer variance – nearly doubled the size of standard errors on average, across ten activity categories with a maximum inflation factor of 3.36. Classification of open reports may be compromised because, for example, coders may fail to consider key information such as the size of different categories and, thus, the probability of membership. Base rates are just one consideration when all else is equal or there is no way to choose between alternative classifications. Alternatively, the quality of coding may be degraded because the coding categories and the rules for using them are too rigid to adequately address the ambiguity inherent in people's descriptions; for example, respondents' descriptions may fit well into more than one category or may not fit well into any category but must be assigned to one nevertheless.

There are various techniques for coding open-ended responses into different categories. The majority of these techniques can be classified into three distinct groups: manual coding, computer-assisted coding, and automated coding. In manual coding, respondents provide their response in free text and coders assign codes based on a standardized classification system. In computer-assisted coding, coders provide codes by means of assistance dictionaries, which provide lists of possible answers for the coder to choose from (Bushnell 1995; Lyberg and Kasprzyk 1997). In automated coding, the software codes some open-ended responses without human intervention, and manual (or computer-assisted) coding is used to code the unresolved balance (Lyberg and Dean 1992; Macchia and D'Orazio 2001; Esuli and Sebastiani 2010).

Using occupation as an example, in the current work we explore how (1) the characteristics of respondents' open-ended reports – in particular the length of the reports and

the difficulty of classifying descriptions that include particular words – and (2) the practices of coders may affect the quality of *computer-assisted* coding of open-ended answers.

The occupation coding task that we study here involves assigning respondents' occupation descriptions, collected from open-ended reports of their occupation and duties, to categories in the US Census occupational classification system, which is derived from the Standard Occupational Classification System (SOC) and has been updated several times. The SOC is one of several standard classification systems established by the US Office of Management and Budget, and is used to publish comparable occupational data for statistical purposes across the U.S. federal statistical agencies. The data set used in this study involved codes from the 1990 Census system, derived from the 1977 SOC. The 1990 Census system included 501 detailed categories, expressed as three-digit codes, 13 major occupation groups, expressed as two-digit codes, and six summary groups, expressed as one-digit codes. There are some differences between the 1977 SOC and the most recent version, the 2010 SOC. The number of major occupation groups has increased from 21 to 23, and the number of detailed occupations has increased from 662 to 840. A detailed description of the revision history and process can be found online (Bureau of Labor Statistics, 2014).

While occupation is one important domain in which the data are based on coded open-ended responses, coding open responses is ubiquitous throughout survey research, producing important data in domains ranging from public opinion (e.g., most serious problems facing the country) to time use (activities) to academic fields of study. To the extent that the coding of open responses across domains and question types is similar to the coding of occupation descriptions, what we learn about occupational coding may inform how we think about coding open responses in other applications.

2. Classification and Measurement

The coding process may introduce error in several ways having to do with (1) the "in or out" requirement of a formal classification system in a world where categories are ever changing and where membership is a matter of degree, (2) lack of category size information to help inform coder decisions when a description seems to fit two categories equally well, (3) ambiguity of particular words that respondents use in occupational descriptions, and (4) length of occupation descriptions.

In everyday classification tasks, people can modify categories to accommodate instances that are atypical. However, formal classification systems are more rigid than this. In a demonstration of the flexibility of everyday categories (as opposed to a formal classification system like the SOC), Kunda and Oleson (1995) found that when presented with descriptions of cases that deviate from the stereotype (e.g., an introverted lawyer), under the right circumstances people preserved the integrity of the main category (lawyer) by creating subtypes (a category for introverted lawyers). Under other circumstances, they redefined the main category to include deviant cases (lawyers in general were rated as more introverted than they were if no deviant case had been presented). A coder using an established classification scheme would not be able to accommodate the deviant case in either of these ways, but instead would have to assign it to a category despite the poor fit.

When respondents' answers are ambiguous, that is, could be assigned to more than one category, coders may systematically assign answers to the wrong category. Tversky and Kahneman (1983) found that if an instance sounds like it could belong to the conjunction of two everyday categories (e.g., a feminist bank teller) but could also be assigned to one of the individual categories (bank teller), people are more likely to judge such instances to be members of the conjunction than the individual category. They call this the *conjunction fallacy* because there are more bank tellers than feminist bank tellers in the world, yet people seem to give more weight to similarity of the description to the category prototype (sounds more like a feminist bank teller than a bank teller) than to the size of the category (there are more bank tellers than feminist bank tellers). In another demonstration of this general tendency, Tversky and Kahneman (1974) found that people were more likely to judge someone who seemed like an engineer to be an engineer than to be a lawyer, even though they were told that, in the experimental scenario, there were more lawyers than engineers. They call this the *representativeness heuristic* and, while it can be a useful guideline in making some classification judgments, it desensitizes people to the base rate or size of categories when making these judgments. Coders may be similarly oblivious to category size and probability of membership when faced with descriptions that sound like particular categories, even if instances of these categories are relatively rare. This is not to say base rates and probability of category membership should be the *only* consideration that informs a coder's decision. If a description could fit equally well into two categories with very distinct meanings – for example, "secretary" could refer to a senior official of an organization or to an office assistant – a coder could be instructed to choose the category for which the odds of membership are greater rather than flipping a coin. There are more office assistants than senior officials in the world so in the absence of any additional information, considering the size of the categories would be a rational – if imperfect – strategy.

By another view, it is not flaws in coder decision making as much as the descriptions themselves that lead to lower-quality codes. Within a particular domain, some terms in respondents' descriptions may be inherently hard to code, for example, they may fit poorly into existing categories or may fit well into multiple categories. Coders may address this by developing specialized rules for classifying descriptions with problematic terms (see, e.g., Hak and Bernts 1996; Martin et al. 1995). While the use of such rules should increase agreement among coders, this could well happen without any increase in the "validity" of codes. It could be the case that a rule leads to incorrect – but consistent – codes on some occasions because it may lack the means to adjust the code on the basis of subtle changes in context. Such rules may actually lower agreement among coders if they are not defined by the group or, for other reasons, not unanimously endorsed. This is particularly likely when rules are not explicitly documented. Furthermore, one rule may conflict with another even though both seem to apply to a particular case; this too might lead to disagreement.

In addition to the inherent difficulty in coding certain terms, the length of respondents' answers may also affect how well they are coded. Couper and Conrad (1996) asked a national sample standard questions about their occupation ("What kind of work do you do, that is, what is your occupation?") and duties ("What are your usual activities or duties at this job?") from the Current Population Survey (CPS), and asked half of the sample an additional question about their job title ("What is your job title?"). When coders were able

to consider the extra response in their coding decision, their agreement with each other was lower than when they only had the initial response to work with. Cantor and Esposito (1992) report that coders prefer less information: they asked coders to listen to the interviews and indicate where additional probes would have helped them code the response. The coders virtually never asked for additional probes, suggesting they recognized that longer descriptions are harder to code than shorter ones.

More information could harm coder agreement for much the same reason that in everyday classification people prefer categories at intermediate levels of abstractness – a concept known as the *basic level* (e.g., Rosch et al. 1976). The idea is that categories that are neither too abstract nor too concrete are most useful, for example, "dog" (basic level) versus "Welsh Terrier" (more concrete) or "mammal" (more abstract). Thus description length may be a proxy for level of abstraction: longer descriptions will facilitate coding to the extent that they refer to basic-level jobs but will confuse matters if they describe overly specific categories.

On the other hand, longer descriptions are likely to be more specific than shorter ones – more words probably convey more detail. Perhaps for this reason, longer open responses are often assumed to be of higher quality than shorter ones across a variety of domains (e.g., Andrews 2005; Smyth et al. 2009; Israel 2010). Moreover, according to the 1997 CPS Field Interviewers Manual, as well as the current manual (U.S. Census Bureau, 2013), interviewers are told that

> One-word responses to the question on occupation (for example, clerk, manager, nurse, engineer, teacher) are usually far too general to be coded accurately. Whenever very brief responses are given, probe to obtain a more specific response.

So, one can imagine agreement would be higher for longer, or at least more detailed, descriptions: with more detail, there is less opportunity for two coders to interpret the description differently.

2.1. *Measures of Coding Quality*

Just as in assessing the quality of closed responses, *validity* and *reliability* are generally used to characterize the quality of open responses. However, the notion of validity is not as straightforward when applied to coded open responses as it is with respect to closed responses, at least for facts and behaviors such as one's job title or one's duties at work. Validity of closed responses can be determined, in principle, by comparing the responses to a gold standard such as a set of administrative records; with coded data, validity is typically operationalized as agreement with an expert. (With automated coding, validity is typically defined as matching an open response to text in a reference dictionary that maps text examples to categories, e.g, Macchia and D'Orazio 2001). This has much of the character of an agreement or reliability measure: a valid code matches another code that is treated as the gold standard; if there is not agreement the response is considered not valid. This lacks the potential for objective verifiability that is part of response validity for closed (factual) responses.

Reliability is simpler in concept – agreement between two or more classifications of an open response – but it is less definitive than a validity measure in that two or more coders

can agree with each other without necessarily being "correct." They can both be "wrong", assuming the correct category is known or is knowable.

3. Current Study

In the current study we focus on characteristics of respondents' occupation descriptions from the CPS that might affect the quality of codes. As noted above, the CPS descriptions come from one question about occupation, "What kind of work do you do, that is, what is your occupation?" and one about duties, "What are your usual activities or duties at this job?" After filtering out "special-case" occupations for which direct mappings between descriptions and codes are provided and "combined occupations" for which, again, direct mappings between descriptions and categories are provided, coders are instructed to consider both the occupation and duties responses together in assigning a single numeric occupation code to the description (see U.S. Census Bureau 2014). Consider the following example:

OCC – Credit Manager

DUTIES – Directing operation of credit department

In a case like this, the coder is instructed to combine the word "department" from the DUTIES line with the content of the OCC line ("Credit Manager") and code "Manager, Credit Department." Much of the instruction concerns direction on how to proceed beyond an impasse. For example, if the occupation and duties lines contain contradictory information, coders are taught to use whichever is more specific. It is our assessment that the training about the actual coding *decision* is not more detailed than instructions of this sort: coders need to be very familiar with the occupation categories and use their judgment about which words are important to consider and which ones are not. In the end, the coding task relies more on coders' knowledge of the job definitions and their aptitude for determining which parts of the description to consider than on particular training in the coding procedure. Although coders' expertise in occupational classification is at the center of the classification process, the coders searched electronic indices for occupation categories corresponding to particular terms contained in the description by entering those terms into a computer. The tool did not classify the description for the coder but returned possible categories given the input. It was still the coders' decision what category best fit the description. Although not in place at the time of the current study, the Census Bureau introduced an autocoder in 2012 that provided coders with suggested classifications for particular descriptions.

The study we report has three parts. The first is designed to explore what characteristics of occupation descriptions reduce coding reliability. We analyzed twelve months of CPS occupation descriptions (March 1997 to February 1998); note that although these data were collected and coded many years before the current article was written, the Census Bureau confirmed that they currently process and code open-ended responses in essentially the same way they did in 1997–1998. These descriptions represent 32,362 cases, each of which was independently classified by two coders. More specifically, about ten percent of all industry and occupation (I&O) descriptions were double coded, that is, independently classified by a second coder. This process was conducted for quality

assurance (QA), not production purposes; that is, the original code was not affected by the second code. Once the second (QA) coder assigned a code, the initial (production) code was revealed, and the coder had to decide whether to change his or her code to the original code (assuming a discrepancy) or refer the case back to the field for more information. The first part of the study consisted of analyses of this data set, including the effects of the length of descriptions.

Second, we investigated how characteristics identified in the first part of the study jointly affect coding agreement. To do this, we created a data set of occupation descriptions that varied systematically on several characteristics and asked pairs of coders to classify them. The experiment explicitly tested the joint effect of coding difficulty of particular words in the description and the length of descriptions. To do so, we created a set of 800 occupation descriptions systematically varied on the following dimensions:

(1) Length: one, two, and three or more words
(2) Difficulty of "primary" word: easy versus hard
(3) Difficulty of "secondary" word: easy versus hard
(4) Order of primary word: first, not first.

The easy primary words were selected by taking the eight words from the QA dataset with the highest agreement ratio. The eight words chosen were: secretary, cashier, driver, cook, teacher, nurse, waitress, and carpenter. A similar process was used to select the hard primary words (high ratio of disagreement to agreement), resulting in the following selection: owner, operator, laborer, director, technician, clerk, supervisor, and administrator. The secondary words were chosen using similar procedures, that is, high ratio of agreement to disagreement and vice versa, conditioning on each of the eight easy and eight difficult primary words first. This produced equal numbers of easy-easy, easy-hard, hard-easy, and hard-hard word pairs, for example, "school nurse" would be an easy-easy word pair. We then randomly selected existing descriptions from the QA data containing these word pairs. While the QA data set contained a large number of descriptions (over 30,000), there were some word pairs for which no description existed in the data set. In these cases, we created new descriptions by adding or removing words from descriptions that partially matched the word pair. For example, if the word pair "research supervisor" was not found in a description but "laboratory supervisor" was, we used that description, including the duties, but substituted "research" for "laboratory."

These 800 descriptions were then seeded into the ongoing production coding process, using the same procedures as regular CPS coding, but with all of the experimental cases being flagged for QA coding. In this way we obtained two codes for each of the experimental descriptions from coders who were blinded to which cases came from the experimental corpus.

Finally, we examined the coders' strategies and the kind of information they brought to bear while performing the coding task. In this third part of the study, we asked coders to think aloud while classifying occupation descriptions excerpted from the set created for the second part of the study. More specifically, we selected 100 cases from the experimental corpus, and observed four coders each coding 50 cases. Multiple-word descriptions from the experiment just described were overselected as these tend to produce higher levels of overall disagreement. The authors interacted with the coders while they

were coding, asking them to think out loud about their decision-making process, and probing for reasons for specific actions, roughly following the procedure outlined by Ericsson and Simon (1993).

4. Results

The analyses are reported separately for each of the three parts of the study. In the first part, which concerned coder agreement in the QA dataset, we first report descriptive statistics about agreement and disagreement, then examine whether disagreement is concentrated at certain digits in the occupation codes. This is followed by analyses of disagreement by occupation category. Finally we examine how attributes of the description, in particular the length of the description, affects agreement. Although the second coders could change their classifications once the first coders' classification was revealed, our focus here is on the initial code assigned to each case by the two coders, a cleaner measure of agreement. While the data we examined included both industry and occupation descriptions, we only analyze agreement on occupation, that is, not industry classification.

In the second part (coding experiment) we test whether any length effects observed in the first part are replicated in the experiment. We also test whether length interacts with the "difficulty" (agreement to disagreement ratio) of words in the descriptions. In the third part (coder observation) we analyze the coders' verbal reports, in particular, monitoring for evidence of what knowledge and conventions they use to facilitate coding of ambiguous cases.

4.1. *Analysis of Agreement in Quality Assurance Data*

Table 1 contains more details on the outcome of the double-coding process. A referral implies that the coder has insufficient information to classify the case, and refers the case back to the field for more information. Our main focus is on the 2,749 occupation descriptions (8.5% of all double-coded descriptions) where both coders assigned a code

Table 1. Occupation code agreements, referrals and disagreements

Outcome	Number	Percent of all cases	Percent of nonreferred disagreements
Agreements:	27,518	85.0	
Agreement on substantive code	23,116	71.4	
Agreement on referral	4,402	13.6	
One coder refers	2,095	6.5	
Disagreements:	2,749	8.5	100.0
Disagreement on first digit	1,251	3.9	45.5
Disagreement on second digit	888	2.7	32.3
Disagreement on third digit	610	1.9	22.2
Total	32,362	100.0	

but they disagree. In 22.2% of these cases ($n = 610$), the disagreements were relatively trivial, involving only the last of three digits in the code. However, for the balance of cases the disagreements are more severe, with 45.5% ($n = 1{,}251$) involving the first digit of the code, and a further 32.3% ($n = 888$) involving the second digit. In other words, 3.9% of all cases yield genuine and major substantive disagreements between pairs of coders. As indicated before, agreement does not guarantee accuracy – both coders could be wrong – but disagreement guarantees at least one coder is wrong – both cannot be right.

Given that the coders are evaluated on both speed and accuracy, we suspected that some of the errors may be due to "slips" (e.g., Norman 1981) such as transpositions (e.g., 234 versus 243) or single-digit offsets (e.g., 123 versus 223). We found that simple transposition errors account for a very small fraction (0.2%) of discrepancies. While one-digit offsets account for almost seven percent of discrepancies, many of these may be intended: at the first digit, the substantive difference between categories is large, for example, *legal* vs. *health care*. The majority of the descriptions whose code differed by one digit (247 out of 318) involved a discrepancy on the last digit. In the occupation coding system this represents a minor distinction in the detailed coding scheme, for example, between bartenders (code 434) and waiters/waitresses (code 435). We concluded that slips of this sort are a negligible source of error in occupation coding.

Another issue we explored was whether disagreements were more likely to occur between certain occupation groups. Restricting our focus to those cases where both coders assigned a substantive code and disagreed on the summary group (i.e., the first digit), we found that 29.4% of all these disagreements occurred between two summary groups: (1) managerial and professional specialty occupations, and (2) technical, sales and administrative support occupations. A further 14.8% occurred between (5) precision, production, craft and repair occupations and (6) operators, fabricators and laborers. However, groups (1) and (2) account for only 11% of all occupation codes, while (5) and (6) account for 4.7%. So while there appears to be some clustering of disagreements, the majority of disagreements occur between all summary (first-digit) occupation groups.

While some job categories may be particularly prone to disagreement, the descriptions themselves may affect agreement. One attribute of the descriptions that is potentially relevant to coding agreement is their length. Figure 1 shows the relationship between the number of words in the occupation description (the combined responses to both the occupation and duties questions) and the percent of all cases that result in disagreements

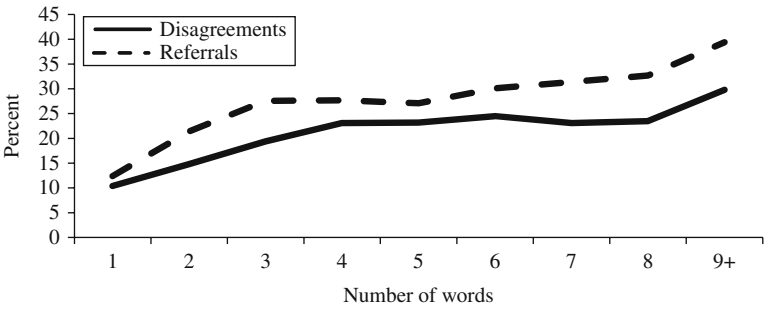

Fig. 1. Length of occupation description and disagreement and referral rates; percent is out of 32,362 cases

(including any disagreements, irrespective of the digit at which the disagreement takes place) and referrals (where at least one coder referred the case back to the field) respectively.

It is clear from this graph that the disagreement rate increases with increasing length of the occupation description. Another way to see this relationship is to compare the mean number of characters and words for the agreements and disagreements. The mean number of characters in descriptions where coders agreed is 15.3 (s.d. 8.69) compared to a mean of 18.6 (s.d. 9.92) for disagreements, a statistically significant ($t = 23.82$; $p < .001$) difference. Similarly, the mean number of words is 2.10 (s.d. 1.20) for agreement cases and 2.56 (s.d. 1.52) for disagreement cases, which is also a statistically significant difference ($t = 20.56$; $p < 0.001$). Given that referrals to the field simply requested *more* information (although coders now indicate *what* information they need), these results suggest that such an approach may actually have been counterproductive. One explanation is that more words simply create more opportunity for coders to disagree; each word is open to interpretation, and given the variability in how different people interpret the same words, the longer the description the greater the chance of disagreement. Of course there are situations in which more information can help clarify a description, for example, if the response is "teacher" and there are only three possible codes "preschool and kindergarten teacher," "elementary and middle school teacher," and "secondary school teacher," clearly more information could disambiguate the description. But if descriptions are already appropriately detailed then more information – more words – can muddle the picture unless they correspond exactly to the definition. A one-word description may well be too abstract (above the basic level) to be reliably classified, so additional words may help. However, as more words are provided – unless they are very similar to the actual definition of the job category – they are likely to confuse coders.

These results concur with Couper and Conrad's (1996) findings mentioned at the outset: they administered the standard CPS occupation and duties questions and then asked half of the sample an additional question about job title. The first-digit coder agreement rate for the standard CPS questions was 86.4% while that for the group asked the additional question was 82.1%. One explanation offered for this finding was that the addition of the job title question reduced the amount of information provided in the occupation description (combined responses of occupation and duties). In fact, the opposite occurred; when job title was asked before the occupation and duties questions, the occupation description was significantly longer than when job title was not asked (23.5 versus 18.0 characters). Furthermore, the length of the occupation description was negatively associated with coder agreement. For example, the average length of the occupation description was 20.5 characters when the coders disagreed on the first digit of the code, but 17.5 when they agreed on the code.

Similar results are presented in an unpublished report (Westat/AIR 1989): The coder agreement rate on summary (first-digit) occupation group using the standard CPS questions was 88%, but only 75% when additional job identification questions were asked. More specifically, the study compared agreement when the standard CPS questions (occupation and duties) were asked to agreement after two additional questions, including one job title probe about the identity of the respondent's job ("What was . . .'s job at [organization name]? *If necessary, probe:* What was . . .'s job title at [organization name]?") were asked. While it is not clear from the report how often the probe was

Table 2. Words with high ratios of disagreement to agreement and agreement to disagreement (ratios in parentheses)

High disagreement to agreement ratios:
Administrative (3.16); services (2.76); research (2.63), assist (2.34); maintenance (2.16); administrator (2.15); general (2.11); service (2.03)
High agreement to disagreement ratios:
Waitress (18.54); registered (8.24); guard (6.45); carpenter (6.34); electrician (5.24); secretary (5.19); accountant (5.16)

actually administered in this experimental condition, the effect of these extra questions could only have been to increase the length of the description and amount of information compared to the standard CPS approach. This again lends support to the finding that the provision of additional information (either in longer descriptions or through additional questions or probes) is associated with lower levels of coder agreement. These consistent findings that appear to run counter to common practice (seeking more information in the case of uncertainty or disagreement) are certainly worth further exploration.

To elaborate further on the length effect, we speculated that length may interact with particular occupations (or words used to describe them) in affecting agreement. For example, some occupations may be easily described using a single word (e.g., waiter), and adding words to the description may only serve to muddy the decision. On the other hand, some occupations may be inherently complex, and cannot be described adequately using only one or two words.

To test this, we separated the QA data file into agreement cases and disagreement cases, and then measured the frequency of the words in each set of descriptions. We found that words such as "administrative" were three times more likely to appear among the disagreement cases than the agreement cases (a ratio of 3.16). Table 2 provides a list of words with the highest disagreement to agreement ratio, and those with the highest agreement to disagreement ratio. Thus, for example, when the word "waitress" appears in the occupation description, there are over 18 coder agreements for every disagreement. The words for which the disagreement ratio was highest are clearly more abstract and general than the words for which the agreement ratio was highest. The more abstract a term, the larger the number of legitimate interpretations.

Given that both length and the presence of certain words affected the likely coder agreement, we sought to examine the combined effects of these two factors. We found this difficult to do as some words (e.g., "assist") rarely occurred alone, while other words (e.g., "waiter") rarely occurred in combination with other words. However, an examination of a set of selected words supports the possibility of an interaction effect. For example, when the word (or part of word) "manage" appears in the description, coder agreement declines with increasing length of description. On the other hand, coder agreement is higher when the word "operate" appears with more words. Unfortunately, the QA data set was not perfectly suited to this kind of analysis. In particular, the data set did not contain enough comparable descriptions that were both long and short and involved the same easy (high agreement ratio) and difficult (low agreement ratio) words. To address this issue we carried out an experiment, which allowed us to control the characteristics of the descriptions.

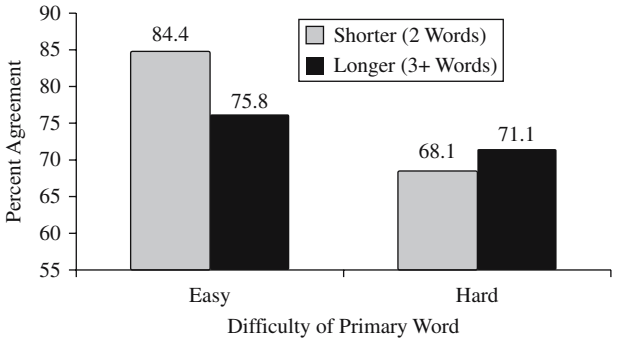

Fig. 2. Coder agreement by length of description and difficulty of primary word

4.2. Occupation Coding Experiment

The results from the experiment confirmed our earlier hypothesis that the length effect depends on the presence of certain words. When the primary word is easy, longer occupation descriptions (three or more words) decrease coder agreement, but when the primary word is difficult, longer descriptions were coded with marginally greater reliability. This is illustrated in Figure 2. When the interaction term is included in a logistic regression model, the model is statistically significant (Wald chi-square $= 10.8$, d.f. $= 3$, $p = 0.012$) and the interaction term marginally so (Wald chi-square $= 2.74$, d.f. $= 1$, $p = .098$). The secondary words' difficulty had no impact on coding reliability.

While the experimental results enabled us to elaborate on the earlier finding regarding length of occupation description, they provide little insight into *why* this pattern occurs. Our intuition was that we might gain some insight by examining coders' thinking while they make their classification decisions. Thus the last step in our investigation was to ask a small number of coders to think aloud while coding a small number of the experimental descriptions.

4.3. Coder Observations

All four coders reported following specialized coding rules that pertained to descriptions with specific characteristics or that included specific terms. The coders could not produce any of these rules in writing, nor could they provide a clear rationale for the rules or comment on their origin. These rules tended to be concerned with superficial aspects of the descriptions, rather than the concepts behind the relevant occupation or the logic of the overall classification system. Such rules are likely to increase coder agreement to the degree that they are followed by all coders on all occasions. However, there is no reason that they should improve validity, given their atheoretical character, and in fact their use may contribute to correlated coder variance (Martin et al. 1995; Sturgis 2004).

One of the rules that coders reported using could help to explain the interaction between length and difficulty that we observed in the experiment:

> When multiple occupations and multiple duties are described, select the occupation that corresponds to the duty listed first, even if it is not the first occupation.

Given the following description, this rule would dictate that the case should be coded as a driver: although "driver" is listed second in the occupation description, the corresponding duties ("delivery") are listed first.

OCC: COOK, DRIVER

DUTIES: DELIVERY, COOKING

If we assume that occupation descriptions with multiple occupations and duties are longer than descriptions with only one occupation, using such a rule could increase reliability for long descriptions, irrespective of the inherent difficulty of their component words. In contrast, for descriptions with a single (shorter) occupation, the rule does not apply; descriptions with difficult terms will therefore be less reliably coded than those with easy terms.

If such a rule is consistently applied it will improve reliability, but it may well degrade validity: in the CPS interview, respondents are first asked about their occupation, then about their duties in that occupation. This could lead to identical occupation and duty answers, which respondents might feel is uncooperative in the sense of Grice (1975). To avoid redundancy (i.e., provide different answers to the different questions), respondents may simply reverse the order of the duties relative to the occupations, irrespective of which duty best describes their occupation. Thus the order of the duties may have little substantive meaning.

Another frequently mentioned type of rule involves directly coding specific terms:

If the word "secretary" appears in the occupation line, code to secretary and ignore all other information.

The rule would require that the following description be coded as secretary, regardless of the other information it contains:

OCC: SECRETARY, CUSTOMER SERVICE ADVISOR

DUTIES: BILLING CUSTOMERS, SCHEDULING SERVICE, ADVISING

This rule seems to be given priority over other rules, so even though the first rule for multiple occupations and duties could apply here, the direct coding rule for secretary is applied. In other cases the priority is less clear. For example, one rule was:

If you see "assistant anything," drop the "assistant" and code to the other word.

But another rule stated

If you see "teacher's assistant," drop the "teacher" and code "assistant."

Yet another rule applied to "assistant to . . . ," in which case the coder was to look at the duties rather than the occupation line. Hence, the rules themselves may be contradictory under certain circumstances. Furthermore, the above suggests that the rules depend on the order in which the words appear in the occupation description.

5. Conclusions and Implications

The current study of occupation coding produced three main findings. First, we found that longer occupational descriptions were less reliably coded than shorter ones. The pattern

appeared to depend on the particular occupation terms involved. Second, for easy occupation terms longer descriptions were less reliably coded, but for difficult occupation terms longer descriptions were slightly more reliably coded than short descriptions.

The third main finding was that coders rely on the use of arbitrary coding rules based on superficial features of the description. It is possible coders' use of one of these rules – the rule for coding descriptions with multiple occupations and duties – could explain the interaction of length and difficulty observed in the experiment because the rule is most likely to be applicable to longer descriptions, making the difficulty of the words in those descriptions less important than in shorter descriptions. While such rules are likely to promote the reliability of coding, they are unlikely to improve validity. Although the rules are likely to be specific to a particular survey operation, the general phenomenon seems to be widespread (Hak and Bernts 1996; Campanelli et al. 1997).

We see several areas where concerted effort might directly improve the accuracy of coding in most survey operations. The first involves the training for interviewers and coders. The interviewers (who collect the occupation descriptions) can become more skilled at eliciting descriptions that are not unnecessarily long or overly specific, particularly for easy occupation terms. This should increase intercoder reliability and reduce the effect of length of description (as it was most problematic for descriptions containing easy terms). Concerning coders, if all coders are regularly exposed to a set of cases producing high disagreement (or low quality by some other measure), the group can discuss these cases and reach consensus on how to code them based on sound, theoretical reasons. This should increase agreement for descriptions that would previously have led to different codes from different coders. In addition, the informal coding rules of the sort we observed should be carefully evaluated and, if deemed to improve valid classification, should be formalized and made explicit; coders should be instructed to use them consistently. It should be possible to identify and document exceptions and develop ways to resolve conflicts among rules. Rules that are not found to improve the validity of classification should be explicitly discouraged. In fact, since the time of our coder observation in which we observed the use of many informal rules, some of these rules have been made explicit as "Job Aids" in the coder instructional materials (U.S. Census Bureau, 2014).

A second area ripe for improvement concerns the occupational coding software system used by coders. The CPS coding system in use when we observed coders suffered from numerous usability problems that could be identified and fixed with proper usability evaluation. As with any software, usability engineering can greatly improve the speed, accuracy and satisfaction of use. This is important because the way in which the software provides results from searches to coders could facilitate the application of incorrect rules. The appropriate rules used to resolve complex or ambiguous cases could be formally built into the software system.

Similarly, a more usable and flexible coding system might allow coders to assign a description that seems to belong to more than one job category to all appropriate categories, in the way that respondents are sometimes allowed to choose more than one race category to describe themselves. But this would involve a major departure from current practice about occupation data, where traditionally a job is classified just once. If jobs can be assigned to multiple categories, the number of such "composite" jobs would be

vast, given that there are 501 occupations that can potentially be combined with each other in contrast to, say, race categories for which this practice would result in far fewer composite categories (see Jones and Bulock 2012).

Finally, the data collection process can be honed to improve the quality of descriptions by more fully engaging interviewers in the coding process. Campanelli et al. (1997, 450) remark that using interviewers for I&O coding may not achieve the same levels of accuracy as specialized office coders, but interviewers who are responsible for coding occupation should have a better sense of what constitutes a good occupational description and probe accordingly for more information. At the very least, the interviewers should be trained on the logic and rationale behind the coding structure so that they have a better sense of the kinds of decisions coders need to make. In addition, decision criteria can be implemented as part of the data collection software, making them available during the interview to support the coding task. For example, when the interviewer types in a term that is known to be problematic, the system would propose particular probes that should resolve the coding problems.

The current study was restricted to interviewer administration of occupation and duties questions and transcription of respondents' descriptions – as is the case in most government surveys that collect such data. But mode may matter. Self-administered questionnaires that are used to collect occupation descriptions (e.g., the American Community Survey is administered both online and by mail, as well as via telephone and personal interviews) require respondents to type or write their answers. Because writing and typing require more effort for most people than speaking, it could be the case that occupation descriptions tend to be shorter in self-administered (visual) modes. If so, the kinds of length-of-description effects we observed here might be reduced when responses are textual. This is an area – especially with the growth of online survey administration – that certainly warrants further study.

Another way in which the current study was restricted was the lack of information about individual coders, in particular their experience and competence. This might affect agreement and moderate the patterns observed here. Unfortunately we did not have access to any information about individual coders, so we could not quantify such effects. Future studies might extend the current findings by including coder information in analyses of coding quality.

Coding open-ended responses is an overlooked source of survey error. More accurate coding, rather than just more reliable coding, should be a priority. If this is achieved, then more reliable coding will follow.

6. References

Andrews, M. 2005. "Who is Being Heard? Response Bias in Open-Ended Responses in a Large Government Employee Survey." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, Minneapolis, MN, 3760–3766.

Anker, R. 1998. *Gender and Jobs: Sex Segregation of Occupations in the World*. Geneva: International Labour Office.

Barsalou, L.W. 1985. "Ideals, Central Tendency and Frequency of Instantiation as Determinants of Graded Structure in Categories." *Journal of Experimental Psychology: Learning, Memory and Cognition* 11: 629–654.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Bjerk, D. 2007. "The Differing Nature of Black-White Wage Inequality Across Occupational Sectors." *The Journal of Human Resources* 42: 398–434. Doi: http://dx.doi.org/10.3368/jhr.XLII.2.398.

Bureau of Labor Statistics. 2014. "Revising the Standard Occupational Classification." Working Paper. Available at: http://www.bls.gov/soc/revising_the_standard_occupational_classification_2018.pdf (accessed September 11, 2015).

Bushnell, D. 1995. "Computer Assisted Occupation Coding." Working Paper. Available at: http://www.blaiseusers.org/1995/papers/bushne95.pdf (accessed September 11, 2015).

Campanelli, P.C., K. Thomson, N. Moon, and T. Staples. 1997. "The Quality of Occupational Coding in the United Kingdom." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, and C. Dippo, 437–457. Hoboken, NJ: Wiley-Interscience.

Cantor, D. and J. Esposito. 1992. "Evaluating Interviewer Style for Collecting Industry and Occupation Information." In Proceedings of the Section on Survey Research Methods, August, Boston, MA, 661–666.

Cawley, J.C. and G.T. Homce. 2003. "Occupational Electrical Injuries in the United States, 1992–1998, and Recommendations for Safety Research." *Journal of Safety Research* 34: 241–248. Doi: http://dx.doi.org/10.1016/S0022-4375(03)00028-8.

Couper, M.P. and F.G. Conrad. 1996. "Collecting Data to Facilitate the Classification of Occupations Using a Skill-Based Approach." Paper presented at Fourth International Social Science Methodology Conference, July, Essex, UK.

Ericsson, A. and H. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*, rev. ed. Cambridge, MA: MIT Press.

Esuli, A. and F. Sebastiani. 2010. "Machines that Learn to Code Open-Ended Survey Data." *International Journal of Market Research*, 52: 775–800. Doi: http://dx.doi.org/10.2501/S147078531020165X.

Grice, H.P. 1975. "Logic and Conversation." In *Syntax and Semantics: Volume 3, Speech Acts*, edited by P. Cole and J.L. Morgan, 41–58. New York: Academic Press.

Hak, T. and T. Bernts. 1996. "Coder Training: Theoretical Training or Practical Socialization?" *Qualitative Sociology* 19: 479–501. Doi: http://dx.doi.org/10.1007/BF02393420.

Hammond, S.K., G. Sorensen, R. Youngstrom, and J.K. Ockene. 1995. "Occupational Exposure to Environmental Tobacco Smoke." *Journal of the American Medical Association* 274: 956–960. Doi: http://dx.doi.org/10.1001/jama.1995.03530120048040.

Heywood, J.S. and P.L. O'Halloran. 2005. "Racial Earnings Differentials and Performance Pay." *The Journal of Human Resources* 40: 435–452. Doi: http://dx.doi.org/10.3368/jhr.XL.2.435.

Israel, G.D. 2010. "Effects of Answer Space Size on Responses to Open-Ended Questions in Mail Surveys." *Journal of Official Statistics* 26: 271–285.

Jones, N.A. and J. Bullock. 2012. "The Two or More Races Population: 2010." 2010 Census Briefs. Available at: https://www.census.gov/prod/cen2010/briefs/c2010br-13.pdf (accessed April 10, 2015).

Kauppinen, T., J. Toikkanen, D. Pedersen, R. Young, W. Ahrens, P. Boffetta, J. Hansen, H. Kromhout, J.M. Blasco, D. Mirabelli, V. Orden-Rivera, B. Pannett, N. Plato, A. Savela, R. Vincent, and M. Kogevinas. 2000. "Occupational Exposure to Carcinogens in the European Union." *Occupational and Environmental Medicine* 57: 10–18. Doi: http://dx.doi.org/10.1136/oem.57.1.10.

Kromhout, H., E. Symanski, and S.M. Rappaport. 1993. "A Comprehensive Evaluation of Within- and Between-Worker Components of Occupational Exposure to Chemical Agents." *The Annals of Occupational Hygiene* 37: 253–270. Doi: http://dx.doi.org/10.1093/annhyg/37.3.253.

Kunda, Z. and K.C. Oleson. 1995. "Maintaining Stereotypes in the Face of Disconfirmation: Constructing Grounds for Subtyping Deviants." *Journal of Personality and Social Psychology* 68: 565–579. Doi: http://dx.doi.org/10.1037/0022-3514.68.4.565.

Layne, L.A. 2004. "Occupational Injury Mortality Surveillance in the United States: An Examination of Census Counts from Two Different Surveillance Systems, 1992–1997." *American Journal of Industrial Medicine* 45: 1–13. Doi: http://dx.doi.org/10.1002/ajim.10308.

Lettau, M.K. 2003. "New Estimates for Wage Rate Inequality Using the Employment Cost Index." *The Journal of Human Resources* 38: 792–805. Doi: http://dx.doi.org/10.3368/jhr.XXXVIII.4.792.

Lyberg, L. and P. Dean. 1992. *Automated Coding of Survey Responses: An International Review*. R&D Reports, No. 2. Stockholm, Sweden: Statistics Sweden.

Lyberg, L. and D. Kasprzyk. 1997. "Some Aspects of Post-Survey Processing." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 353–370. New York: Wiley.

Martin, J., D. Bushnell, P. Campanelli, and R. Thomas. 1995. "A Comparison of Interviewer and Office Coding of Occupations." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August, Orlando, FL, 1122–1127.

Macchia, S. and M. D'Orazio. 2001. "A System to Monitor the Quality of Automated Coding of Textual Answers to Open Questions." *Research in Official Statistics* 4: 7–21.

Moscarini, G. and K. Thomsson. 2007. "Occupational and Job Mobility in the US." *The Scandinavian Journal of Economics* 109: 807–836. Doi: http://dx.doi.org/10.1111/j.1467-9442.2007.00510.x.

Norman, D.A. 1981. "Categorization of Action Slips." *Psychological Review* 88: 1–15. Doi: http://dx.doi.org/10.1037/0033-295X.88.1.1.

Reichard, A.A. and L.L. Jackson. 2010. "Occupational Injuries among Emergency Responders." *American Journal of Industrial Medicine* 53: 1–11. Doi: http://dx.doi.org/10.1002/ajim.20772.

Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. 1976. "Basic Objects in Natural Categories." *Cognitive Psychology* 8: 382–439. Doi: http://dx.doi.org/10.1016/0010-0285(76)90013-X.

Rosenthal, N. 1992. "Evaluating the 1990 Projections of Occupational Employment." *Monthly Labor Review* 115: 32–48.

Shniper, L. 2005. "Occupational Mobility, January 2004." *Monthly Labor Review* 128: 30–35.

Smyth, J.D., D.A. Dillman, L.M. Christian, and M. Mcbride. 2009. "Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly* 73: 325–337. Doi: http://dx.doi.org/10.1093/poq/nfp029.

Sturgis, P. 2004. "The Effect of Coding Error on Time Use Survey Estimates." *Journal of Official Statistics* 20: 467–480.

Sudman, S., N. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass Publishers.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tversky, A. and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185: 1124–1131. Doi: http://dx.doi.org/10.1126/science.185.4157.1124.

Tversky, A. and D. Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90: 293–315. Doi: http://dx.doi.org/10.1037/0033-295X.90.4.293.

U.S. Census Bureau. 2013. Current Population Survey Interviewing Manual. Available at: http://www.census.gov/prod/techdoc/cps/CPS_Manual_June2013.pdf (accessed September 11, 2015).

U.S. Census Bureau. 2014. *Current Population Survey (CPS) and American Community Survey (ACS): Coding Instructions for 2007/2010/2012 Industry and Occupation (I&O) Coding*. Revision 2. October 1, 2014.

Westat/AIR. 1989. *Research on Industry and Occupation Questions in the Current Population Survey. Final Report to the Bureau of Labor Statistics*. Washington, DC: Westat, Inc. and American Institutes for Research.

Received February 2014
Revised September 2015
Accepted October 2015