ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 2, *pp.* 379–407

Occupation coding during the interview

Malte Schierholz,

University of Mannheim, and Institute for Employment Research, Nuremberg, Germany

Miriam Gensicke and Nikolai Tschersich

Kantar Public, Munich, Germany

and Frauke Kreuter

University of Maryland, College Park, USA, University of Mannheim and Institute for Employment Research, Nuremberg, Germany

[Received April 2016. Final revision April 2017]

Summary. Currently, most surveys ask for occupation with open-ended questions. The verbal responses are coded afterwards, which is error prone and expensive. We present an alternative approach that allows occupation coding during the interview. Our new technique uses a supervised learning algorithm to predict candidate job categories. These suggestions are presented to the respondent, who in turn can choose the most appropriate occupation. 72.4% of the respondents selected an occupation when the new instrument was tested in a telephone survey, entailing potential cost savings. To aid further improvements, we identify some factors for how to increase quality and to reduce interview duration.

Keywords: Coding; Interview coding; Measurement error; Occupation; Open-ended questions; Supervised learning

1. Introduction

Occupation is a core organizational principle in our society. Researchers from many disciplines have an interest in measuring occupation, e.g. to capture individuals' tasks and duties for economic studies, to measure the health risk from a person's job or to determine the person's status in society for sociological research, for example in terms of the '*Standard international occupational prestige scale*', the *class scheme of Erikson, Goldthorpe and Portocarero* or the '*International socio-economic index*' (see Hoffmeyer-Zlotnik and Warner (2012), page 191). Many data collections ask for occupation, including the *UK census*, which yielded almost 30 million verbal answers on employment in 2001 (Office for National Statistics, 2003), and the register-based *German 2011 census* with 3.6 million verbal answers (Loos *et al*., 2013). The *American Community Survey* also contains questions on occupation, collecting approximately 2 million responses annually (Thompson *et al*., 2014). Similar questions are common within many other surveys.

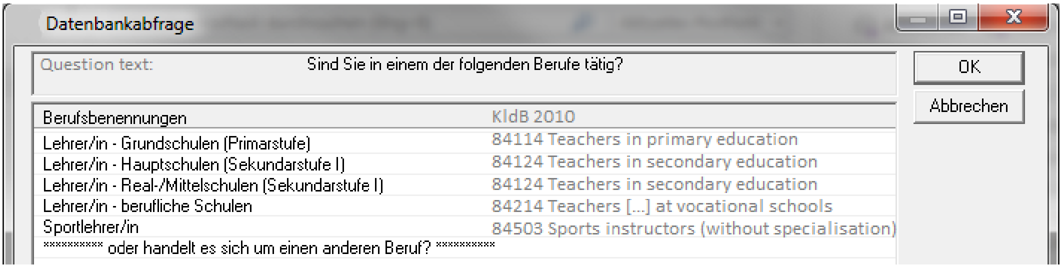
Unfortunately, the measurement of occupation is costly, time consuming and prone to errors. The standard approach is to ask one or two open-ended questions during the interview and sub-

Address for correspondence: Malte Schierholz, Mannheim Centre for European Social Research, University of Mannheim, 68131 Mannheim, Germany.
E-mail: Malte.Schierholz@mzes.uni-mannheim.de

sequently to code the verbal answers in a classification scheme with hundreds of categories and thousands of jobs. This coding task is non-trivial. Conrad *et al*. (2016) discussed various reasons why quality may be compromised. For example, many verbal responses are ambiguous and fit well into more than one category. Furthermore, some respondents have occupations for which no appropriate category exists. Because the target classification is fixed in advance, category modifications that could account for such difficulties are not feasible. Still, coders are typically required to decide on a single, most appropriate, job category. Several studies review the quality of coding occupational information under a variety of conditions (e.g. language, target classification, coding rules and procedures, and coder's experience) and report agreement rates for different people coding the same answers. Campanelli *et al*. (1997) employed three British expert coders to validate original codes from a number of non-experts, obtaining accuracies between 69% and 85%. Elias (1997) listed several British studies with intercoder reliabilities between 70% and 78%, with one exception from Slovenia reaching only 56%, and an international review by Mannetje and Kromhout (2003) mentioned reliabilities between 44% and 89%. Thus, the coding process entails a high degree of uncertainty that is usually ignored during data analysis. Higher quality in occupational data is clearly desirable—even more so if the new technique that we suggest here allows data collection at reduced costs.

Before going into detail, we briefly illustrate the technique proposed: consider a respondent who answers 'vice director [*sic*] and teacher' when asked about his job activities. On the basis of this *verbatim* answer and, if desired, further input from the interview, a computer algorithm searches for possible occupations and calculates associated probabilities at the time of the interview. The job titles that were found to be most likely are then suggested in closed-ended question format to the interviewer, who in turn asks the respondent to select the most appropriate occupation among these suggestions. The suggestions for the above-mentioned example are shown in Fig. 1. Since we cannot guarantee that the algorithm will always suggest an accurate job title, suggestions are complemented by a last answer option 'or do you work in a different occupation?'. If this option is chosen, further questions should be asked to gather additional details about the person's job; if not, coding is complete. In the example, the job title 'Teacher—elementary school' was selected, capturing a detail, the school type, that was not provided in the original verbal response.

With this new approach, we pursue three fundamental objectives to improve current shortcomings in the data collection process. First, we aim to reduce *coding errors* that arise from missing data or contradictory information provided by respondents. Respondents' verbal answers are sometimes ambiguous and difficult to code, in particular when survey questions are not aligned with the theoretical concepts that underlie occupational classification systems.



Question text: Sind Sie in einem der folgenden Berufe tätig?

Berufsbenennungen	KldB 2010
Lehrer/in - Grundschulen (Primarstufe)	84114 Teachers in primary education
Lehrer/in - Hauptschulen (Sekundarstufe I)	84124 Teachers in secondary education
Lehrer/in - Real-/Mittelschulen (Sekundarstufe I)	84124 Teachers in secondary education
Lehrer/in - berufliche Schulen	84214 Teachers [...] at vocational schools
Sportlehrer/in	84503 Sports instructors (without specialisation)
********** oder handelt es sich um einen anderen Beruf? **********	

Fig. 1. Screenshot from the interview with the 'vice director and teacher': job titles in black fount were suggested to the interviewer; the text in grey fount was not shown during the interview and only added for this paper to illustrate underlying categories from the 2010 German classification of occupations; category titles are shown in abbreviated form (this example is discussed in Section 4.4.1)

Answer options often help to clarify the meaning of survey questions. Suggesting a limited number of answer options from the occupational classification based on initial verbal responses thus is expected to improve the measurement, while limiting respondent burden. Second, we seek to maximize the number of interview-coded answers to *minimize efforts for coding the residual cases* after the interview. Third, we aim to *save valuable interview time*, thereby reducing the respondent burden. The closed-ended question that is shown in Fig. 1 can often replace the additional open-ended question about occupation that is used in many questionnaires so that the total interview duration decreases. A final key advantage of the new instrument is the supervised learning algorithm that predicts possible job titles. The predictions are based on training data from past studies and can be *improved as more data become available*.

The new approach was tested in a computer-assisted telephone survey and codes occupations according to the *2010 German classification of occupations* (GCO) (Bundesagentur für Arbeit, 2011a, b), which is a detailed official classification and consists of 1286 well-documented categories subsuming 24000 job titles. Simultaneous coding according to the *2008 'International standard classification of occupations'* (ISCO) (International Labour Office, 2012) is supported in theory; in practice, the algorithm relies on a database that was not prepared for ISCO coding. Adaptions to Web surveys and other computer-assisted modes of data collection are possible, showing that many applications beyond telephone surveys and German occupations exist.

In this paper we describe the test of the new approach. To provide sufficient background and rationale, we review in Section 2 ('Background') literature on occupational coding and survey methodology more generally. In Section 3 ('Data and methods') we describe the data that were used for the test, the technical underpinnings of the new approach and the procedures to evaluate the new method. In Section 4 ('Results and evaluation') we describe the results from our test and discuss extensively the strengths and weaknesses of our approach, giving a special focus on possible modifications of the instrument to achieve even better results. Section 5 serves as a summary and compiles recommendations.

Owing to German privacy regulations, we cannot make our data public. Researchers who are interested in analysing the data on site at the Institute for Employment Research are invited to contact the first author. The computer code of our analysis is available from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. Background

Many processes are related to occupation coding during the interview. First, we characterize briefly the scientific coding tradition in Germany. Second, we argue in detail why our new instrument is expected to improve current practice. Third, we outline the main techniques that are similar to our approach.

2.1. Occupation coding in Germany

Geis and Hoffmeyer-Zlotnik (2000) have provided an overview of occupation coding in Germany and its difficulties. They argued for scientific standards in occupation coding that require coding to be carried out systematically and reliably. Consequently, Geis (2011) published a German coding manual for the international classifications from 1968 and 1988 (both outdated), which contains coding rules, conventions and a case collection to help to achieve high reliability between codings done by different coders. However, reliability does not guarantee validity and coding procedures that are optimized to achieve high reliability carry the danger of introducing systematic biases. For example, one of Geis's rules requires the coder to select the least skilled of the plausible job categories. As a consequence, this coding procedure will underestimate the degree of professionalization in the workforce. Paulus and Matthes (2013) provided shorter instructions for coding into the 2010 German national classification. Dictionaries are available for both classifications to automate the coding process. If the *verbatim* answer from the interview matches an entry in the dictionary, then the corresponding code is assigned. With our technique of coding during the interview we do not follow this German coding tradition with its emphasis on coding rules and reliability; however, we employ professional coders who have a history of coding within the German context and we compare the new approach with current practice.

2.2. Motives for interview coding

We pursue three objectives with our new instrument:

(a) minimizing coding costs,
(b) increasing data quality and
(c) reducing the duration of the interview.

To ensure high data quality, researchers from various countries have discussed and compared various procedures for occupation coding (Biemer and Caspar, 1994; Campanelli *et al.*, 1997; Bushnell, 1998; Biemer and Lyberg, 2003; Maaz *et al.*, 2009; Svensson, 2012; Belloni *et al.*, 2016). Although they focused on errors that arise from coding after the interview, all these researchers also observed that insufficient, low quality verbal answers from the interview are another possible source of errors. According to Hoffmann *et al.* (1995), page 13, ‘the largest source of error lies in shortcomings of the verbatim raw material’, as opposed to errors resulting from coding. Coming from a different perspective, the data editing literature (e.g. Granquist and Kovar (1997) and de Waal *et al.* (2011)) pointed out that it is overly expensive to correct all errors and inconsistencies in a processing step after data collection and advised researchers to improve measurement during the interview instead.

Why is it difficult to elicit suitable information from respondents concerning their occupations? Occupational classifications have organizing principles to cluster the variety of occupations into categories of ‘similar’ occupations. The classification then specifies each category in terms of illustrative occupations, typical tasks and boundaries to related categories. The classification thus dictates which job details are necessary for accurate coding. The current measurement process, however, is not aligned with theoretical concepts from the classification: the United Nations and International Labour Office (2010) recommended asking two open-ended questions, thereby hoping to collect sufficient details for coding according to the 2008 ISCO international classification. It is common practice in many surveys both to ask two questions and to give further instructions, requesting full details about the ‘occupation’ or ‘job title’ as well as the tasks and activities in the job (Tijdens, 2014a). We are sceptical about these standards for two reasons.

(a) No efforts are made in questionnaires to provide information about the classification structure and the boundaries between categories to make respondents’ verbal answers more relatable to specific categories. Respondents can only guess which details about their jobs are requested and ambiguous answers are thus inevitable.
(b) The exact wording of an answer matters to coders, but it is unlikely that every respondent would use exactly the same words when answering the same question in a repetition study. In fact, the belief sampling model (Tourangeau *et al.*, 2000), which is usually applied to attitude questions, suggests otherwise. According to this model, respondents base their answer on a small number of considerations drawn from a larger body of accessible

knowledge about his or her job. The considerations are retrieved at random depending on his or her current state of mind, with randomness leading to variability in answers. If the questions were more precise, the respondent could understand better what kind of knowledge is requested and randomness should decline.

Taking both points together, ambiguous and variable answers are thus a direct consequence of overly general questions that do not provide sufficient clues concerning what kind of information is required.

If respondents provide only vague occupational titles or if their task descriptions are not sufficiently detailed, many surveys demand from interviewers that they ask for a full job title or description (Hoffmann *et al.*, 1995; Tijdens, 2014a). The optimal kind and extent of probing are controversial. On the one hand, probes to open-ended questions in general can increase respondents' understanding of a question or encourage them to clarify their answers and to provide more complete information (e.g. Billiet and Loosveldt (1988), Conrad and Schober (2005) and Holland and Christian (2009)). On the other hand, when interviewers have some freedom concerning when and how to carry out probing, there is concern that respondents' answers are prone to interviewer effects, biasing all responses that are elicited by the same interviewer in a certain direction. To reduce this error, the standardized interviewing literature recommends avoiding probing whenever possible by using improved question wordings (e.g. Fowler and Mangione (1990), Mangione *et al.* (1992) and Schaeffer *et al.* (2010)).

Aside from this general discussion, the specific evidence for occupation suggests that probes are sometimes counterproductive. Since chances for conflicting information that is more difficult to code are higher when more information is available, coders disagree more often about the correct code when answers are longer (Conrad *et al.*, 2016). Results from Cantor and Esposito (1992), drawn from coders' comments on the interviewers' questioning strategy, point in a similar direction. Their coders criticized that interviewers probe too much in some cases, which results in adding ambiguity to the answer rather than in resolving conflicting information, whereas in other cases no probes are asked at all although coders actually desire more specific information about a certain aspect of the occupation. The bottom line is that generating useful probes is extremely difficult for interviewers. They would need to be well trained in probing and experienced with the occupational classification to elicit more useful answers, or—and this is the route that we follow in this paper—the questionnaire should provide interviewers with additional, classification-related questions that prescribe the exact wording for standardized probes.

Any effort to obtain more information about a respondent's job increases the total duration of the interview. Longer interviews have been associated with an increased respondent burden, lower rates of survey participation, more satisficing behaviour and reduced quality of data at the end of long questionnaires (e.g. Bradburn (1978), Holbrook *et al.* (2003), Galesic and Bosnjak (2009) and Roberts *et al.* (2010)). Asking two open-ended questions and possible additional probes cost a considerable amount of time. Although these efforts are necessary to obtain precise information from some respondents, valuable interview time is wasted for others who have given a precise answer already to the first question. To reduce the length of the interview, our proposed instrument is adaptive: we suggest asking only one open-ended question, whereupon the interview software evaluates the answer and decides which question is asked next.

2.3. Related techniques and instruments

Our new instrument combines two specific developments that have been implemented separately from each other.

(a) Post-interview coding of occupations is increasingly carried out by using machine learning algorithms and training data rather than simple matches of answers to prespecified words from a dictionary.
(b) Some researchers have described mechanisms for coding during the interview that enable the respondent to specify his or her response more precisely, if required.

In what follows we provide a brief overview of both approaches.

Several researchers have proposed computer systems to automate the post-survey coding process (e.g. Speizer and Buckley (1998) for a review, Measure (2014) and Gweon *et al.* (2017)). Software for computer-assisted coding suggests possible categories to the coder to make human work more efficient. Other algorithms known as 'automated coding' independently assign categories without human supervision. More difficult residual cases are left for professional manual coding to keep the error level from automated coding below some prespecified threshold. Conceptually, both systems are generally based on coding rules and large databases that contain codes for recurrent job titles (e.g. the prominent 'Computer-assisted structural coding tool' program CASCOT that was described by Elias *et al.* (2014) implements all the above mentioned). However, coding rules are created by hand, and complex coding systems require quality checks before they can be used for production. Creecy *et al.* (1992) have challenged such hand-crafted rule systems, which are expensive to develop, with their own algorithm that learns from training data, consisting of 132 247 observations. By doing so, previously coded verbal answers are used to learn coding rules automatically. Their software outperforms another coding system that was based on hand-crafted rules and used for production in the 1990 US census. More recent proposals learn from even larger training data with more than 1.5 million observations each (Jung *et al.*, 2008; Thompson *et al.*, 2014; Javed *et al.*, 2015). The novel algorithm that we use in this study combines the learning from training data and the usage of hand-crafted databases.

Different strands of research try to code occupations directly during the interview. Hoffmeyer-Zlotnik *et al.* (2006) asked for occupation with a sequence of three filter questions: the first asks for broad occupational groups, the second specifies the occupation further and the final third question refers to specific 1988 ISCO categories. Tijdens (2014b, 2015) also avoided verbal answers with a similar 'search tree' for Web surveys. A different strategy is employed by the job portal offered on line at `http://jobboerse.arbeitsagentur.de/` by the German Federal Employment Agency, which uses textual input to autosuggest possible job titles from a database. These titles are linked to job categories from the 2010 GCO. Hacking *et al.* (2006) and Svensson (2012) also mentioned occupation coding during the interview, but their descriptions lack details.

Coding during the interview is not limited to occupation coding but is applicable to any question with a large number of answer options. Bobbitt and Carroll (1993), for example, tested a system for coding 'major field of study'. In a telephone survey, they implemented a fuzzy text search algorithm that suggests possible codes, allowing interviewers to verify codes directly with the respondents. Couper and Zhang (2016) asked for prescription drugs in Web surveys and compared three question formats: a text box for later coding, a drop box menu containing a list of 4768 drug names in alphabetical order and an autosuggested list that narrows down the large number of drugs on the basis of textual input and simple text matching. They concluded that each format has its own strengths; nevertheless, a careful design of the instrument is worthwhile.

3. Data and methods

The new tool was tested in the survey '*Selectivity effects in address handling*' that was commissioned by the German Institute for Employment Research and conducted by Kantar Public. In

October and November 2014, Kantar Public conducted in total 1208 valid computer-assisted telephone interviews; 1064 verbal answers for occupation were collected. The questionnaire covered, among others, several topics related to the respondents' current occupation and work history, the use of social media for private and professional purposes, and volunteering activities.

3.1. Sampling and data collection

A random sample of 17001 people—some of them with multiple addresses; others without phone numbers, which needed to be identified—was drawn from a German federal database that is used in social security administration (vom Berge *et al.*, 2013). Since the primary purpose of the survey was to explore possible selectivity effects, a random subsample of 10000 people was asked for consent to address transfer and the consenters' addresses as well as control group addresses were transferred from the German Institute for Employment Research to the survey operator. Details of the experiment are described in Sakshaug *et al.* (2016). Before the fieldwork, a notification letter was sent to 7183 available addresses.

The sampling frame covers employees, unemployed people, jobseekers, recipients of unemployment benefit II and participants in active labour market programmes. It thus accounts for a large share of the German working population. However, people who never paid contributions for social security insurance and have never received benefits from the German Federal Employment Agency are not included. This implies a specifically strong undercoverage of civil servants and self-employed people.

All 67 interviewers, the local fieldwork managers and the supervisors were trained by the central project management team of Kantar Public. The new tool was an essential part of this training.

3.2. Integration into the questionnaire

The coding process starts by asking one open-ended question about the occupation ('Please tell me your occupational activity'), followed by a sequence of approximately eight additional job-related questions that are intended to collect as many details as possible about the respondent's job. These further questions are used

(a) for manual coding to evaluate the new instrument and
(b) as covariates in our model to improve predictions.

The answer to the first open-ended question would suffice to suggest possible job titles and that the additional questions could be skipped if they were not needed to evaluate the instrument. The exact German wording and its corresponding English translation are available in the on-line appendix.

Immediately following these questions, interviewers are prompted to read the following text:

> 'We now try to classify your occupation. A database query is made for this purpose. This can take a short moment.'

The interviewer then starts the query and the algorithm computes at most five job titles to be suggested to the respondent. After a few seconds, the question generated is shown in a pop-up window (Fig. 1). The interviewer then asks into which of the following categories (job titles) the job falls, or whether the answer option 'different occupation' would be most appropriate. A random subset of less than 10% received an additional answer option 'similar occupation'. Because of the small sample size we discuss this portion of the study only in the on-line appendix. All interviewers received a quick debriefing question on the flow of the interaction after the coding module. Details of those debriefings are also available in the on-line appendix, for brevity.

Every suggested job title (shown on the left-hand side in Fig. 1) corresponds to one category from the *Dokumentationskennziffer*, which is an internal job classification that is used by the German Federal Employment Agency in its daily operations (see Paulus and Matthes (2013) for details). This classification subdivides the 1286 categories from the 2010 GCO in 11 194 *Dokumentationskennziffer* categories. Conversely, this means that every job title is linked to exactly one category in the 2010 GCO. Thus, when a job title is selected during the interview, the *Dokumentationskennziffer* code is saved and a 2010 GCO code is automatically assigned as well. For illustration, we include these associated GCO categories in the grey fount on the right-hand side of Fig. 1. All evaluations provided below will be done on the scale of the 2010 GCO, as this is the official and well-documented German national classification. The *Dokumentationskennziffer* itself is used only as an auxiliary classification that provides the job titles for our instrument, links these job titles to the 2010 GCO and makes available a large database of search words.

Many researchers do not use the national 2010 GCO but work with the 2008 ISCO instead. As this study explores technical possibilities, we test our technology only on the 2010 GCO. However, it is worth noting that many—but not all—*Dokumentationskennziffer* categories are linked to specific ISCO categories, making it conceptually feasible to code in the 2008 ISCO and 2010 GCO at the same time during the interview. Since the ISCO with its 436 categories corresponds to only about a third the size of the GCO, we also expect improved quality evaluations if the analysis below is carried out for the 2008 ISCO.

3.3. Prediction algorithm

Possible job categories are predicted with a supervised learning algorithm that learns from training data, i.e. from verbal answers whose classification codes are already known from manual coding. Our training data come from the survey '*Working and learning in a changing world*' (Antoni *et al.* (2010); Drasch *et al.* (2012) documented the coding process). This survey interviewed 9 227 people about their employment biographies, i.e. all the jobs that they have held during their lifetime, yielding a total of 32 887 job records. Compared with other supervised learning algorithms for occupation coding, this number is exceptionally small. Because of the tiny size of our training data, 433 out of 1 286 job categories from the 2010 GCO are not covered, implying that these categories would never be suggested if the predictions were based only on these training data.

In principle, training data should be as large as possible to account for a high variety of possible verbal inputs, including misspellings, and it should also cover all contingencies how a specific input text can be coded in different categories. Such large training data were not available to us; as a consequence, many respondents provide verbal answers that cannot be matched to the training data. To obtain predictions for these respondents still, we use two databases of job titles in addition to our training data and search for possible job categories in all three sources. Resorting to additional databases should mitigate our problem of small training data, but more training observations could certainly further improve our results.

Schierholz (2014) developed the underlying prediction algorithm and evaluated its performance. To integrate Schierholz's (2014) algorithm into our new coding approach, the target classification was changed (*Dokumentationskennziffer* instead of the 2010 GCO) and some scores (see below) were streamlined. The algorithm works in three steps; the exact calculations in each step are described below. The rest of this paper can easily be understood without these technical details.

(a) Calculate scores $\theta_{lj}^{(m)} = f_m(x_l, c_j)$ for a given respondent l and all *Dokumentationskennziffer* job categories c_j, $j = 1, \ldots, 11194$. We use 26 predefined matching methods f_m, $m = 1, \ldots, 26$, that link the respondent's answers x_l to databases and training data.
(b) Predict correctness probabilities for all categories using a function $\hat{g} : (\theta_{lj}^{(1)}, \ldots, \theta_{lj}^{(26)}) \mapsto \hat{p}(c_j|l)$. This function was estimated beforehand from training data.
(c) Suggest the five most probable job categories (under some restrictions) to the respondent.

3.3.1. *Calculate scores*

To be useful, the scores $\theta_{lj}^{(m)}$ should be predictive of the true probability $p(c_j|l)$ that category c_j is correct for respondent l. Any supervised learning technique might be used to estimate functions f_m, the more the better, as long as the number of scores is far below the number of observations that are used. For simplicity, we use only a small set of 26 matching methods to define the functions f_m. Several scores are built on each other and, because predictions improve when more scores are used, we included all of them. The matching methods are summarized in Table 1. For example, our first matching method, f_1, selects all training data observations in which the full texts from respondent l and from the training data are identical and calculates the frequencies of each category c_j in this subset. By construction, the most frequent code that is found with this matching method is likely to have the highest probability of being correct.

To develop additional matching methods, we vary four dimensions as shown in Table 1.

(a) The input is either the respondent's answer to one of the closed questions, the *full text* (i.e. the first verbal answer after removing some special characters and replacing letters with their upper-case equivalents), or a *phrase* (i.e. the subsequence of words from the full text that has the highest frequency of appearance in a single category in the training data). In our example, the full text is 'VICE DIRECTOR AND TEACHER' and the derived phrase is 'TEACHER'.
(b) For comparison, our input texts must be either *identical* to or a *substring of* another text. The naive Bayes statistic is based on a *word-by-word* comparison.
(c) The input is compared with *training data*, to an *alphabetic dictionary* of job titles (the '*Berufs- und Tätigkeitsverzeichnis*' that is part of the 2010 GCO; Bundesagentur für Arbeit (2011a)), or to an index of *search words* (created by the German Federal Employment Agency for operative purposes; Bundesagentur für Arbeit (2013)).
(d) The statistic dimension prescribes how to calculate category-specific scores θ_{lj} from large numbers of matching entries.

The most basic statistic is the *code frequency*, i.e. the absolute frequency $\#\{\text{answer}, c_j\}$ of each code c_j that appears in the selected subset.

'*Posterior expectation*' is the posterior expectation for some category c_j and '*posterior probability*' is the posterior probability that $\text{parameter}_j > 0.05$. The underlying Bayesian model consists of a subset-specific multinomial likelihood to model the observed code frequencies $\#\{\text{answer}, c_j\}$ and a Dirichlet prior that depends on relative code frequencies in the complete training data, $\text{Dirichlet}(0.5 \cdot \#\{c_1\}/N, \ldots, 0.5 \cdot \#\{c_J\}/N)$. The posterior is thus a $\text{Dirichlet}(\#\{\text{answer}, c_1\} + 0.5 \cdot \#\{c_1\}/N, \ldots, \#\{\text{answer}, c_J\} + 0.5 \cdot \#\{c_J\}/N)$ having posterior expectation $\theta_{lj}^{(3)} = \omega \cdot \#\{\text{answer}, c_j\}/\#\{\text{answer}\} + (1-\omega) \cdot \#\{c_j\}/N$, a weighted average with weights $\omega = \#\{\text{answer}\}/(\#\{\text{answer}\} + 0.5)$ that shrinks the relative code frequencies in the selected subset towards the prior expectations.

For closed questions, we calculate the *proportions* $\hat{p}(\text{input}|c_j) = \#\{\text{answer}, c_j\}/\#\{c_j\}$, which is the subset-specific code frequency divided by the absolute frequency of each code in the complete training data. On the basis of the *naive Bayes* assumption, we esti-

Table 1. Overview of matching methods

m	*Input*	*Comparison*	*Compared with*	*Statistic*	*Mean*†
Open-ended questions					
1	Full text	Identical to	Training data	Code frequency	0.00261
2	Full text	Substring of	Training data	Code frequency	0.00439
3	Full text	Identical to	Training data	Posterior expectation	0.00009
4	Full text	Identical to	Training data	Posterior probability	0.00019
5	Full text	Word by word	Training data	Naive Bayes	0.00008
6	Full text	Identical to	Search words	Code frequency	0.00019
7	Full text	Substring of	Search words	Code frequency	0.00400
8	Full text	Substring of	Alphabetic dictionary	Code frequency	0.01264
9	Phrase	Identical to	Training data	Code frequency	0.00261
10	Phrase	Identical to	Training data	Posterior expectation	0.00009
11	Phrase	Identical to	Training data	Posterior probability	0.00019
12	Phrase	Word by word	Training data	Naive Bayes	0.00008
13	Phrase	Identical to	Search words	Code frequency	0.00049
14	Phrase	Substring of	Search words	Code frequency	0.14391
15	Phrase	Substring of	Alphabetic dictionary	Code frequency	0.02347
Closed questions					
16	Occupational status	Identical to	Training data	Proportion	0.11747
17	Differentiated occupational status	Identical to	Training data	Proportion	0.05755
18	Number of staffers	Identical to	Training data	Proportion	0.20734
19	Superviser	Identical to	Training data	Proportion	0.10926
20	Number of employees supervised	Identical to	Training data	Proportion	0.12083
21	Education required	Identical to	Training data	Proportion	0.04996
22	Industry	Identical to	Training data	Proportion	0.15182
23	Company size	Identical to	Training data	Proportion	0.05623
Other					
24	Multiply scores 5 and 16–23, and relative code frequency in training data				0.03127
25	Are full text and phrase identical?			Yes or no	0.41820
26	Number of suggested categories from all matching methods			Count	184.0031

†Mean $=(1/N)(1/J)\Sigma_{l=1}^{N}\Sigma_{j=1}^{J}\theta_{lj}^{(m)}$ is the mean score over all categories and respondents for the matching method m. It is based on a subset of 958 respondents where the algorithm finds possible categories.

mate probabilities to observe the observed text under any given job category by using the formula

$$\hat{p}(\text{input text}_l|c_j) \propto \prod_{v=1}^{V} \{0.95\hat{p}(T_v|c_j) + 0.05\hat{p}(T_v)\}.$$

This is a product over all words T_v that appear in the input text. $\hat{p}(T_v|c_j)$ and $\hat{p}(T_v)$ are both relative frequencies as calculated from the training data. $\hat{p}(\text{input text}_l|c_j)$ is standardized to sum to 1. The proportion statistic and the naive Bayes statistic were originally developed as a by-product to estimate $p(c_j|l)$ as in $\theta_{lj}^{(24)} = \hat{p}(c_j|l) := \hat{p}(c_j)\,\hat{p}(l|c_j)/\hat{p}(l)$ with

$$\hat{p}(l|c_j) := \hat{p}(\text{input text}_l|c_j) \prod_{m=16}^{23} \theta_{lj}^{(m)}.$$

The final score $\theta_{lj}^{(26)}$ is a person-specific (equal to category-independent) variable that counts how many different categories were found that have code frequency greater than 0 in one of the matching methods. If this number is high, many different categories appear possible, and the final probability of picking the correct category might be lower.

Schierholz (2014) explained the scores and the underlying reasoning in more detail.

3.3.2. *Predict correctness probabilities*

We have now 26 different scores that are expected to correlate with the true probability $p(c_j|l)$ and that can be interpreted, in the case of $\theta^{(3)}_{lj}$, $\theta^{(10)}_{lj}$ and $\theta^{(24)}_{lj}$, as an estimate for this probability. We write all scores from a single person in a data frame as depicted in Table 2 and concatenate the data frames from all training observations to form a single data frame. How can we *combine* the different scores to form a single more accurate prediction $\hat{p}(c_j|l)$? We need to estimate a function $\hat{g}: (\theta^{(1)}_{lj}, \ldots, \theta^{(26)}_{lj}) \mapsto \hat{p}(c_j|l)$. Although the outcome is multinomial, we regard it as a binary problem and aim to predict the probabilities $p\,(c_j\,\text{correct}|l)$ instead. A similar problem of combining predictions ('stacking') was studied by Stone (1974), LeBlanc and Tibshirani (1996) and Breiman (1996) who restricted themselves to linear combinations g of the predictors $\theta^{(m)}_{lj}$. Stone (1974), LeBlanc and Tibshirani (1996) and Breiman (1996) assumed that predictors $\theta^{(m)}_{lj}$ are in themselves estimates of the outcome variable that may be obtained from any supervised learning model; however, we see no reason why their method should be limited to linear combinations of other models' predictions. The key problem is that estimated parameters would be biased if the same training observations were used twice: a first time to estimate the functions f_m and a second time to estimate g. To avoid double usage, we apply leave-one-out cross-validation, i.e. the first-stage predictions $\theta^{(m)(-l)}_{lj} = f_m$ are not based on the observed outcome of respondent l. The observed outcome from the training data is used only afterwards for estimation of g, together with the leave-one-out estimates $\theta^{(m)(-l)}_{lj}$. To estimate the function g, we train gradient-boosted trees as implemented by Hothorn *et al.* (2010), which is a more flexible tool than linear regression that allows for non-linearities and high order interactions. In doing so, a sequence of decision trees is trained iteratively, each iteration focusing on examples that the previous iteration got wrong. The final prediction is a sum over the different trees.

Training the gradient boosting model on a data set with $32\,887 \times 11\,194 \approx 368$ million rows is computer intensive and time consuming. Furthermore, computers need to have a very large random-access memory to load a boosted model if the training data consist of many observations. This is a shortcoming of our approach and three workarounds are needed to make this manageable.

(a) We keep only the rows in the data frame in which at least one score obtained via the verbal answer indicates that this category could be correct. If the text does not indicate

Table 2. Illustrative data frame for person l with correct job category $c_j = 01104101$

c_j	c_j *correct*	$Score^{(1)}_{lj}$	$Score^{(2)}_{lj}$	$\cdots$
$c_1 = 01104100$	False	$\theta^{(1)}_{l,1}$	$\theta^{(2)}_{l,1}$	$\cdots$
$c_2 = 01104101$	True	$\theta^{(1)}_{l,2}$	$\theta^{(2)}_{l,2}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c_{11194} = 99998115$	False	$\theta^{(1)}_{l,11194}$	$\theta^{(2)}_{l,11194}$	$\cdots$

the possibility of correctness (as operationalized in $\theta_{lj}^{(26)}$), the row is removed. The resulting data set has 461 816 rows, many of them still highly unlikely to be correct.

(b) We randomly split the data by rows into 10 disjoint sets and estimate five separate boosting models, leaving the other five sets aside because of performance restrictions. To predict a new code for a given response l, we can then
  (i) predict the scores $\theta_{lj}^{(m)}$ by using the complete training data,
  (ii) predict probability vectors for 'c_j correct' with each of the five boosting models and
  (iii) average over the five predictions.

(c) To speed up model training, we use only 45 iterations, which are fewer than recommended. To reach close-to-optimal solutions after 45 iterations, we increase the step size. Tuning parameters (maximum tree size, 9; step size, 0.5) are chosen according to extensive exploratory bootstrap-type cross-validation.

Although we are confident that these design decisions do not negatively affect the algorithm much, more efficient solutions are clearly desirable.

3.3.3. *Suggest job categories*

At this point, the algorithm enables us to calculate a number of possible *Dokumentationskennziffer* categories and corresponding estimated correctness probabilities for new data. For most responses, dozens of categories are found—more than would be convenient to ask in a survey. We therefore restrict the maximum number of suggested categories to five, which is a suitable number for unordered response options in telephone surveys (Schnell (2012), page 94). It is desired to suggest job titles that cover a range of GCO categories. For this we select (up to) five *Dokumentationskennziffer* categories with the highest correctness probabilities under the condition that not more than two of the selected *Dokumentationskennziffer* categories may belong to the same GCO category. Only if we cannot fill the five available spaces according to this rule are additional *Dokumentationskennziffer* categories from the same GCO category added according to their correctness probability (highest first). Finally, the categories suggested are ordered by GCO code numbers and the answer option for 'other occupation' is added.

3.4. *Quality analysis*

We evaluate the quality of our new approach by comparing the machine-learning-assisted within-interview coded answers with professional post-survey coding, as is traditionally done in Germany. The analysis includes two steps:

(a) manual coding and
(b) double-checking of answers for which at least one code might be wrong.

For the first step, two professional coders were asked to code the verbal answers independently from each other and without knowledge about the interview-assigned codes. Both are experienced coders and offer this service on a paid basis. Their respective coding documentations show that both coders have different coding procedures and, for ambiguous answers, different decision rules are used. In addition, one of the coders provides a special indicator describing which verbal answers have multiple possible codes. To guarantee the anonymity of the coders, the differences cannot be described in more detail.

In a second step, all observations for which there was disagreement between any two of the three codes (from interview coding and both professional codings) was subject to additional examination. This also includes observations for which one of the professional coders expressed his uncertainty. Observations that were not interview coded are excluded. Two student assistants checked the correctness of the different codes for each of the 368 observations. Both assistants worked independently of each other. They were provided with the same source material as the two coders (verbal answers and additional answers from the interview; see the on-line appendix) and with the codes from the professional coding and interview coding. Their task was to categorize each coding decision in one of the following three categories.

(a) Acceptable: there is a good argument for the coding decision to be considered correct. This is independent of the fact that other plausible arguments may lead to different coding decisions that may be considered correct as well.
(b) Wrong: it is obvious that the coding decision is erroneous and other codes are clearly more appropriate.
(c) Uncertain: this is the residual category to be assigned when a code is not obviously erroneous and at the same time there is no good argument for it to be correct. Three reasons are most common why a category is classified as uncertain:
    (i) the job title that is selected during the interview appears correct at a first glance, but a different category definition from the GCO, volume 2, describes the job activities more precisely;
    (ii) the interview-coded job category requires a level of skill that is contradictory to the answers from the interview (i.e. to the questions on the vocational training that is usually required or the differentiated occupational status);
    (iii) the answers from the interview suggest a different thematic focus, but at the same time the code is not entirely wrong.

The complete instructions including examples, which were given to the student assistants, are provided in the on-line appendix.

3.5. *Interviewer behaviour*

To evaluate the new approach further we coded the interviewer behaviour itself. This allows us to analyse the extent how often interviewers correctly applied standardized interviewing techniques that are prescribed for the new question on occupation (see Ongena and Dijkstra (2016) for an overview on behaviour coding). At the beginning of the interview, all respondents were asked for permission to record the conversation, obtaining an 87.5% rate of consent. Of the consenters who provided answers to the occupation questions and whose audio recordings did not contain personal identifiers, a total of 211 were randomly selected for behaviour coding. An independent coder from Kantar Public recorded whether the interviewer read the question text and each answer option as instructed, or if he or she diverged from the text or omitted suggested job categories. In addition, the coder noted what the respondent said as a first reaction. The complete coding instructions are provided in the on-line appendix. Two audio files were excluded, because the recordings only start after the occupational questions have been asked. In the course of the analysis, the first author listened to several recordings and felt reassured that the coder delivered high quality. Various interpretations of the interviewer–respondent interaction in the result section were also obtained from listening to the recordings with careful attention to the specified aspects.

4. Results and evaluation

This section starts with three key criteria to assess the tested system: productivity, interview duration and quality, followed by two examples to explain some particularities of our instrument. We then report from the detailed analysis of the audio recordings to understand how interviewers and respondents interact, and discuss the strengths and weaknesses of the prediction algorithm. We close this section with an examination of errors resulting from the classification material. Throughout all descriptions, we highlight shortcomings in the tested system and mention possible modifications to obtain even better results in a future version of the instrument.

4.1. Productivity analysis

Table 3 provides an overview of the productivity of our system. Among the 1064 people who responded to the survey questions about occupation, the algorithm found possible categories for 90.0%, leaving only 10.0% for whom the algorithm did not suggest a single job category. This happens if the algorithm cannot relate the text that is entered by the interviewer to any previous input from the training data or from the job title databases. This is often due to misspelled job titles and could be reduced by using spell checking algorithms.

72.4% of the respondents selected a job title from the list generated. This number is highly important, because it shows that nearly three-quarters of the coding task could be carried out during the interview, which considerably reduces the work for post-interview coding.

13.6% of the respondents did not find an appropriate job title among those suggested by the algorithm and declared that they have a different occupation instead. This was expected, as the algorithm is optimized to suggest appropriate job titles, but it is impossible to guarantee that it will always propose correct job categories. In fact, the matching methods in our algorithm often find dozens or even hundreds of possible job titles. For usability, we restrict the maximal number of suggested job titles to five. When filtering out the five best-suited job titles, frequently relevant categories are missed, whereas irrelevant categories are suggested. The quality of the suggestions depends on the availability of training data and details in the algorithm. With additional training data and improved algorithms for prediction, we thus expect to decrease the proportion of answers for which no code is assigned and to increase the productivity of our system.

For respondents who answer that no job title is appropriate and who indicate that they have an ‘other occupation’ (applicable for 13.6%, as shown above), two additional lists are generated automatically and suggested to them. The first contains titles from the more general occupational subgroups (four-digit GCO). The respondent can then select a subgroup or terminate the procedure by saying that no subgroup is appropriate. When selecting a subgroup, *Dokumentationskennziffer* job titles only from the chosen subgroup are suggested to the respondent. This demanding follow-up process was implemented because the algorithm usually finds dozens of possible job titles and, although it is desired that respondents can navigate to the best fitting job title during the interview, it is impossible to suggest all of them within a single question. Contrary to our expectations, 79% of the eligible respondents did not select an occupation during

Table 3. Productivity of the coding system

Number of respondents who give a job description	1064		100.0%	
Algorithm provides no job suggestion	106		10.0%	
Algorithm finds possible categories: thereof,	958		90.0%	
. . . Respondent chooses a job title		770		72.4%
. . . Respondent chooses ‘other occupation’		145		13.6%
. . . Item non-response		3		0.3%
. . . Other experimental conditions		40		3.8%

this process. In case they did, this interview-coded occupation is not in agreement with manual coding in 77% of cases. Fig. A1 and Table A1 in the on-line appendix provide additional details. We conclude that these follow-up questions yield unsatisfactory results and should be dropped. If respondents select 'other occupation', responses should be referred to manual coding. They are thus excluded from the subsequent analysis.

Table 3 also shows that three of the 1064 people did not respond to the new instrument. The remaining 3.8% were due to the following experimental artefact: if the algorithm finds only a single job title or more than 250 possible job titles, job titles were not suggested within the regular closed question on occupation, but different question wordings were tested instead. Results are shown in Tables A2 and A3 in the on-line appendix. Both experimental conditions were not worthwhile for our research because the number of observations falls below our expectations. Standard procedures, as if 2–250 categories were suggested, would probably have worked equally well.

4.2. Interview duration

If coding during the interview is to replace the present procedure that asks two or three open-ended questions about a respondent's job, it is of high relevance that the duration of the interview does not increase. Longer interviews are more expensive and tiresome for the respondent. For respondents who select an occupation during the interview, our additional question takes 37 s on average. As further open-ended questions can be avoided (a standard question in German surveys is 'Please describe this occupational activity precisely', which takes 44 s on average), the total interview duration is reduced for these respondents. Conversely, for respondents who do not select an occupation but instead choose the category 'other occupation', additional open-ended answers are still necessary for coding after the interview, increasing the total duration of the interview. The objective must therefore be to minimize the number of respondents who choose 'other occupation'.

4.3. Quality analysis

Nearly three-quarters of the respondents select a job title during the interview. Although this is auspicious, the quality of the interview-coded categories is even more relevant. Two specific aspects of quality are analysed: the agreement between and the evaluation of the different coding procedures. Both measures enable conclusions about the quality.

Table 4 (the first three rows) provides the intercoder reliabilities for the professional coders (coder 1 and coder 2) and their respective rates of agreement when compared with the codes from the interview. Agreement between five-digit categories from the 2010 GCO is highest with 66.23% when comparing coder 2 with interview coding. All agreement rates improve for broader classifications with fewer digits, but coder 2 and interview coding again have the highest rates of agreement. Agreement between both professional coders is lowest with almost 39% of disagreement, leaving room for improvement.

An explanation for the lower agreement between professional coders might be that it is easy to find a correct code for some job descriptions, whereas it is not for others (e.g. Cantor and Esposito (1992) and Conrad *et al.* (2016)). For example, if respondents find their previous verbal answer in the list of suggested job titles, they often select this job title, acting similarly to what professional coders do. We assume that people with more complex job descriptions are more hesitant to choose one of the suggested job titles, as the titles are less likely to be appropriate. Consequently, simpler job descriptions are more often interview coded. In contrast, professional coders are required to code all occupations, regardless of the selection process during the interview, including also the more complex job descriptions, on which professional

Table 4. Agreement rates between the two professional coders (coder 1 and coder 2) and interview coding (interview)†

Agreement between	*Number of codes*‡	*First . . . digits are in agreement (%)*				
		1	*2*	*3*	*4*	*5*
	(All available)					
Coder 1 and coder 2	1039	87.20	79.40	74.98	67.56	61.11
Coder 1 and interview	754	87.67	80.37	75.46	67.77	61.80
Coder 2 and interview	770	89.09	82.21	77.53	71.56	66.23
	(Subset)					
Coder 1 and coder 2	754	89.26	82.76	79.18	72.68	65.78
Coder 1 and interview	754	87.67	80.37	75.46	67.77	61.80
Coder 2 and interview	754	88.86	81.83	77.19	71.35	66.05

†The 2010 GCO consists of five-digit codes; aggregates for broader classifications with fewer digits are shown for convenience.
‡'Number of codes' shows how many codes are available for each comparison. Coder 1 provides codes for 1041 out of 1064 occupations. For three occupations, the 'qualification is unknown', one occupation is a worker without further specification, and for 19 occupations 'multiple codes [are considered] possible'. Coder 2 provides codes for 1062 out of 1064 occupations, whereas the other two occupations are 'not codable'. Interview coding provides codes for 770 occupations. The quotes stem from the respective coding documentation.

coders presumably agree less often. This argument is supported by the fact that agreement between coder 1 and coder 2 increases from 61.11% to 65.78% when this number is calculated only for the subset of the 754 occupations that were also coded during the interview.

To allow for more meaningful comparisons of the different coding procedures, the bottom part of Table 4 is based on a subset of respondents for whom we have codes that are available from all three procedures. Looking at agreement rates of 1–4 digits, agreement is highest between coder 1 and coder 2. For five-digit codes, this pattern changes with coder 2 and interview achieving the highest rate of agreement. Hypothesizing that interview-coded answers are more often miscoded, they would agree less often with accurate codes from professional coding. This could explain that agreement between coder 1 and coder 2 is highest for 1–4 digits, suggesting lower quality of data of interview coding. However, the differences between the coding procedures are small and even non-existent for five-digit codes. Furthermore, higher miscoding rates of interview coding are only one possible explanation for the observed pattern. Another explanation assumes that professional coders lack relevant information because it was not provided during the interview. In this case, they both would assign wrong codes, leading to agreement between professional coders but to disagreement with the respondent's own choice, who knows better. Taken together, the evidence provided so far is inconclusive about the validity of each coding procedure and suggests that differences between the coding procedures are only small, if existent at all. The second step of our analysis will elucidate this further.

For 402 out of 754 respondents (53.32%), the professional coders both agree with the respondents' own choices. For these cases we can be highly certain that interview coding yields a code with quality that is comparable with manual coding. In what follows, we assume that these codes are 'acceptable'. More problematic are the $770 - 402 = 368$ cases in which at least one human coder deviates from the code that was obtained via interview coding.

Table 5 shows the results from the students' evaluation of the quality of coding. For the

Table 5. Student assistants' evaluation of the coding process quality†

		Student 2			
		Acceptable	*Uncertain*	*Wrong*	Σ
Correctness for coder 1					
Student 1	Acceptable	(402 +) 232	6	19	257
	Uncertain	45	3	20	68
	Wrong	19	3	21	43
	Σ	296	12	60	368
Correctness for coder 2					
Student 1	Acceptable	(402 +) 194	8	23	225
	Uncertain	35	7	20	62
	Wrong	27	12	42	81
	Σ	256	27	85	368
Correctness for interview coding					
Student 1	Acceptable	(402 +) 189	13	13	215
	Uncertain	54	12	16	82
	Wrong	33	15	23	71
	Σ	276	40	52	368

†Cases are analysed only if at least one human coder deviates from the code that was obtained via interview coding.

majority of the 368 problematic codes, both student assistants agreed that the codes are acceptable. Viewed as a proportion of all 770 interview-coded answers, coder 1, coder 2 and interview coding are acceptable in 82.3%, 77.4% and 76.8% of the cases. Coder 1 is rated best: 232 (plus 402 acceptable by assumption) of his assignments are considered acceptable, which is significantly more than the 194 (plus 402) acceptable codes from professional coder 2 (H_0: equal proportions of acceptable codes from coder 1 and coder 2; $\chi^2 = 5.53$; $p = 0.019$) and more than from interview coding (189 + 402). Our aspired goal to increase data quality with interview coding was not achieved. Coder 1 also produced the lowest number of wrong codes (21) among the three different coding procedures. All other codes from coder 1 are somewhere in between acceptable and wrong, with little agreement between the student assistants. The frequent disagreement between students cautions not to overinterpret these results.

To conclude, we find it difficult to determine a clear winner concerning optimal quality of data. Both the analysis of agreement rates and the students' evaluation provide evidence that differences in quality of data among coding procedures are quite small and may be negligible for practical purposes. Furthermore, there is frequent disagreement on the correct code, reflecting the complexity of occupation coding (see part C in the on-line appendix for coding examples). The students' evaluation shows that it is possible to identify obvious errors in all coding procedures, but these errors account for only a small proportion of the overall disagreement. In the next section, we illustrate why disagreement may occur.

4.4. Illustrative examples

Occupation coding in general and interview coding in particular have several particularities that are worth discussing in detail. The following two examples help us to understand weaknesses and to consider ways of improvement. The example of the 'vice director and teacher', which was introduced in Fig. 1, was chosen because it offers various insights. The second example ('truck salesman') was chosen because it is symptomatic for a type of error that we observed more than once in interview coding.

4.4.1. *Vice director and teacher*

In interview coding, the category 84114 'Teacher—elementary school' is selected, which is plausible given the last word from the text written down by the interviewer. Audiorecording confirms that this person is a vice principal and teacher at an elementary school. A professional coder would not have known that this person works at an elementary school because the interviewer failed to write down the complete information, making this answer a candidate for error.

Two professional coders were asked to code the textual answer. Both decided on the category 84194 'Managers in school of general education'. This category is the most appropriate from a post-interview coding perspective, for three reasons.

(a) Additional questions from the interview show that this person supervises 14 employees, indicating that managerial responsibilities may dominate his professional tasks as a teacher. This would favour the category 84194, because the main focus of activities performed in the job is, according to the 2010 GCO, the criterion to decide for the best-suited category.
(b) The alphabetic dictionary that is part of the 2010 GCO assigns 'vice principal' to code 84194. Note, however, that our algorithm does not recognize synonyms.
(c) The respondent answered 'vice director' before 'teacher'. Coding rules often determine that the first job title is coded if multiple titles are provided in the *verbatim* response and the other titles do not specify the first title.

The school manager category 84194, which both professional coders prefer, is missing in the list of suggested job titles. Only if the respondent had had the chance to choose this category, would one know whether he actually had preferred this category instead or still the category that he selected. The algorithm fails to find this or any other managerial category because the calculated phrase for text matching is 'TEACHER', which is not linked in any database to category 84194. The word 'director', however, could theoretically be linked via some databases to the desired category (and to many more managerial categories), but the text matching methods that we applied to those databases do not work if there are additional words besides the key term in the textual input. It is by no means an exception that relevant answer options are missing in the dialogue: the category which was selected by the professional coder 1 is missing for 36.0% of the eligible respondents. Upgraded algorithms and/or larger training data would be needed for improvement, although it still cannot be guaranteed that all appropriate categories will be suggested to the respondent.

Although one relevant category is missing in Fig. 1, other suggested job titles are less relevant: 'Sports teacher' is clearly implausible and the job titles 'Teacher—*Hauptschulen*' and 'Teacher—*Real-/Mittelschulen*' are repetitive. Both are associated with a single GCO category (84124), allowing respondents a detailed choice between two different school types. Yet, the GCO does not distinguish between both and it would be sufficient to ask for a single overarching category 'secondary school teacher' instead (non-existent in the *Dokumentationskennziffer*). As we restricted the number of suggested job titles to a maximum of five, such a reduction of answer options would create space for other possible categories.

4.4.2. *Truck salesman*

The second example illustrates a major mechanism of how interview coding leads to wrong and uncertain codes. Consider a person who sells trucks. Our algorithm for coding during the interview is not sufficiently intelligent to suggest the correct job title 'motor vehicle seller', which would lead to the correct job code 62272. Instead, the respondent chooses the more general job title 'salesman' which appears correct to him. Unfortunately, this job title is associated with category 62102 titled 'Sales occupations in retail trade (without product specialization)', which is the wrong code for this person's actual job. The point here is that job titles from the *Dokumentationskennziffer* are not well suited to support coding during the interview. Textbooks recommend using the most specific, unambiguous terms for question design and avoiding an overlap between the answer options (e.g. Tourangeau *et al.* (2000) and Krosnick and Presser (2010)); yet, this is quite difficult to accomplish for a question about occupation. In the *Dokumentationskennziffer*, many general job titles, such as 'salesman', exist. This causes the risk that people might select a job title which appears to be correct but which leads, in fact, to a wrong code. To eliminate this type of error, one might try to reword or delete all general job titles in the *Dokumentationskennziffer* so that the meaning becomes clearer and the respondents will in no case prefer an incorrect answer option over the alternative 'other occupation'. In doing so, quality is likely to improve, but the proportion of interview-coded answers will probably decrease.

4.5. Interviewer behaviour

Our new technique was tested in a telephone survey. Compared with self-administered surveys in which the respondents can be confronted directly with the answer options suggested, the telephone survey has an extra level of interaction between respondents and interviewers. Interviewers are trained to follow the rules of standardized interviews, i.e. they are supposed to read the questions and answers exactly as worded and respondents are supposed to select the most appropriate answer without any help from the interviewer (e.g. Fowler and Mangione (1990) and Schaeffer *et al.* (2010)). This general training was not repeated for our particular survey. Since interviewers often spontaneously choose to violate these guidelines for the proposed question on occupation, it is relevant to describe how the interview-coded occupations are obtained.

Immediatly before the job titles are suggested, the algorithm needs a few seconds to calculate the most plausible job titles. Although interviewers are provided with a standardized text to explain the situation, interviewers may feel the need to keep the conversation running and to fill the gap by explaining what comes next in their own words. When the answer options pop up, it is often not necessary to read the exact question text ('Are you employed in one of the following occupations?') to proceed with the interview. In 177 out of 209 interviews (85%) that were subject to behaviour coding, the question text was not read.

Frequently, job titles are automatically suggested although they are definitely not appropriate. In the above-mentioned example (see Fig. 1), the interviewer knows from the preceding conversation that the list of suggestions contains only one job title that is appropriate in her view. Not reading out inappropriate suggestions saves time and prevents possibly confusing the respondent. This makes it attractive for interviewers to skip inappropriate job suggestions. In 97 out of 209 interviews (46%), at least one suggested job title was not read. In 10 cases (10%) this happened because the algorithm found a job title that is identical to the verbal answer that was previously provided by the respondent, in 35 cases (36%) because the job titles suggested are definitely inappropriate, and in 23 cases (24%) because of both reasons. Some interviewers steer respondents towards a specific answer: in 27 out of 209 interviews (13%), the interviewer reads out only a single job title, typically formulated in the form of a question (e.g. 'Here we have.... Is this correct?'), but sometimes also formulated as a statement, so that the respondent is not required to confirm this job title. In eight interviews (4%), the interviewers did not read out loud the suggestions at all but independently selected the most appropriate answer option.

It is also very common for interviewers to skip the answer option 'other occupation', which was read to 37 out of 209 respondents (18%) only. Reasons for this might be that the answer option is highlighted in the question text or because interviewers think that an appropriate job title had already been found.

Every question should usually be followed by an appropriate answer from the respondent. In a first reaction, 156 out of 209 respondents (75%) provided such an answer, either interrupting the interviewer (21 people) or naming it after the interviewer had finished reading out the entire question (135 people). Normally, this answer marks the end of the occupation coding process unless the respondent chooses 'other occupation' or the interviewer starts to reason with the respondent about a more appropriate category, as we have observed in a few interviews. Cases in which the respondents do not give an appropriate answer immediately are more problematic. If no job title is appropriate at first sight, respondents hesitate to answer. Because of this confusion, 17 respondents (8%) mentioned additional details about their jobs and, as a result, 'other occupation' was most often selected. Another 18 respondents (9%) were confused or asked the interviewer to explain or repeat the job title suggested. 14 out of the 18 respondents eventually agreed with one of the suggestions. In 18 additional interviews (9%), the respondents did not have a chance to speak because the interviewer was thinking or saying something without asking a question. It is then typically the interviewer who selects the most appropriate answer option.

In summary, our exercise in behaviour coding shows that many interviewers did not closely follow the rules for standardized interviews. It is the exception that an interviewer reads out the exact question text and all answer options, including the last option for 'other occupation'. When an interviewer skips a job title, decides all by himself or herself without asking the respondent, or starts a discussion with the respondent about the most appropriate answer option, one might worry that interviewer effects can be large for this question. However, these problems should not be exaggerated. Many skipped job titles are definitely inappropriate, typically respondents and not interviewers make the decision and it is not clear whether data quality is diminished when interviewers play an overly active part, as they often have a good understanding of the respondent's job. Instead, they often have good reasons for departures from the script. For future improvements of the instrument, the interplay between interviewer, question (length, number of categories, formulation) and respondent should be considered an important issue.

4.6. Algorithm analysis

Another element contributing to the overall success is the algorithm itself. The prediction algorithm should provide job category suggestions for as many respondents (i.e. verbal answers) as possible. Furthermore, these categories should be of high quality so that the respondents find their own jobs in the suggested list. In what follows, we analyse how well the algorithm currently performs regarding both objectives and search for possible ways of improvement.

Any algorithm must match the verbal responses given by respondents with some database containing possible categories. To find possible job categories for a maximal number of respondents, we apply three different databases: our training data consist of 14 912 unique entries, the search word catalogue has 153 588 entries and there are 24 000 entries in the alphabetic dictionary which is part of the 2010 GCO. However, a larger size of the database does not imply more matches. Matching respondents' answers with identical entries in the respective database provides job category suggestions for 45.7%, 46.5% and 40.8% of the 1064 respondents who answered the open-ended questions on employment. Despite the different sizes of the databases, these numbers are remarkably similar, probably because the alphabetic dictionary and the search word catalogue were not constructed for our purpose.

Many respondents reply to the open question with common and precise one-word job titles that can easily be matched with any database. These people are easy to code, either during or after the interview. In our sample, 33.6% of the respondents provided answers that enable identical matching with any database, showing that the different databases have an enormous overlap.

However, all databases fail to make suggestions via exact matching for at least half of the respondents. To overcome this limitation, two additional inexact matching methods were implemented. Results for all the different text matching methods and all databases are shown in Table 6. When the verbal answer is not required to be identical with a database record but only needs to be a substring of it, more matches are found (49.2% *versus* 45.7% and 51.8% *versus* 46.5%), but the gains are relatively small. This is because this matching technique is appropriate only for short answers. 349 respondents (32.8%), however, provided longer answers with at least three words (operationalized by two blank characters), of which only 45 can be matched with the above-mentioned identical and substring matching methods.

The second inexact matching method is more promising for longer answers: when searching for a meaningful subsequence of words (equal to a phrase as defined above) in the original verbal answer, which is then again matched to the different databases, the number of matches increases considerably, as can be seen in the lower half of Table 6.

Column (2), 'Percentage of respondents for whom at least one suggested category was also coded by at least one professional coder', confirms that we find suitable matches with all methods. For most respondents and any matching method, categories are suggested that are relevant in

Table 6. Descriptive results for various matching methods and databases†

m	*Matching method*	*(1) (%)*	*(2) (%)*	*(3)‡*		
				Median	*Mean*	*Maximum*
	Answer matches with training data					
1,3,4	Identical	45.7	39.9	2	4.2	45
2	Answer is substring	49.2	43.1	4	8.0	122
	Answer matches with file of search words					
6	Identical	46.5	39.7	2	3.8	66
7	Answer is substring	51.8	46.9	5	12.6	187
8	Answer matches with alphabetic dictionary	40.8	38.9	GCO/DKZ 2/23	GCO/DKZ 4.8/71.3	GCO/DKZ 69/1012
	Phrase matches with training data					
9–11	Identical	73.9	57.0	3	7.0	45
—§	Answer is substring	82.1	69.8	8	57.3	1479
	Phrase matches with file of search words					
13	Identical	71.4	52.3	3	6.8	82
14	Answer is substring	83.7	72.5	12	133.6	3878
15	Phrase matches with alphabetic dictionary	57.2	52.3	GCO/DKZ 2/30	GCO/DKZ 7.5/94.7	GCO/DKZ 96/1190

†Column (1), percentage of respondents for whom the matching method suggests at least one category. Column (2), percentage of respondents for whom at least one suggested category was also coded by at least one professional coder. Column (3), average number of categories, provided that at least one category is suggested.

‡GCO/DKZ: the alphabetic dictionary links job titles only to categories from the 2010 GCO. All *Dokumentationskennziffer* categories that are associated with the so-found GCO categories are possible candidates for suggestion. We thus provide the number of GCO suggestions first and the number of *Dokumentationskennziffer* suggestions second.

§This matching method was not included in the production software.

Table 7. Productivity of the coding system under various hypothetical situations†

Ask first inquiry only if …	*(1)*	*(2)*	*(2)/(1) (%)*	*(3)*	*(3)/{(1)–(2)} (%)*
Condition (a):… identical match with training data and match with alphabetic dictionary	386	12	3.1	312	83.4
Condition (b):… no shorter phrase is found	532	27	5.1	416	82.4
Condition (c):… no shorter phrase is found or phrase matches with alphabetic dictionary	712	60	8.4	511	78.4
Condition (d): always (actual condition in this study)	915	145	15.8	574	74.5

†Column (1), number of respondents who would be asked under the given condition. Column (2), number of respondents who answer ‘other occupation’ under the given condition. Column (2)/(1), column (2) divided by column (1). Column (3), number of respondents under the given condition who select a code that is in agreement with at least one professional coder. Column (3)/{(1)−(2)}, column (3) divided by the difference between columns (1) and (2).

the sense that professional coders usually select one of the suggested categories independently. This is not self-evident—especially in the case of the phrase matching methods it does happen that the phrase itself is meaningless for coding (e.g. words like ‘in’ or ‘and’) and matching such a phrase certainly brings no improvement.

The downside of inexact matching is summarized in column (3), ‘Average number of categories if at least one category is suggested’. Identical matching methods usually suggest small numbers of possible categories and inexact matching methods find larger numbers. Obviously, not all suggested categories are always appropriate for a given occupation and it is also prohibitive to suggest dozens or hundreds of categories to a respondent during the interview. The overall performance of the system shows that these difficulties are well absorbed by the gradient boosting algorithm, which calculates correctness probabilities for all categories that are suggested by any matching method. Boosting thus integrates the different matching methods to a single prediction algorithm and allows finding the most probable categories.

These descriptions suggest a trade-off with each additional matching method. On the one hand, adding a matching method offers the possibility that additional categories can be suggested to the respondents. On the other hand, suggesting more categories can also mean suggesting more unsuitable categories, which may protract the interview, induce more people to choose ‘other occupation’, or lead to inaccurate coding. Therefore, system improvements might be expected if candidate job categories are not suggested to all possible respondents but only to a subgroup for which the matching methods meet specific criteria. Residual respondents would not come in contact with our proposed system. We searched for corresponding criteria and found three possible conditions to be particularly meaningful. Table 7 presents the hypothetical results for a modified algorithm, i.e. it shows what would have happened if these conditions had been applied in the field. The conditions are as follows.

(a) Answers have identical matches in both the training data and the alphabetic dictionary.
(b) No shorter phrase is found. This condition comprises all cases from the first condition with only two exceptions.
(c) The second condition holds or, alternatively, a phrase is found that must match with the alphabetic dictionary. A match with the alphabetic dictionary confirms that the phrase is a job title which makes this term especially relevant for coding.

Column (1) in Table 7 shows that the number of respondents who are presented with job category suggestions increases when the conditions are loosened, allowing more respondents to code their occupation during the interview. At the same time, not only the absolute number (column (2)) but also the proportion (column (2)/(1)) of respondents who select ‘other occupation’ increases. This is detrimental to the original goal of keeping interview times in check because those respondents would be asked an additional open question. Furthermore, the proportion of respondents who select a code that is in agreement with at least one professional coder (column (3)/$\{(1) - (2)\}$) decreases when the conditions are loosened, suggesting that the quality of interview coding is also affected. The trade-off hypothesis is thus confirmed.

Which condition should be chosen to find an optimal balance between both objectives? In our opinion, condition (c) is best. (712−60)/1064 = 61.3% of the respondents would have chosen a job title during the interview under this condition, which is still a considerable proportion. At the same time, only 60/1064 = 5.6% of the population would have selected ‘other occupation’, which is a substantial improvement. It is not acceptable to have $(145 - 60)/(915 - 712) = 41.9\%$ of the respondents who do not fulfil condition (c) select ‘other occupation’, as it was implemented in the tested system.

This result also has implications for our algorithm. Job category suggestions are satisfactory when verbal answers are short and can be matched by identical or substring matching to any database. The predictions are still sufficiently accurate if the algorithm can extract a phrase from a multiworded verbal answer that is a job title from the alphabetic dictionary. The remaining verbal answers require more attention to improve the algorithm further. They may be characterized as follows. The algorithm finds a shorter phrase that is not listed in the alphabetic dictionary for 203 verbal answers. These answers contain at least two words—often more—but frequently lack a single job title that would be most relevant for coding. Algorithms that exploit interactions between words can prove useful here but were not employed so far. For 106 answers the algorithm does not find a single match in any database. These answers usually consist of a single word. Spelling errors and compound words are frequent reasons why matching is not possible. Future improvements of the algorithm should address these problems.

To motivate our algorithm, we claimed that better predictions can be achieved when the algorithm learns not only from training data but also from existing databases. Did we succeed? We compare our algorithm with predictions from multinomial regression with elastic net regularization as implemented by Friedman *et al.* (2010). We use the same training data with identical covariates as before to train the multinomial regression model. The full text, as obtained from the first verbal answer, is converted to a document term matrix that counts how often each word appears in the respondents’ answers. As the software package requires that each category in the outcome variable occurs at least twice, we remove 680 cases from the training data whose *Dokumentationskennziffer* codes occur only once. All covariates are dummy coded. The problem is thus to predict one of 1600 *Dokumentationskennziffer* codes from a sparse predictor matrix of size 32 275 observations times 10 930 columns. We explore various tuning parameters to estimate the model and obtain best results when setting $\alpha = 0.05$ and $\lambda = 0.001$, implying only weak regularization. This model is used to predict job titles and associated correctness probabilities in the current study for all 1041 respondents for whom we have codes from coder 1 available. After running the same procedure as in our original algorithm to select five job titles, multinomial regression suggests at least one job title from the same GCO category that was chosen by coder 1 for 557 respondents (54%). Our own algorithm reaches a slightly better performance, suggesting at least one job title from the same GCO category for 578 respondents (56%). Although a sceptic may argue that this small improvement does not justify the complexity of our algorithm, we are more optimistic and suggest including the predictions that are obtained via multinomial regression as another covariate in the boosting procedure. This should improve the performance of our own algorithm even further.

4.7. Classification material

Two features of the classification material should be highlighted as contributing factors to coding errors, both in interview coding and in traditional post-survey coding efforts.

First, a classical strategy for automatic occupation coding is to search for a given job title in a database and to assign the associated category accordingly. We matched the first verbal answer from the interview to a database that we prepared from the alphabetic dictionary of 24 000 job titles that is part of the 2010 GCO. Although we matched job titles only if they were clearly associated with a single category, successful exact database matches were found for 418 out of 1064 verbal responses. For these people, it was then possible to compare the codes with those obtained from manual coding (coder 1 and coder 2), with the following results: all three codes are identical for 307 responses (73.4%), only one manual coder agreed with the code from the database for 88 responses (21.1%) and both disagreed with the database for 23 responses (5.6%). These numbers show that a substantial proportion of respondents mention job titles that can be coded automatically in some category with the alphabetic dictionary, though this does not mean that these categories are the only possible categories. Manual coders frequently disagree with those codes and base their decision on more information, which they retrieve from additional answers. Many job titles exist whose semantic content is vague and does not uniquely determine a single correct job category. If a coding technique relies on vague job titles—and the proposed system for coding during the interview does so excessively, like many other approaches—we cannot hope for an optimal quality of coding which guarantees that every respondent will be classified in the category that describes his or her occupational tasks and duties best.

Another source of error that leads to low intercoder reliabilities can be found in both manual and interview coding. Coders are usually required to select a single correct category; multiple categories are not permitted, even if appropriate. The decision for a single category can be difficult, either because information from the respondent to determine a precise category is missing or because categories from the job classification are not pairwise disjoint and, as a consequence, the occupational activity does not belong to a single category. The following numbers indicate that this issue requires further attention. When looking only at the subset of respondents for which both student assistants agreed that the assigned codes from coder 1 and coder 2 both are acceptable, we can have high confidence that both codes for this subset of 137 respondents are correct. However, for 52 respondents in this subset, both codes are different and it appears that more than one category may be considered correct.

5. Summary and conclusion

Traditional coding of occupations is costly and time consuming. In our study, two independent coders obtained a reliability of 61.11%: a number that is low but by no means an exception. We implemented and tested a technical solution with increased interaction during the interview to counter these challenges. After a verbal answer has been entered in the interview software, the computer automatically calculates a small set of possible job categories and suggests them to the respondent, who in turn can select the most appropriate. Our results show that this strategy for interactive coding during the interview is technically feasible.

Our system achieves high productivity: 72.4% of the respondents choose an occupation during the interview. The proportion for which manual coding is still necessary is thus reduced to 27.6%. This result is promising because coding costs can be saved and data are available directly after the interview.

The quality of interview coding was compared with that of two professional coders and was found to be slightly lower than the quality of the first coder and comparable with the quality of the second. We also find frequent disagreement between both coders, which can be partly attributed to a lack of information provided by the respondents and to the fact that both coders observed different coding rules. Our desire to increase the quality of the coded occupations by collecting more information already during the interview was not fulfilled for several reasons: categories that are suggested by the algorithm are sometimes inappropriate, the two generated follow-up questions are unsuited to elicit more appropriate codings and respondents occasionally select overly general job titles, which lead to incorrect categories.

For respondents whose occupations are coded successfully during the interview, the duration of the interview is reduced by a few seconds; others who do not select one of the categories suggested will have to bear the burden of slightly longer interviews with an additional question. This is a major drawback of the tested system, affecting 13.6% of the population.

Our system was optimized to achieve high productivity. This may not be the best strategy because marginal gains at high levels of productivity imply larger costs in terms of the number of people who will have to endure longer interviews. We instead suggest a different strategy that finds an optimal balance between both objectives. For this, we identify four conditions that are easy to implement in the current algorithm. One condition, which would decrease the productivity rate from 72.4% to 61.3%, is recommended in particular because, under this condition, fewer respondents (5.6% compared with 13.6% now) would have to bear the burden of longer interviews.

These results are satisfactory for the first trial of a complex instrument. The key component of the system proposed—a machine learning algorithm that suggests possible answer options during the interview—works well. Other minor features were tested but their results are discouraging. Some obvious adaptions would be necessary for future application. In addition, it would be useful to estimate whether the instrument proposed leads to cost reductions in the coding process. At the very least, our results show that coding during the interview can become a viable technique that may partly replace traditional post-interview coding in the future.

Before implementing the new instrument in a production environment, we recommend further testing in more practical settings. Our study has some limitations and survey operators may want to ask the following questions for their own application. First, is it possible to achieve a similar or better performance if some occupations were not underrepresented, as we have reported in our study? Second, what would happen if the interviewer and the algorithm had less information to predict a person's job from occupation-related questions preceding the new tool? Third, what would have to be changed in the proposed instrument if the researcher was not interested in self-reported occupations from telephone interviews, but in other types of occupation (e.g. job aspirations of adolescents and occupations of spouses and parents) that might be collected via different modes of operation (e.g. Internet surveys or computer-assisted personal interviewing)?

Throughout this paper, we described the strengths and weaknesses of the proposed instrument. For future developments, we have identified the following factors how to improve the process.

A supervised learning algorithm was used to generate plausible job category suggestions for the respondents. With an improved algorithm and additional training data, it is likely that the productivity of the system can be further increased. In the frequent situation that a verbal answer comprises more than one word and does not contain a predefined job title, we suspect largest gains in productivity. Spelling correction and the splitting of compound words may also prove to be helpful.

When respondents choose one of the job titles suggested, it is too often not the most appropriate. Respondents frequently select general job titles that are not entirely wrong but link to suboptimal GCO categories. These inappropriate job titles stem from the *Dokumentationskennziffer*, which is therefore not well suited for coding during the interview. To preclude the possibility that respondents select an incorrect category, we recommend the development of an auxiliary classification that describes answer options more precisely. All answer options from this auxiliary classification should map to a single category in both classifications, national (2010 GCO) and international (2008 ISCO), for simultaneous coding.

Interviewers frequently did not act according to the rules of standardized interviews at the question proposed but often preferred rewording the question text and skipping suggested answer options. Although this behaviour leads to concerns about interviewer effects, we must not forget the positive effect: respondents are not confused by strange answer options and the duration of the interview is reduced. For an improved instrument, one may even try to provide interviewers with a medium-sized number of answer options (say 10). Since respondents cannot intellectually process so many answer options in a telephone interview, one would also explicitly request interviewers to skip inappropriate job categories. This procedure could partly remedy the current problem that the algorithm finds many possible job titles, but the most appropriate job category is not suggested to about 36% of the respondents. Furthermore, extended interviewer training will be necessary to ensure that interviewers know when they must follow the script and to reduce the risk of omitting relevant answer options.

Some answers in reply to the first open-ended question about occupation are very general and one would need to suggest a huge number of possible categories. Instead, our vision is to recognize these general answers automatically. An additional open-ended question would then be asked to collect more details and this second answer could be used as input for coding during the interview. Additionally, future research should consider the possibility that more than one job category may be appropriate.

In summary, such a system for occupation coding during the interview promises an increase of quality of data while reducing costs of data collection.

Acknowledgements

Funding for this work has been provided by the German Institute for Employment Research and the Mannheim Centre for European Social Research, and by grant KR 2211/3-1 from the German Research Foundation to Frauke Kreuter. The idea for this study originates from the Master's thesis written by Malte Schierholz. Miriam Gensicke and Nikolai Tschersich contributed with valuable comments and supervised the implementation in the survey software, which was technically demanding. We thank Alexandra Schmucker for helping us to test the proposed technique in a survey commissioned by the Institute for Employment Research and operated by Kantar Public (formerly TNS Infratest Sozialforschung). This study would not have been possible without Ariane Wickler and Gerd Döring, who implemented the interface between the interview software and the predictive system that was developed by the first author. Valuable comments from Josef Hartmann led to improvements in the questionnaire. We sincerely thank the Associate Editor and the referees for their constructive comments. We further thank the three coders, our student assistants (Max Hansen and Sebastian Baur) for quality checking, Hannah Laumann for proofreading and colleagues at the German Institute for Employment Research and the University of Mannheim for helpful comments.

References

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M. and Trahms, A. (2010) Arbeiten und Lernen im

Wandel ∗ Teil 1: Überblick über die Studie. *Methodenreport 05/2010*. Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Belloni, M., Brugiavini, A., Meschi, E. and Tijdens, K. (2016) Measuring and detecting errors in occupational coding: an analysis of share data. *J. Off. Statist.*, **32**, 917–945.

vom Berge, P., König, M. and Seth, S. (2013) Sample of integrated labour market biographies (SIAB) 1975-2010. *Datenreport 01/2013*. Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Biemer, P. and Caspar, R. (1994) Continuous quality improvement for survey operations: some general principles and applications. *J. Off. Statist.*, **10**, 307–326.

Biemer, P. and Lyberg, L. (2003) *Introduction to Survey Quality*. Hoboken: Wiley.

Billiet, J. and Loosveldt, G. (1988) Improvement of the quality of responses to factual survey questions by interviewer training. *Publ. Opin. Q.*, **52**, 190–211.

Bobbitt, L. G. and Carroll, C. D. (1993) Coding major field of study. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 177–182.

Bradburn, N. M. (1978) Respondent burden. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 35–40.

Breiman, L. (1996) Stacked regressions. *Mach. Learn.*, **24**, 49–64.

Bundesagentur für Arbeit (2011a) *Klassifikation der Berufe 2010*, vol. 1, *Systematischer und Alphabetischer Teil mit Erläuterungen*. Nuremberg: Bundesagentur für Arbeit.

Bundesagentur für Arbeit (2011b) *Klassifikation der Berufe 2010*, vol. 2, *Definitorischer und Beschreibender Teil*. Nuremberg: Bundesagentur für Arbeit.

Bundesagentur für Arbeit (2013) Index of search words. Bundesagentur für Arbeit, Nuremberg. (Available from `http://download-portal.arbeitsagentur.de/files/`.)

Bushnell, D. (1998) An evaluation of computer-assisted occupation coding. In *New Methods for Survey Research* (eds A. Westlake, J. Martin, M. Rigg and C. Skinner), pp. 23–36. Southampton: Association for Survey Computing.

Campanelli, P., Thomson, K., Moon, N. and Staples, T. (1997) The quality of occupational coding in the United Kingdom. In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz and D. Trewin), pp. 437–453. New York: Wiley.

Cantor, D. and Esposito, J. (1992) Evaluating interviewer style for collecting industry and occupation information. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 661–666.

Conrad, F. G., Couper, M. P. and Sakshaug, J. W. (2016) Classifying open-ended reports: factors affecting the reliability of occupation codes. *J. Off. Statist.*, **32**, 75–92.

Conrad, F. G. and Schober, M. F. (2005) Promoting uniform question understanding in today's and tomorrow's surveys. *J. Off. Statist.*, **21**, 215–231.

Couper, M. and Zhang, C. (2016) Helping respondents provide good answers in web surveys. *Surv. Res. Meth.*, **10**, 49–64.

Creecy, R. H., Masand, B. M., Smith, S. J. and Waltz, D. L. (1992) Trading mips and memory for knowledge engineering. *Communs ACM*, **35**, 48–64.

Drasch, K., Matthes, B., Munz, M., Paulus, W. and Valentin, M.-A. (2012) Arbeiten und Lernen im Wandel ∗ Teil V: Die Codierung der offenen Angaben zur beruflichen Tätigkeit, Ausbildung und Branche. *Methodenreport 04/2012*. Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Elias, P. (1997) Occupational classification (ISCO-88): concepts, methods, reliability, validity and cross-national comparability. *Labour Market and Social Policy Occasional Paper 20*. Organisation for Economic Cooperation and Development Publishing, Paris. (Available from `http://dx.doi.org/10.1787/304441717388`.)

Elias, P., Birch, M. and Ellison, R. (2014) CASCOT international version 5 user guide. Institute for Employment Research, University of Warwick, Coventry. (Available from `http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/internat/`.)

Fowler, F. J. and Mangione, T. W. (1990) *Standardized Survey Interviewing: Minimizing Interviewer-related Error*. Newbury Park: Sage.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Galesic, M. and Bosnjak, M. (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. *Publ. Opin. Q.*, **73**, 349–360.

Geis, A. (2011) Handbuch für die Berufsvercodung. *Coding Documentation*. GESIS–Leibniz-Institut für Sozialwissenschaften, Mannheim. (Available from `http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/handbuch_der_berufscodierung_110304.pdf`.)

Geis, A. and Hoffmeyer-Zlotnik, J. H. (2000) Stand der Berufsvercodung. *ZUMA-Nachr.*, **24**, 103–128.

Granquist, L. and Kovar, J. (1997) Editing of survey data: how much is enough? In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz and D. Trewin), pp. 415–435. New York: Wiley.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017) Three methods for occupation coding based on statistical learning. *J. Off. Statist.*, **33**, 101–122.

Hacking, W., Michiels, J. and Janssen-Jansen, S. (2006) Computer assisted coding by interviewers. In *Proc. 10th Int. Blaise Users Conf.* (ed. J. Bethlehem), pp. 283–296. Arnhem: International Blaise User Group.

Hoffmann, E., Elias, P., Embury, B. and Thomas, R. (1995) What kind of work do you do?: Data collection and processing strategies when measuring "occupation" for statistical surveys and administrative records. *STAT Working Paper. 95–1*. Bureau of Statistics, International Labour Office, Geneva. (Available from http://www.ilo.org/public/libdoc/ilo/1995/95B09_135_engl.pdf.)

Hoffmeyer-Zlotnik, J. H., Hess, D. and Geis, A. J. (2006) Computerunterstützte Vercodung der International Standard Classification of Occupations (ISCO-88): Vorstellen eines Instruments. *ZUMA-Nachr.*, **30**, 101–113.

Hoffmeyer-Zlotnik, J. H. and Warner, U. (2012) *Harmonisierung Demographischer und Sozioökonomischer Variablen: Instrumente für die International Vergleichende Surveyforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Holbrook, A. L., Green, M. C. and Krosnick, J. A. (2003) Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. *Publ. Opin. Q.*, **67**, 79–125.

Holland, J. L. and Christian, L. M. (2009) The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Socl Sci. Comput. Rev.*, **27**, 196–212.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010) Model-based boosting 2.0. *J. Mach. Learn. Res.*, **11**, 2109–2113.

International Labour Office (2012) *International Standard Classification of Occupations: ISCO-08*. Geneva: International Labour Organisation.

Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M. and Kang, T. S. K. (2015) Carotene: a job title classification system for the online recruitment domain. In *Proc. 1st Int. Conf. Big Data Computing Service and Applications, Redmond City*, pp. 286–293. New York: Institute of Electrical and Electronics Engineers.

Jung, Y., Yoo, J., Myaeng, S.-H. and Han, D.-C. (2008) A web-based automated system for industry and occupation coding. In *Web Information Systems Engineering* (eds J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim and X. Wang), pp. 443–457. Berlin: Springer.

Krosnick, J. and Presser, S. (2010) Question and questionnaire design. In *Handbook of Survey Research* (eds P. V. Marsden and J. D. Wright), pp. 263–313. Bingley: Emerald.

LeBlanc, M. and Tibshirani, R. (1996) Combining estimates in regression and classification. *J. Am. Statist. Ass.*, **91**, 1641–1650.

Loos, C., Eisenmenger, M. and Bretschi, D. (2013) Das Verfahren der Berufskodierung im Zensus 2011. *Wirtsch. Statist.*, 173–184.

Maaz, K., Trautwein, U., Gresch, C., Lüdtke, O. and Watermann, R. (2009) Intercoder-Reliabilität bei der Berufscodierung nach der ISCO-88 und Validität des sozioökonomischen Status. *Zeits. Erziehungs.*, **12**, 281–301.

Mangione, T. W., Fowler, F. J. and Louis, T. A. (1992) Question characteristics and interviewer effects. *J. Off. Statist.*, **8**, 293–307.

Mannetje, A. T. and Kromhout, H. (2003) The use of occupation and industry classifications in general population studies. *Int. J. Epidem.*, **32**, 419–428.

Measure, A. (2014) Automated coding of worker injury narratives. *Proc. Gov. Statist. Sect. Am. Statist. Ass.*, 2124–2133.

Office for National Statistics (2003) Quality of data capture and coding: evaluation report. *Census 2001 Review and Evaluation Report*. Office for National Statistics, Titchfield. (Available from http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/processing/quality-of-data-capture-and-coding—evaluation-report.pdf.)

Ongena, Y. P. and Dijkstra, W. (2016) Methods of behavior coding of survey interviews. *J. Off. Statist.*, **22**, 419–451.

Paulus, W. and Matthes, B. (2013) Klassifikation der Berufe ∗ Struktur, Codierung und Umsteigeschlüssel. *Methodenreport 08/2013*. Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Roberts, C., Gillian, E., Allum, N. and Lynn, P. (2010) Data quality in telephone surveys and the effect of questionnaire length: a cross-national experiment. *Research Working Paper 2010-36*. Institute for Social and Economic Research, University of Essex, Colchester. (Available from https://www.iser.essex.ac.uk/publications/working-papers/iser/2010-36.)

Sakshaug, J. W., Schmucker, A., Kreuter, F., Couper, M. P. and Singer, E. (2016) Evaluating active (opt-in) and passive (opt-out) consent bias in the transfer of federal contact data to a third-party survey agency. *J. Surv. Statist. Methodol.*, **4**, 382–416.

Schaeffer, N. C., Dykema, J. and Maynard, D. W. (2010) Interviewers and interviewing. In *Handbook of Survey Research* (eds P. V. Marsden and J. D. Wright), pp. 437–470. Bingley: Emerald.

Schierholz, M. (2014) Automating survey coding for occupation. *Methodenreport 10/2014*. Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Schnell, R. (2012) *Standardisierte Befragungen in den Sozialwissenschaften*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Speizer, H. and Buckley, P. (1998) Automated coding of survey data. In *Computer Assisted Survey Information Collection* (eds M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II and J. M. O'Reilly), pp. 223–243. New York: Wiley.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.

Svensson, J. (2012) Quality control of coding of survey responses at Statistics Sweden. In *Proc. Eur. Conf. Quality in Official Statistics*. Athens: Hellenic Statistical Authority–Eurostat.

Thompson, M., Kornbau, M. E. and Vesely, J. (2014) Creating an automated industry and occupation coding process for the American Community Survey. *Federal Economic Statistics Advisory Committee Meet*. US Census Bureau, Suitland. (Available from `http://www.census.gov/about/adrm/fesac/meetings/june-13-2014-meeting.html`.)

Tijdens, K. (2014a) Reviewing the measurement and comparison of occupations across Europe. *Working Paper 149*. Amsterdam Institute for Advanced Labour Studies, University of Amsterdam, Amsterdam. (Available from `http://hdl.handle.net/11245/1.432281`.)

Tijdens, K. (2014b) Dropout rates and response times of an occupation search tree in a web survey. *J. Off. Statist.*, **30**, 23–43.

Tijdens, K. (2015) Self-identification of occupation in web surveys: requirements for search trees and look-up tables. In *Survey Insights: Methods from the Field* (eds H. Best, P. Farago, D. Joye, L. Kaczmirek, C. Vandenplas, M. Vettovaglia and C. Wolf). Lausanne: Swiss Foundation for Research in Social Sciences–GESIS–Leibniz Institute for the Social Sciences.

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

United Nations and International Labour Office (2010) *Measuring the Economically Active in Population Censuses: a Handbook*. New York: United Nations and International Labour Office.

de Waal, T., Pannekoek, J. and Scholtus, S. (2011) *Handbook of Statistical Data Editing and Imputation*. Hoboken: Wiley.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix to: Occupation coding during the interview'.