Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

IAB-Discussion Paper 17/2016

Articles on labour market issues

Occupation coding during the interview

Malte Schierholz
Miriam Gensicke
Nikolai Tschersich

ISSN 2195-2663

Occupation Coding During the Interview

Malte Schierholz (IAB and Mannheim Centre for European Social Research, University of Mannheim)
Miriam Gensicke (TNS Infratest Sozialforschung)
Nikolai Tschersich (TNS Infratest Sozialforschung)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

Abstract

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

Zusammenfassung

Die Erfassung des Berufs geschieht in Umfragen üblicherweise mithilfe offener Fragen. Anschließend ist eine Kodierung der Freitextantworten notwendig, was teuer und fehleranfällig ist. Wir beschreiben einen alternativen Ansatz, bei dem die Kodierung bereits während des Interviews erfolgt. Die neue Methode verwendet Algorithmen des maschinellen Lernens um mögliche Berufskategorien automatisch vorherzusagen. Die so erzeugten Vorschläge werden dem Befragten vorgelegt, der dann sofort die am besten passende Kategorie auswählen kann. 72.4% der Teilnehmer einer Telefonbefragung haben auf diese Weise ihren Beruf direkt während des Interviews kodiert, was mögliche Kosteneinsparungen impliziert. Um weitere Verbesserungen des neuen Instruments zu ermöglichen, identifizieren wir verschiedene Faktoren, wie auch die Qualität der Kodierung erhöht und die Dauer der Interviews verkürzt werden kann.

JEL classification: C830, J400

Keywords: Coding, Interview Coding, Measurement Error, Occupation, Open-ended Questions, Supervised Learning

Acknowledgements: Funding for this work has been provided by the German Institute for Employment Research (IAB), the Mannheim Centre for European Social Research (MZES), and by grant KR 2211/3-1 from the German Research Foundation (DFG). The idea for this study originates from the master's thesis written by Malte Schierholz. Thanks to Frauke Kreuter and Alexandra Schmucker, it was possible to test the proposed technique in a survey commissioned by the Institute for Employment Research (IAB) and operated by TNS Infratest Sozialforschung. Miriam Gensicke and Nikolai Tschersich contributed with valuable comments and supervised the implementation in the survey software, which was technically demanding. 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 three anonymous coders, our student assistants (Max Hansen and Sebastian Baur) for quality checking, Hannah Laumann for proof-reading, and colleagues at the IAB for helpful comments.

1 Introduction

Occupation is a core organizational principle in our society. Researchers from many different disciplines have an interest in measuring occupation, for example to capture individual's 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, e.g., in terms of the *Standard International Occupational Prestige Scale* (SIOPS), the *class scheme of Erikson, Goldthorpe and Portocarero* (EGP), or the *International Socio-Economic Index* (ISEI) (c.f. Hoffmeyer-Zlotnik/Warner, 2012: p. 191). Many data collections ask for occupation, including the *United Kingdom Census*, which yielded almost 30 million verbatim answers on employment in 2001 (Office for National Statistics, 2003), and the register-based *German Census 2011* with 3.6 million verbatim answers (Loos/Eisenmenger/Bretschi, 2013). The *American Community Survey* also contains questions on occupation, collecting approximately two million responses annually (Thompson/Kornbau/Vesely, 2014). Similar questions are also 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 to subsequently code these verbatim answers into a classification scheme with hundreds of categories and thousands of jobs. This coding task is nontrivial. Conrad/Couper/Sakshaug (2016) discuss various reasons why quality may be compromised. For example, verbatim responses are sometimes ambiguous and fit well into multiple categories. Some respondents have occupations for which no adequate category exists. Because the target classification is fixed in advance, category modifications that could account for such difficulties are not possible. Still, coders are typically required to decide for a single most adequate job category. Several studies review the quality of coding occupational information under a variety of different conditions (e.g., language, target classification, coding rules and procedures, and coder's experience may differ) and report agreement rates when different persons code the same answers. Campanelli et al. (1997) have three British expert coders validating original codes from a number of non-experts, finding accuracies between 69% and 85%. Elias (1997) lists several British studies with inter-coder reliabilities between 70% and 78% with one exception from Slovenia as low as 56%, and an international review by Mannetje/Kromhout (2003) mentions reliabilities between 44% and 89%. Thus, the coding process introduces 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 we suggest here allows data collection at reduced costs.

Relevant for accurate coding at the most detailed level of the classification are verbatim answers that are tailored to the classification of interest and embody precise information about the job. The United Nations & International Labour Office (2010) therefore recommend asking two open-ended questions to collect sufficient details for coding according to the *International Standard Classification of Occupations 2008* (ISCO-08, International Labour Office, 2012). It is common practice in many surveys both to ask 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, 2014b). While 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. Even worse, this first response (typically the job title) may be contradictory to information collected afterwards (e.g., tasks and activities in the job) and therefore troubling coders who have been shown to disagree more often when more information is available (Cantor/Esposito, 1992; Conrad/Couper/Sakshaug, 2016). To overcome such difficulties, our proposed instrument is adaptive: We suggest asking only a single open-ended question and, by evaluating the answer, the interview software decides which question is asked next. It is our expectation that such a procedure can reduce measurement errors that are due to imperfect data collection at the interview.

Accepting the described challenges, we aim at three ultimate objectives with our adaptive questionnaire: (1) reducing coding errors that arise from missing or contradictory information provided by respondents; (2) maximizing the number of interview-coded answers to minimize efforts for coding the residual cases after the interview; (3) saving valuable interview time by automatic question selection to meet occupation-specific information demands.

The new instrument was tested in a telephone survey (CATI) and codes occupations according to the *German Classification of Occupations 2010* (KldB 2010, Bundesagentur für Arbeit, 2011a,b). The KldB 2010 is a detailed official classification and consists of 1,286 well-documented categories subsuming 24,000 job titles. Simultaneous coding according to the international classification ISCO-08 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.

Before we go into detail, we illustrate the proposed technique: Consider an exemplary respondent who is a “vice director and teacher” according to the first question about his job activities. Based on 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. Our algorithm combines a rule-based approach to automatic job classification and a supervised learning approach in which predictions are based on training data from the past. The job titles found to be most likely are then suggested in closed question format to the interviewer, who in turn asks the respondent to select the most adequate occupation. The suggestions for the “vice director and teacher” are shown in figure 1. Since we cannot guarantee that the algorithm always suggests an accurate job title, suggestions are amended by a last answer option “or do you work in a different occupation?” When this option is chosen, further questions should be asked to gather additional details about the person’s job; otherwise, coding is complete. For the “vice director and teacher”, the job title “Teacher - Elementary School” was selected, capturing a detail that was not provided in the original verbatim response.

When testing a new data collection technique, we want to find out how the instrument *would perform* if it were applied again and what *should be changed* to obtain even better results. *Past performance* from a first test can be useful in its own right, but it is mostly relevant to anticipate the performance in future application. This article intends to answer all three

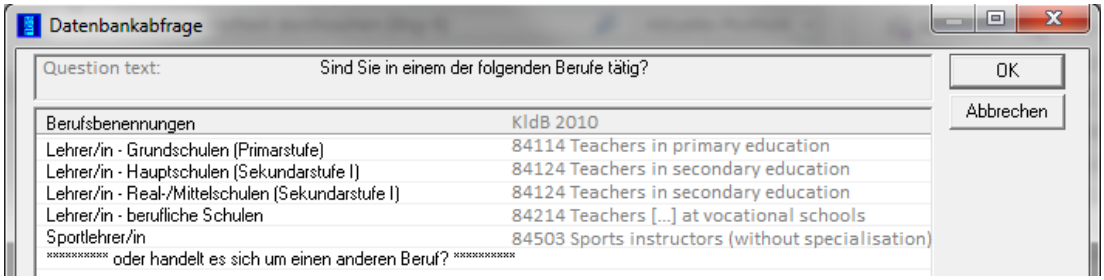

Figure 1: Screenshot from the interview for a “vice director and teacher”. Job titles in black font were suggested to the interviewer. The text in gray font was not shown during the interview and only added for this article to illustrate underlying categories from the German national classification (KldB 2010). Category titles are shown in abbreviated form. This example is discussed in the result section.

objectives. Section 2 (Related Literature) summarizes what other people did in the past to tackle related problems. Section 3 (Methods and Data) describes the mechanics behind the new instrument and how it was tested in practice. Section 4 (Results and Evaluation) focuses on the question which performance would be expected if the same instrument were applied in practice. In addition, we describe relevant weaknesses in order to find possible ways of improvement. Several ineffective features of the original instrument are only mentioned as a side note and the reader is referred to appendix A for a complete description of past performance. Section 5 concludes and compiles recommendations.

2 Related Literature

Geis/Hoffmeyer-Zlotnik (2000) provide an overview on occupation coding in Germany and its difficulties. According to them, scientific standards require that coding is carried out *systematically* and *reliably*. Consequently, Geis (2011) published a coding manual for the ISCO-classifications from 1968 and 1988 (both outdated now), which contains coding rules, conventions, and a case collection to achieve high replicability. However, replicability must not be confused with validity, and in fact coding procedures that are optimized to achieve high replicability carry the danger of introducing systematic biases. For example, one of Geis’ rules requires the coder to select from all plausible job categories the one which is least professionalized. As a consequence, this coding procedure will underestimate the degree of professionalization in the workforce. Paulus/Matthes (2013) provide shorter instructions for coding into the German national classification KldB 2010. Dictionaries are available for both classifications to automate the coding process: If the verbatim answer from the interview matches an entry in the dictionary, 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 replicability. Instead, two specific developments, which have been described in other countries, are relevant for us: (1) Post-interview coding is now sometimes done using machine learning algorithms and training data rather than simple matches of answers to pre-specified words from a dictionary. (2) Some researchers have described mechanisms for occupation coding during the interview to enable the respondent to specify her response more precisely, if needed. Our technique combines both

developments.

Several researchers have proposed computer systems to automate the coding process (see Speizer/Buckley 1998 for a review and Jung et al. 2008, Elias/Birch/Ellison 2014, Measure 2014, Thompson/Kornbau/Vesely 2014, and Javed et al. 2015 for some more recent developments). Software for computer-assisted coding suggests possible categories to the coder to make human work more efficient. Other algorithms known as "automated coding" assign categories all by themselves and without human supervision. More difficult residual cases are left for professional manual coding to keep the error level from automated coding below some pre-specified threshold. Conceptually, both systems are often based on coding rules and large databases that contain codes for frequent job titles (e.g., the prominent CASCOT program described by Elias/Birch/Ellison (2014) implements all of 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. In this way, verbatim answers that were coded before 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 U.S. Census. The algorithm 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/Hess/Geis (2006) ask for occupation with a sequence of three filter questions: the first one is about rough occupational groups, the second specifies the occupation further, and the final third question is about specific ISCO-88 categories. Tijdens (2014a, 2015) also avoids verbatim answers with a similar "search tree" for web surveys. A different strategy is applied for the job portal offered online at http://jobboerse.arbeitsagentur.de/ by the German Federal Employment Agency, which implements an algorithm to auto-complete textual input and helps job seekers to find job titles that exist in a database. These titles are linked to job categories from the KldB 2010. Hacking/Michiels/Janssen-Jansen (2006) and Svensson (2012) also mention occupation coding during the interview but their descriptions are rather rare in details.

Coding during the interview is not limited to occupation coding. Bobbitt/Carroll (1993), for example, test a system for coding "major field of study". They have implemented a fuzzy text search algorithm in a telephone survey that suggests possible codes, allowing interviewers to verify codes directly with the respondents.

3 Methods and Data

The new tool was tested in the survey *Selectivity Effects in Address Handling* commissioned by the Institute for Employment Research (IAB) and conducted by TNS Infratest Sozialforschung. In October and November 2014, TNS Infratest Sozialforschung conducted in total 1,208 valid CATI interviews (Computer-assisted Telephone Interview); 1,064 open verbatim answers for occupation were collected. The questionnaire covered several

topics related to the current occupation and work history, the use of social media for private and professional purposes, and volunteering activities, among others.

3.1 Sampling and Data Collection

A random sample of 17,001 persons - some of them with multiple addresses, others without phone numbers, which were to be researched - was drawn from a German federal database used in the social security administration (vom Berge/König/Seth, 2013). Since a primary purpose of the survey was to explore possible selectivity effects, a subsample of 10,000 persons was asked for consent to address transfer before only the consenters' addresses were transferred from IAB to the survey operator (c.f. Kreuter et al., 2015). Before fieldwork started, a letter of advice was sent out to 7,183 eligible respondents. All 67 interviewers, the local fieldwork managers, and the supervisors were trained by the central project management team of TNS Infratest Sozialforschung. Especially the new tool was an essential part of this training.

The sampling frame covers employees, unemployed persons, job seekers, recipients of unemployment benefit II, and participants in active labor market programs. Although they account for a large share of the German working population, persons who never paid contributions for social security insurance and have never received benefits from the German Federal Employment Agency are not included. This means specifically severe undercoverage of civil servants and self-employed persons. Strictly speaking, our results are only valid for occupations that do not belong to these groups; nevertheless, we see no reason why our conclusions should not generalize to all occupations.

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", number 6.3 in the questionnaire documented in appendix B). In very few cases (4.3% of the respondents), when the answer appears in a predefined list from TNS Infratest Sozialforschung containing overly general job titles (e.g., "salesman", "clerk"), another open-ended question is asked (no. 6.4). We also consider it helpful, although by no means necessary, for the productivity of our system to ask additional closed questions that are predictive for a person's job and common in many surveys (no. 6.2, 6.7-6.10, 6.13-6.18). Based on all these answers, the algorithm suggests possible job categories and the respondent can then select the most adequate one (no. 6.23b by default, but no. 6.23a, 6.24a, and 6.24b are related).

To compare the interview-coded category with professional manual coding, additional questions are asked in between (no. 4.2, 4.3, 6.5, 6.6, 6.11, 6.12 in appendix B). In particular, we follow the recommendations from the demographic standards (Statistisches Bundesamt, 2010) with some minor modifications and ask three open-ended questions (no. 6.3, 6.5, 6.6) to collect as many details as possible about the respondents' job for manual coding. Note that most questions mentioned above are primarily asked to evaluate the quality of interview coding in comparison with manual coding. If our suggested technique were

applied in practice, it would be sufficient to ask questions 6.3 and 6.23b only and skip the additional questions, saving valuable interview time. Only if an occupation cannot be coded during the interview, questions 6.5 and 6.6 would still need to be asked for ex-post manual coding.

3.3 Classification

What is saved when the interviewer clicks on an answer? Every suggested job title (shown on the left side in figure 1) corresponds to one category from the *Dokumentationskennziffer* (DKZ), an internal job classification that is used by the German Federal Employment Agency in its daily operations (see Paulus/Matthes 2013 for details). This classification subdivides the 1,286 categories from the German national classification KldB 2010 even further in 11,194 DKZ categories. Conversely, this means that every job title is linked to exactly one category in the KldB 2010. Thus, when a job title is selected during the interview, the DKZ-code is saved and a KldB 2010 code is automatically assigned as well. For illustrative purposes we include these associated KldB categories in gray font on the right side of figure 1. All evaluations provided below will be done on the scale of the KldB 2010, as this is the official and well-documented German national classification. The DKZ itself is only used as an auxiliary classification, which provides the job titles for our instrument, links these job titles to the KldB 2010, and makes available a large database of search words.

Many researchers do not use the national KldB 2010 but work with the *International Standard Classification of Occupations* (ISCO-08) instead. As this study explores technical possibilities, we only test our technology on the KldB 2010. However, it is worth noting that many - but not all - DKZ categories are linked to specific ISCO-categories, making it conceptually feasible to code in ISCO-08 and KldB 2010 at the same time during the interview. Because ISCO is with 436 categories only about one third the size of the KldB, we also expect improved quality evaluations if the analysis below were carried out for ISCO-08.

3.4 Prediction Algorithm

Possible job categories are predicted with a supervised learning algorithm, which learns from training data, i.e., from verbatim answers whose classification codes are already known from manual coding. Our training data comes from the survey *Working and Learning in a Changing World* (ALWA) ((Antoni et al., 2010; Drasch et al., 2012) documented the coding process). The ALWA survey questioned 9,227 persons about their employment biographies, i.e., all the jobs they have held during their lifetime, yielding a total of 32,887 job records. This is an exceptionally small number compared to other successful supervised learning applications for occupation coding: Thompson/Kornbau/Vesely (2014) test their algorithms with 1.5 million training observations and the studies by Javed et al. (2015) and Jung et al. (2008) have 2 million observations each. Due to the tiny size of our training data, it does not cover 433 out of 1,286 job categories from the KldB 2010, implying that these categories would never be suggested if the predictions were only based on this training data.

In principle, training data should be as large as possible to account for a high variety of possible verbatim inputs, including misspellings, and it should also provide for all contingencies how specific input texts can be coded into different categories. Such large training data were not available to us and, as a consequence, many results presented below may be improved with larger training data. In order to partly compensate for this loss in performance, we developed an algorithm that is tailored to small training data. Additionally, verbatim responses are matched with two job title databases, making it possible for the algorithm to recognize job titles that are not included in the training data.

The algorithm works as follows: Verbatim answers are standardized by removing some special characters and replacing letters with their uppercase equivalents. No spell-checking or further preprocessing takes place. The algorithm then searches for a meaningful subsequence of words from the standardized input text and chooses the subsequence of words that appears most often in a single category in the training data - which is what we call a phrase. In our example, the standardized input text is "VICE DIRECTOR AND TEACHER" and the derived phrase is "TEACHER". Respondents' standardized verbatim answers, the derived phrases, and their answers to closed questions are then matched with the training data and the job title databases. For each of a total of 26 different matching methods, scores are calculated on how much evidence exists in the training data respectively in the database for any job category c_j to be correct. For a given response l, we thus have scores $\theta_{lj}^{(m)}$ for $m = 1, ..., 26$ matching methods and for $j = 1, ..., 11194$ DKZ job categories. All the different scores are suspected to correlate with the true probability $P(c_j|l)$ that category c_j is correct for respondent l. The supervised learning problem is now to predict a binary target variable "c_j correct" from the 26 different scores obtained via matching. Gradient boosting as implemented by Hothorn et al. (2010) is used to predict the variable "c_j correct". Hereby, a sequence of decision trees is trained iteratively, each iteration focusing on examples that the previous ones got wrong. The final prediction is a sum over the different trees. Due to performance limitations, we choose 45 iterations and select the other tuning parameters (maximum tree size = 9, step size = 0.5) after exploratory bootstrap-type cross-validation. Our training data is hereby used twice: initially to predict the scores $\theta_{lj}^{(m)}$ from training data and subsequently to estimate the global boosting model. As LeBlanc/Tibshirani (1996) and Breiman (1996) note, this would make predictions biased. To avoid double usage, the prediction models for initial scores $\theta_{lj}^{(m)}$ do not use observation l for their estimations ("stacking"). Further details about the algorithm are provided in appendix C, which is based on work by Schierholz (2014).

At this point, the algorithm allows us to calculate a number of possible DKZ 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. It is desired to suggest job titles that cover a range of different KldB categories. For this purpose, we select (up to) five DKZ categories with the highest correctness probabilities under the condition that not more than two of the selected DKZ categories may belong to the same KldB category. Only if we cannot fill the five available spaces according to this rule, additional DKZ categories from the same KldB category are added with highest correctness probability first. Finally, the suggested categories are ordered by KldB and the answer option for

“other occupation” is appended.

3.5 Quality Analysis

The quality of our instrument is evaluated by comparing interview-coded answers with professional coding. The analysis proceeds in two steps: 1) manual coding and 2) 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 verbatim answers independent 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 different coding rules are in place to decide for a single code when answers are ambiguous. In addition, one of the coders provides a special indicator describing which verbatim answers have multiple possible codes. To safeguard the anonymity of the coders, the differences cannot be described in more detail.

Both codes differ frequently from the code assigned during the interview. In a second step, only these problematic codes are given to student assistants to check the correctness of the different coding procedures (interview coding and both professional codings). Both assistants worked independently from each other. They were equipped with the same source material as were the two coders (verbatim answers and additional answers from the interview; cf. appendix B) and with the codes from professional coding and interview coding. Their task was to categorize each coding decision in one of the following three categories:

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

3.6 Interviewer Behaviour

To understand to what extent interviewers apply standardized interviewing techniques for the proposed question, their behavior was analyzed (c.f. Ongena/Dijkstra 2016 for an overview on behavior coding). At the beginning of the interview, all respondents were asked if they allow recording of the dialogue, obtaining an 87.5% consent rate. Out of those respondents who answered the question of interest and whose audio recordings do not contain personal identifiers, 211 interviews were randomly selected for a detailed analysis. A professional coder from TNS Infratest Sozialforschung was instructed to code on the one hand for the question text and each answer option separately if the interviewer reads it aloud and on the other hand what the respondent says in a first reaction. The coding instructions, which were prepared by listening to several audio recordings, are provided in appendix E. Two audio files were dismissed, because the recordings only started after the question of interest when the interviewer reconnected to the respondent. In the course of the analysis, the first author listened to several recordings and felt reassured that the coder delivered high quality. Various interpretations 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: (1) productivity, (2) interview duration, and (3) quality. Next, we (4) mention some peculiarities of our instrument, which is best done in the context of the teacher-example from the introduction. Afterwards, we dig deeper (5) to understand how interviewers and respondents interacted, and (6) to determine the strengths and weaknesses of the prediction algorithm. The section closes with some (7) additional results. Throughout all descriptions, we highlight shortcomings in the tested system and note possible modifications in order to obtain even better results in a future instrument.

4.1 Productivity Analysis

Table 1 provides an overview of the productivity of our system. Out of 1064 persons who responded to the survey questions about occupation, the algorithm finds possible categories for 90.0%, leaving only 10.0% for which the algorithm does not suggest a single job category. This happens when the algorithm cannot relate the text entered by the interviewer to any previous input from training data or from data bases of job titles. Frequently, this is due to misspelled job titles and spell-checking algorithms could reduce this source of errors.

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

For another 13.6% of the respondents, the algorithm suggests possible job titles but the respondents do not find their own job in this list and declare that they have a different occupation instead. This is reasonable, as the algorithm is optimized to suggest adequate job

Table 1: 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 nonresponse	3		0.3%
..... Other experimental conditions	40		3.8%

titles, but it is impossible to guarantee that correct job categories are always suggested. In fact, the matching methods in our algorithm often find dozens or even hundreds of possible job titles. For usability reasons, we restrict the maximal number of suggested job titles to five. When filtering out the five best-suited job titles, it frequently occurs that 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, increasing the productivity of our system.

For respondents who answer that no job title is adequate and who indicate that they have an “other occupation” (applicable for 13.6%, as shown above), two additional lists are generated automatically and then suggested to them. The first one contains titles from the more general occupational sub-groups (4-digit from the KldB). The respondent can then select a sub-group or terminate the procedure by saying that no sub-group is adequate. When selecting a sub-group, DKZ job titles only from the chosen sub-group are suggested to the respondent. This demanding follow-up process was implemented because the algorithm usually finds dozens of possible job titles and, while 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 do not select an occupation during this process. In case they do, for 77% this interview-coded occupation is not in agreement with manual coding. Figure A1 and table A2 in the appendix contain additional details. We conclude that these follow-up questions yield unsatisfactory results and therefore are not worth to be included. If respondents select “other occupation”, responses should be referred to manual coding. They are thus excluded from the entire analysis that follows.

Table 1 also shows that three out of 1064 persons do not respond to the coding questions and the remaining 3.8% are due to the following experimental artefact: If the algorithm finds only a single job title or more than 250 possible job titles, job titles are not suggested within the regular closed question on occupation, but different question wordings are tested instead. Results are shown in tables A3 and A4 in the appendix. Both experimental conditions did not pay off 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 would be more expensive and are tiresome for the respondent. For respondents who select an occupation during the interview, our additional question takes 37 seconds at 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 seconds on average), the overall interview time 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 and the total interview time increases. The objective must therefore be to minimize the number of respondents who choose "other occupation" in order to make these calculations about interview duration more advantageous for coding during the interview.

4.3 Quality Analysis

Nearly three quarters of the respondents find an appropriate job title during the interview. While this is auspicious, the quality of the interview-coded categories is even more relevant. Two specific aspects of quality are analyzed: the agreement between and an evaluation of the different coding procedures. Both measures allow drawing conclusions about the quality.

Table 2 contains the inter-coder reliabilities for the professional coders (coder 1 & coder 2) and their respective agreement rates when compared to the codes from the interview. Agreement between 5-digit categories from the KldB 2010 is highest with 66.23% when coder 2 is compared to interview coding. Agreement between both professional coders is lowest. All agreement rates improve for broader classifications with fewer digits, but coder 2 and interview coding again have the highest agreement rates. An explanation for this might be that for some job descriptions a correct code is simple to find, while it is not for others. People with more difficult job descriptions are more hesitant to choose one of the DKZ job titles provided during the interview, which are less likely to be adequate. 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. In particular, this includes the more difficult job descriptions where professional coders 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 754 occupations that were also coded during the interview. The first number (61.11%) shows that coders disagree in almost 40% of the cases, leaving room for improvement. When comparing the different coding procedures, the second number (65.78%) is more informative. It ranges in between both agreement rates for interview coding, thus suggesting that the quality from interview coding is comparable to professional coding. The second step of our analysis will elucidate this further.

For 402 out of 754 (53.32%), the professional coders both agree with the respondent's own

Table 2: Agreement rates between 2 coding professionals (coder 1, coder 2) and interview coding (interview)

Agreement between	*First ... digits are in agreement (%)*					*No of codes*
	1	2	3	4	5	
Coder 1 & coder 2	87.20	79.40	74.98	67.56	61.11	1039
Coder 1 & interview	87.67	80.37	75.46	67.77	61.80	754
Coder 2 & interview	89.09	82.21	77.53	71.56	66.23	770

The KldB 2010 consists of 5-digit codes; aggregates for broader classifications with fewer digits are shown for convenience. The "No 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 2 occupations are "not codable".
- Interview coding provides codes for 770 occupations.

Quotes stem from the respective coding documentations.

choice. For these cases we can be highly certain that interview coding yields a code in a quality that is comparable to manual coding. More problematic are the 770-402=368 cases in which at least one human coder deviates from the code obtained via interview coding. Two student assistants were asked to check the correctness of all job codes for these 368 persons and the three coding procedures (coder 1, coder 2, and interview).

Table 3 contains the results from the coding evaluation. For the majority of the 368 problematic codes, both student assistants agree that the codes are acceptable. The professional coder 1 is rated best. 232 of his assignments are considered acceptable, which is significantly more than the 194 acceptable codes for the professional coder 2. Coder 1 also produced the lowest number of wrong codes (21) among the three different coding procedures. All other codes are located somewhere in between acceptable and wrong, with little agreement between the student assistants if these codes are acceptable, uncertain, or wrong. The comparison between coder 2 and interview coding slightly favors interview coding. The number of acceptable codes is almost identical (194 vs. 189) for both and differences are due to chance. This is not the case for wrong code assignments, whose number is significantly higher for coder 2 compared to interview coding (42 vs. 23). It may be concluded that interview coding brings minimal quality improvements when compared to coder 2 but fails to achieve the quality of coder 1. Note that the differences are rather small with respect to the 770 codes that entered this analysis, and for practical purposes, the observed differences may well be negligible.

To understand one major mechanism of how interview coding leads to wrong and uncertain codes, it is illustrating to go through a specific example: Consider a person who sells trucks. Our algorithm for coding during the interview is not intelligent enough to find 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 the category 62102 titled "Sales Occupations in Retail Trade (without Product Specialization)" which is the wrong code for this person's job. The point here is that job titles from the DKZ are not well-suited to support coding

Table 3: Contingency tables how the two student assistants evaluate the correctness of the three different coding procedures, cross-tabled over rows and columns

Correctness for coder 1

	Acceptable	*Uncertain*	*Wrong*	$\sum$
Acceptable	232	6	19	257
Uncertain	45	3	20	68
Wrong	19	3	21	43
$\sum$	296	12	60	368

Correctness for coder 2

	Acceptable	*Uncertain*	*Wrong*	$\sum$
Acceptable	194	8	23	225
Uncertain	35	7	20	62
Wrong	27	12	42	81
$\sum$	256	27	85	368

Correctness for interview coding

	Acceptable	*Uncertain*	*Wrong*	$\sum$
Acceptable	189	13	13	215
Uncertain	54	12	16	82
Wrong	33	15	23	71
$\sum$	276	40	52	368

during the interview. Many general job titles, such as “salesman”, exist in this classification. This generates the danger 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 DKZ 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 supposed to improve, but the proportion of interview-coded answers will probably decrease.

4.4 Example “vice director and teacher”

Occupation coding in general and interview coding in particular have several peculiarities that are worth discussing in some detail. This is best done in the context of an example. The “vice director and teacher”, who was introduced in figure 1, was chosen for illustrative purposes because he has several lessons to offer.

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. Audio-recording 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, making this answer a candidate for error because the interviewer failed to write down the complete information.

Two professional coders were asked to code the same verbatim answer. Both decided for the category 84194 “Managers in school of general education”. This category is the most adequate one from a post-interview coding perspective, for three reasons: (1) Additional questions from the interview show that this person oversees 14 employees, indicating that

managerial responsibilities may dominate his professional tasks as a teacher. This would favor the category 84194 because, according to the KldB 2010, the main focus of activities performed in the job is the criterion to decide for the best-suited category. (2) The alphabetic dictionary which is part of the KldB 2010 assigns “vice principal” to the code 84194. (3) The respondent answered “vice director” prior to “teacher”. Coding rules often dictate that the first job title is coded when multiple titles are provided in the verbatim response and the other titles do not specify the first title.

The school manager category 84194, which is preferred by both professional coders, is missing in the list of suggested job titles. Only if the respondent had had the chance to choose this category, one would know if he preferred this category or the one now 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 via any database to the category 84194. The word “director”, however, could in principle be linked via some databases to the desired category (and to many more managerial categories), but the text matching methods we applied to those databases do not work if any word is inserted in addition to the key term. 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 one can never guarantee that all plausible categories are suggested to the respondent.

While one relevant category is missing in figure 1, other suggested job titles are less relevant: The “sports teacher” is clearly implausible and the job titles “teacher - Hauptschulen” and “teacher - Real-/Mittelschulen” are repetitive. Both are associated with a single KldB category (84124), allowing respondents a detailed choice between two different school types. Yet, the KldB classification does not distinguish between both and it would be sufficient to ask for a single overarching category “secondary school teacher” instead (nonexistent in the DKZ). Because we restricted the number of shown job titles to a maximum of five, such a reduction of answer options would create room for other plausible categories.

4.5 Interviewer Behavior

Our new technique was tested in a telephone survey. Compared to self-administered surveys in which one could confront the respondents directly with the suggested answer options, 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 should read questions and answers exactly as worded and respondents should select the most adequate answer without any help from the interviewer. This general training was not repeated for our particular survey. Because interviewers frequently violate these guidelines for the proposed question on occupation, it is relevant to describe how the interview-coded occupations are obtained.

Directly 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

fill the gap by explaining with their own words what comes next. 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%), the question text is not read.

Frequently, job titles are automatically suggested although they are definitely not plausible. In the "vice director" example from above (c.f. figure 1), the interviewer knows from the preceding conversation that the list of suggestions contains only a single job title that is appropriate in her view. Not reading implausible suggestions saves time and prevents possible irritation by the respondent. This makes it attractive for interviewers to skip implausible job suggestions. In 97 out of 209 interviews (46%), at least one suggested job title is not read. This happens for 10 interviews (10%) because the algorithm finds a job title that is identical with the verbatim answer provided by the respondent before, for 35 interviews (36%) because suggested job titles are definitely implausible, and for additional 23 interviews (24%) both reasons apply. Some interviewers guide respondents to a specific answer: In 27 out of 209 interviews (13%), the interviewer mentions only a single job title, typically formulated in the form of a question (e.g., "We have here .... Is this correct?"), but sometimes also formulated as a statement, so that the respondent is not required to confirm this job title. In 8 interviews (4%), the interviewers do not read aloud a single job title but decide all by themselves for the most adequate answer option.

It is also very common for interviewers to skip the answer option "other occupation", which is given for 37 out of 209 respondents (18%) only. Interviewers may have skipped this option because it is highlighted in the interview software or because they think that an appropriate job title was already found.

Every question should usually be followed by an appropriate answer from the respondent. In a first reaction, 156 out of 209 respondents (75%) provide such an answer, either interrupting the interviewer (21 persons) or naming it after the interviewer has finished reading the question (135 persons). Normally, this answer marks the end of the occupation coding process unless the respondent chooses "other occupation" or the interviewer starts arguing with the respondent about a more adequate occupation, as we have observed in a few interviews. More problematic are cases in which the respondents do not give an appropriate answer in their first reaction: When no job title is adequate at first sight, respondents are clueless what to answer. This confusion leads 17 respondents (8%) to mention additional details about their jobs and as a result "other occupation" is most often selected. Other 18 respondents (9%) are confused or ask the interviewer to repeat or to explain the suggested job title. 14 out of the 18 respondents eventually agree with one of the suggestions. In 18 additional interviews (9%), the respondents mostly remain silent because the interviewer thinks aloud or in silence without asking a question and it is then typically the interviewer who decides for the most adequate answer option.

In summary, our exercise in behavior coding shows that many interviewers did not consequently follow the rules for standardized interviews. It is the exception that an interviewer reads 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 herself without asking the respondent, or engages in a discussion with the respondent about the most adequate

answer option, one might worry that interviewer effects can be large for this question. However, one should not exaggerate these problems: Many skipped job titles are definitely inadequate; typically respondents and not interviewers make the decision, and, as interviewers often have a good understanding about the respondent's job, it is not clear if data quality is diminished when interviewers play an overly active part. 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: Matching Success

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., verbatim answers) as possible. Furthermore, these categories should be of high quality so that the respondents find their own jobs in the suggested list. In the following, we analyze how well the algorithm currently does regarding both objectives and search for possible ways of improvement.

Any algorithm must match the verbatim responses given by respondents with some database containing possible categories. In order to find possible job categories for a maximal number of respondents, we apply three different databases: Our training data consists 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 KldB 2010. 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 486, 495, and 434 of the 1,064 respondents who answered the verbatim 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 persons are simple to code, either during or after the interview. In our sample, 358 respondents provide answers that allow 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 4. When the verbatim answer is not required to be identical with but only needs to be a substring of a database record, more matches are found (524 vs. 486 and 551 vs. 495), but the gains are relatively small. This happens because this matching technique is only appropriate for short answers. 349 respondents, however, provide 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.

Table 4: Descriptive results for various matching methods and databases

Matching method	*(1)*	*(2)*	*(3)*		
			Median	*Mean*	*Maximum*
Answer matches with training data					
- Identical	486	425	2	4.2	45
- Answer is substring	524	459	4	8.0	122
Answer matches with file of search words					
- Identical	495	422	2	3.8	66
- Answer is substring	551	499	5	12.6	187
Answer matches			KldB/DKZ	KldB/DKZ	KldB/DKZ
with alphabetic dictionary	434	414	2/23	4.8/71.3	69/1012
Phrase matches with training data (*)					
- Identical	786	606	3	7.0	45
- Answer is substring	874	743	8	57.3	1479
Phrase matches with file of search words					
- Identical	760	556	3	6.8	82
- Answer is substring	891	771	12	133.6	3878
Phrase matches			KldB/DKZ	KldB/DKZ	KldB/DKZ
with alphabetic dictionary	609	556	2/30	7.5/94.7	96/1190

(1) Number of respondents for whom the matching method suggests at least one category

(2) Number of respondents for whom at least one suggested category was also coded by at least one professional coder

(3) Average number of categories, provided that at least one category is suggested

(*) These matching methods were not included in the production software

(KldB/DKZ) The alphabetic dictionary links job titles only to categories from the KldB 2010. All DKZ categories that are associated with the so found KldB-categories are possible candidates for suggestion. We thus provide the number of KldB-suggestions first and the number of DKZ-suggestions second.

The second inexact matching method is more promising for longer answers: when searching for a meaningful subsequence of words in the original verbatim answer - here called a phrase -, 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 4.

Column (2) “Number of respondents for whom at least one suggested category was also coded by at least one professional coder” confirms that we find meaningful matches with all methods. For most respondents and any matching method, categories are suggested that are relevant in the sense that professional coders usually select one of the suggested categories independently. This is not self-evident because especially for the phrase-matching methods it does actually happen that the phrase itself is meaningless for coding (e.g., words like “in” or “and”) and matching such a phrase brings certainly no improvement.

The downside of inexact matching is summarized in the column (3) “Average number of categories when at least one category is suggested”, which shows some properties about the number of suggested categories provided that 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 accurate 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 suggested categories that can stem from 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 tradeoff with each additional matching method: on the one hand, adding a matching method opens up the possibility for additional categories that are suggested to respondents. On the other hand, suggesting more categories could also mean suggesting more unsuitable categories, which may prolong the interview duration, induce more people to indicate “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 5 contains the hypothetical results if the algorithm were changed, i.e., what would have happened if these conditions were applied in the field. The conditions are as follows:

1. Answers have identical matches in both the training data and in the alphabetic dictionary.
2. No shorter phrase is found. This condition includes all cases from the first condition with only two exceptions.
3. 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 5 shows that the number of respondents who receive job category suggestions increases when the conditions are loosened, allowing more respondents to code themselves 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 to keep 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 3 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 an impressive 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 3 select “other occupation”, as it was implemented in the tested system.

Table 5: Productivity of the coding system under various hypothetical situations

Ask first inquiry only if ...	*(1)*	*(2)*	*(2)/(1)*	*(3)*	*(3)/(1-2)*
Condition 1: ... identical match with training data and match with alphabetic dictionary	386	12	3.1%	312	83.4%
Condition 2: ... no shorter phrase is found	532	27	5.1%	416	82.4%
Condition 3: ... no shorter phrase is found or phrase matches with alphabetic dictionary	712	60	8.4%	511	78.4%
Condition 4: always (this was actually done in this study)	915	145	15.8%	574	74.5%

(1) Number of respondents who would be asked under the given condition

(2) Number of respondents who answer “other occupation” under the given condition

(2)/(1) Column (2) divided by column (1)

(3) Number of respondents under the given condition who select a code that is in agreement with at least one professional coder.

(3)/(1-2) Column (3) divided by the difference between columns (1) and (2)

This result also has implications for our algorithm. Job category suggestions are satisfactory when verbatim answers are short and can be matched by identical or substring matching to any database. The predictions are still accurate enough if the algorithm can extract a phrase from a multi-worded verbatim answer that is a job title from the alphabetic dictionary. The remaining verbatim answers require more attention to further improve the algorithm. They may be characterized as follows: For 203 verbatims the algorithm finds a shorter phrase that is not listed in the alphabetic dictionary. These answers are at least two words long - often longer - and frequently contain no single job title but more than one word that is 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 are usually one word long. Spelling errors and compound words are frequent reasons why matching is not possible. Future improvements of the algorithm should address these problems.

4.7 Additional results

Two additional features were part of the test software, but the results are discouraging and both features are not recommended for future use: (1) We tried if an additional answer option “similar occupation” makes it less likely for respondents to select one of the suggested job titles. (2) We tested if interviewers can detect from their observations how accurate the selected job title is. For completeness we describe the techniques and results in appendix A (tables A5 and A6).

Apart from the analysis of coding during the interview, another result is informative. A classical strategy for automatic occupation coding is to search for a given job title in a database and assign the associated category accordingly. We matched the first verbatim answer from the interview (number 6.3 or 6.4 in appendix B) to a database we prepared from the alphabetic dictionary of 24,000 job titles that is part of the KldB 2010. Although we only matched job titles for this analysis if they were clearly associated to a single category,

successful exact database matches were found for 418 out of 1064 verbatim responses. For these persons it was then possible to compare the codes with those obtained from manual coding (coder 1 & coder 2), with the following results: for 307 responses (73.4%) all three codes are identical, for 88 responses (21.1%) only one manual coder agrees with the code from the database, and for 23 responses (5.6%) both disagree with the database. These numbers show that a substantial proportion of respondents mention job titles that can be coded automatically in some category with the alphabetic dictionary, while this does not mean that these categories are the only possible ones. 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 - one cannot hope for an optimal coding quality which guarantees every respondent to be classified into the category that describes her occupational tasks and duties best.

Another error source that leads to low inter-coder reliabilities can be found in both manual and interview coding. Coders are usually required to select a single correct category and multiple categories are not permitted, even if plausible. The decision for a single category can be hard, 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 agree 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 Conclusion

Traditional coding of occupations is costly and time-consuming. In our study, two independent coders obtain a reliability of 61.11%, a number that is low but by no means an exception. We described and tested a technical solution with increased interaction during the interview to counter these challenges. After a verbatim answer is entered in the interview software, the computer automatically calculates a small set of possible job categories and suggests them to the respondent, who can in turn select the most adequate one. 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 is available directly after the interview.

The quality was compared to the work of two professional coders. It is slightly lower than

the quality of the first coder and comparable to the quality of the second one. 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 different rules were followed by the coders. Our desire to increase the quality by collecting more information already during the interview was not fulfilled. This has several reasons: Categories that are suggested by the algorithm are sometimes inadequate, the two generated follow-up questions are unsuited to elicit more adequate 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 will be shortened by a few seconds; others who do not select one of the suggested categories will have to bear the burden of slightly longer interviews with an additional question, which does not produce relevant data. 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 largest 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 purpose, 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% and the proportion of respondents with prolonged interviews from 13.6% to 5.6%, is recommended in particular.

These results are satisfactory for the first trial of a complex instrument. Although several features, which did not meet our expectations at all, were tested in the current survey, we are confident that obvious adaptions - which would be required for regular usage - would not change our key results substantially. In addition, it would be useful to estimate if the proposed instrument would lead to cost reductions in the coding process. At the very least, our results show that coding during the interview, in future, can become a viable technique that may partly replace traditional post-interview coding. For future developments, we have identified a number of factors how to improve the process.

5.1 Recommendations

When respondents choose one of the suggested job titles, it is too often not the most adequate one. Respondents frequently select general job titles that are not wrong but link to suboptimal KldB categories. These inadequate job titles stem from the DKZ, which is therefore not well-suited for coding during the interview. In order 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 (KldB 2010) and international (ISCO-08), for simultaneous coding.

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 to be expected that the productivity of the system can be further increased. In the frequent

situation that a verbatim 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.

Interviewers frequently did not act according to the rules of standardized interviews at the proposed question but often preferred rewording the question text and skipping suggested answer options instead. While this behavior leads to concerns about interviewer effects, one must not forget the positive impact: Respondents are less hassled with strange answer options and the duration of the interview is shortened when implausible answer options are skipped. 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 implausible job categories. This procedure could partly remedy the current problem that the algorithm finds many possible job categories, but for $\sim 36\%$ of the respondents, a relevant job title is missing in the subset which is provided to the interviewer. Furthermore, extended interviewer training will be necessary to ensure that interviewers know when they have to follow the script and to reduce the risk that they skip relevant answer options.

Some answers in reply to the first open-ended question about occupational activities are very general and one would need to suggest a huge number of possible categories. Our vision is instead to recognize these general answers automatically. An additional open-ended question would then be asked to collect more details and use this second answer as input for coding during the interview.

Additionally, future research should consider the possibility that more than one job category may be adequate.

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

References

Antoni, Manfred; Drasch, Katrin; Kleinert, Corinna; Matthes, Britta; Ruland, Michael; Trahms, Annette (2010): Arbeiten und Lernen im Wandel * Teil 1: Überblick über die Studie. FDZ-Methodenreport 05/2010, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Bobbitt, Larry G.; Carroll, C. Dennis (1993): Coding Major Field of Study. In: Proceedings of the Survey Research Methods Section: American Statistical Association, p. 177–182.

Breiman, Leo (1996): Stacked Regressions. In: Machine Learning, Vol. 24, No. 1, p. 49–64.

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

Bundesagentur für Arbeit (2011b): Klassifikation der Berufe 2010. Band 2: Definitorischer und beschreibender Teil. Bundesagentur für Arbeit, Nuremberg.

Campanelli, Pamela; Thomson, Katarina; Moon, Nick; Staples, Tessa (1997): The Quality of Occupational Coding in the United Kingdom. In: Lyberg, Lars; Biemer, Paul; Collins, Martin; DeLeeuw, Edith; Dippo, Cathryn; Schwarz, Norbert; Trewin, Dennis (Eds.) Survey Measurement and Process Quality, New York: John Wiley & Sons, Inc., p. 437–453.

Cantor, David; Esposito, James (1992): Evaluating Interviewer Style for Collecting Industry and Occupation Information. In: Proceedings of the Survey Research Methods Section: American Statistical Association, p. 661–666.

Conrad, Frederick G.; Couper, Mick P.; Sakshaug, Joseph W. (2016): Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes. In: Journal of Official Statistics, Vol. 32, No. 1, p. 75–92.

Creecy, Robert H.; Masand, Brij M.; Smith, Stephen J.; Waltz, David L. (1992): Trading MIPS and Memory for Knowledge Engineering. In: Commun. ACM, Vol. 35, No. 8, p. 48–64.

Dowle, Matt; Short, T; Lianoglou, S (2012): data.table: Extension of data.frame for Fast Indexing, Fast Ordered Joins, Fast Assignment, Fast Grouping and List Columns. URL `https://cran.r-project.org/package=data.table`, r package version 1.8.6.

Drasch, Katrin; Matthes, Britta; Munz, Manuel; Paulus, Wiebke; Valentin, Margot-Anna (2012): Arbeiten und Lernen im Wandel * Teil V: Die Codierung der offenen Angaben zur beruflichen Tätigkeit, Ausbildung und Branche. FDZ-Methodenreport 04/2012, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Elias, Peter (1997): Occupational Classification (ISCO-88): Concepts, Methods, Reliability, Validity and Cross-National Comparability. OECD Labour Market and Social Policy Occasional Papers 20, OECD Publishing, Paris.

Elias, Peter; Birch, Margaret; Ellison, Ritva (2014): CASCOT International version 5 * User Guide. Institute for Employment Research at the University of Warwick, Coventry, URL http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/internat/, accessed: 2016-04-08.

Feinerer, Ingo; Hornik, Kurt; Meyer, David (2008): Text Mining Infrastructure in R. In: Journal of Statistical Software, Vol. 25, No. 1, p. 1–54.

Geis, Alfons (2011): Handbuch der Berufsvercodung. GESIS. Survey Design and Methodology, Mannheim.

Geis, Alfons; Hoffmeyer-Zlotnik, Jürgen H.P. (2000): Stand der Berufsvercodung. In: ZUMA-Nachrichten, Vol. 24, No. 47, p. 103–128.

Hacking, Wim; Michiels, John; Janssen-Jansen, Saskia (2006): Computer Assisted Coding by Interviewers. In: Proceedings of the IBUC 2006 10th International Blaise Users Conference, p. 283–296.

Hoffmeyer-Zlotnik, Jürgen H.P.; Hess, Doris; Geis, Alfons J. (2006): Computerunterstützte Vercodung der International Standard Classification of Occupations (ISCO-88): Vorstellen eines Instruments. In: ZUMA-Nachrichten, Vol. 30, No. 58, p. 101–113.

Hoffmeyer-Zlotnik, Jürgen H.P.; Warner, Uwe (2012): Harmonisierung demographischer und sozio-ökonomischer Variablen: Instrumente für die international vergleichende Surveyforschung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hothorn, Torsten; Bühlmann, Peter; Kneib, Thomas; Schmid, Matthias; Hofner, Benjamin (2010): Model-based Boosting 2.0. In: Journal of Machine Learning Research, Vol. 11, p. 2109–2113.

International Labour Office (2012): International Standard Classification of Occupations: ISCO-08. International Labour Organization, Geneva, URL http://labordoc.ilo.org/record/441501?ln=en.

Javed, Faizan; Luo, Qinlong; McNair, Matt; Jacob, Ferosh; Zhao, Meng; Kang, Tae Seung Kang (2015): Carotene: A Job Title Classification System for the Online Recruitment Domain. In: Proceedings of the First IEEE International Conference on Big Data Computing Service and Applications, Redmond City, p. 286–293.

Jung, Yuchul; Yoo, Jihee; Myaeng, Sung-Hyon; Han, Dong-Cheol (2008): A Web-Based Automated System for Industry and Occupation Coding. In: Bailey, James; Maier, David; Schewe, Klaus-Dieter; Thalheim, Bernhard; Wang, XiaoyangSean (Eds.) Web Information Systems Engineering - WISE 2008, Vol. 5175 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, p. 443–457.

Kreuter, Frauke; Sakshaug, Joe; Schmucker, Alexandra; Couper, Mick; Singer, Eleanor (2015): Privacy, Data Linkage, and Informed Consent. Presentation at the 70th Annual Conference of the American Association for Public Opinion Research.

LeBlanc, Michael; Tibshirani, Robert (1996): Combining Estimates in Regression and Classification. In: Journal of the American Statistical Association, Vol. 91, No. 436, p. 1641–1650.

Loos, Christiane; Eisenmenger, Matthias; Bretschi, David (2013): Das Verfahren der Berufskodierung im Zensus 2011. In: Wirtschaft und Statistik, p. 173–184.

Mannetje, Andrea 't; Kromhout, Hans (2003): The Use of Occupation and Industry Classifications in General Population Studies. In: International Journal of Epidemiology, Vol. 32, No. 3, p. 419–428.

Measure, Alexander (2014): Automated Coding of Worker Injury Narratives. In: Proceedings of the Government Statistics Section, American Statistical Association, p. 2124–2133.

NIPO Software (2014): NIPO Fieldwork System. NIPO Software, Amsterdam.

Office for National Statistics (2003): Quality of Data Capture and Coding: Evaluation Report. Census 2001 Review and Evaluation, URL 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, accessed: 2016-04-08.

Ongena, Yfke P.; Dijkstra, Wil (2016): Methods of Behavior Coding of Survey Interviews. In: Journal of Official Statistics, Vol. 22, No. 3, p. 419–451.

Oracle Corporation (2014): MySQL. Oracle Corporation, Redwood City.

Paulus, Wiebke; Matthes, Britta (2013): Klassifikation der Berufe * Struktur, Codierung und Umsteigeschlüssel. FDZ-Methodenreport 08/2013, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

R Core Team (2014): foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... URL https://CRAN.R-project.org/package=foreign, r package version 0.8-66.

R Core Team (2012): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, ISBN 3-900051-07-0.

Ripley, Brian; Lapsley, Michael (2013): RODBC: ODBC Database Access. URL https://CRAN.R-project.org/package=RODBC, r package version 1.3-10.

Schierholz, Malte (2014): Automating Survey Coding for Occupation. FDZ-Methodenreport 10/2014, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Speizer, Howard; Buckley, Paul (1998): Automated Coding of Survey Data. In: Couper, Mick P.; Baker, Reginald P.; Bethlehem, Jelke; Clark, Cynthia Z. F.; Martin, Jean; Nicholls II, William L.; O'Reilly, James M. (Eds.) Computer Assisted Survey Information Collection, New York: John Wiley & Sons, Inc., p. 223–243.

Statistisches Bundesamt (2010): Demographische Standards. Statistisches Bundesamt, Wiesbaden.

Svensson, Jörgen (2012): Quality Control of Coding of Survey Responses at Statistics Sweden. In: Proceedings of the European Conference on Quality in Official Statistics - Q2012.

Thompson, Matthew; Kornbau, Michael E.; Vesely, Julie (2014): Creating an Automated Industry and Occupation Coding Process for the American Community Survey. Background Material for a meeting of the Federal Economic Statistics Advisory Committee, URL http://www.census.gov/about/adrm/fesac/meetings/june-13-2014-meeting.html, accessed: 2016-04-08.

Tijdens, Kea (2015): Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-up Tables. In: Survey Insights: Methods from the Field, URL http://doi.org/10.13094/SMIF-2015-00008.

Tijdens, Kea (2014a): Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey. In: Journal of Official Statistics, Vol. 30, No. 1, p. 23–43.

Tijdens, Kea (2014b): Reviewing the Measurement and Comparison of Occupations Across Europe. AIAS Working Paper 149, Universiteit van Amsterdam.

Trappmann, Mark; Beste, Jonas; Bethmann, Arne; Müller, Gerrit (2013): The PASS Panel Survey After Six Waves. In: Journal for Labour Market Research, Vol. 46, No. 4, p. 275–281.

United Nations & International Labour Office (2010): Measuring the Economically Active in Population Censuses: A Handbook. United Nations & International Labour Office, New York.

Urbanek, Simon (2013): Rserve: Binary R server. Urbanek, Simon, URL http://CRAN.R-project.org/package=Rserve, r package version 1.7-3.

vom Berge, Philipp; König, Marion; Seth, Stefan (2013): Sample of Integrated Labour Market Biographies (SIAB) 1975-2010. FDZ-Datenreport 01/2013, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nuremberg.

Wickham, Hadley (2015): stringr: Simple, Consistent Wrappers for Common String Operations. URL https://CRAN.R-project.org/package=stringr, r package version 1.0.0.

Recently published

No.	Author(s)	Title	Date
2/2016	Van den Berg, G. Hofmann, B. Uhlendorff, A.	The role of sickness in the evaluation of job search assistance and sanctions	1/16
3/2016	Bossler, M.	Employment expectations and uncertainties ahead of the new German minimum wage	2/16
4/2016	Fuchs, J. Kubis, A. Schneider, L.	Replacement migration from a labour market perspective: Germany's long-term potential labour force and immigration from non-EU member countries	2/16
5/2016	Garloff, A. Wapler, R..	Labour shortages and replacement demand in Germany: The (non)-consequences of demographic change	2/16
6/2016	Garloff, A. Roth, D.	Regional age structure and young workers' wages	2/16
7/2016	Stockinger, B. Wolf, K.	The productivity effects of worker mobility between heterogeneous firms	2/16
8/2016	Bruckmeier, K. Wiemers, J.	Differences in welfare take-up between immigrants and natives: A microsimulation study	3/16
9/2016	Möller, J.	Lohnungleichheit - Gibt es eine Trendwende?	3/16
10/2016	Bossler, M. Gerner, H.-D.	Employment effects of the new German minimum wage: Evidence from establishment-level micro data	3/16
11/2016	Bossler, M. Grunau, P.	Asymmetric information in external versus internal promotions	4/16
12/2016	Mendolicchio, C. Pietra, T.	A re-examination of constrained Pareto inefficiency in economies with incomplete markets	4/16
13/2016	Hamann, S. Niebuhr, A. Peters, C.	Benefits of dense labour markets: Evidence from transitions to employment in Germany	4/16
14/2016	Bender, S. Bloom, N. Card, D. Van Reenen, J. Wolter, S.	Management practices, workforce selection, and productivity	5/16
15/2016	Bossler, M. Broszeit, S.	Do minimum wages increase job satisfaction? Micro data evidence from the new German minimum wage	5/16
16/2016	Dengler, K. Stops, M. Vicari, B.	Occupation-specific matching efficiency	5/16

As per: 2016-05-17

For a full list, consult the IAB website
http://www.iab.de/de/publikationen/discussionpaper.aspx

Imprint

IAB-Discussion Paper 17/2016
18 May 2016

Editorial address
Institute for Employment Research
of the Federal Employment Agency
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Ricardo Martinez Moya, Jutta Palm-Nowak

Technical completion
Gertrud Steele

Website
http://www.iab.de

Download of this Discussion Paper
http://doku.iab.de/discussionpapers/2016/dp1716.pdf

ISSN 2195-2663

For further inquiries contact the author:

Malte Schierholz
Phone +49.911.179 6022
E-mail malte.schierholz@iab.de