Gesche Brandt / Susanne de Vogel / Steffen Jaksztat / Carola Teichmann / Kerstin Lange / Percy Scheller / Sandra Vietgen

DZHW PhD Panel 2014

Data and Methods Report on the PhD Panel 2014 (1st and 2nd Survey Waves)

Data and Methods Report

May 2018

fdz.DZHW
Forschungsdatenzentrum
für Hochschul- und Wissenschaftsforschung

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Author(s):
Gesche Brandt
Susanne de Vogel
Steffen Jaksztat
Carola Teichmann
Kerstin Lange
Percy Scheller
Sandra Vietgen

With assistance from:
Kolja Briedis

Editor:
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW)
German Centre for Higher Education Research and Science Studies (DZHW)
Lange Laube 12 | 30159 Hannover | Germany | www.dzhw.eu
Tel.: +49 511 450670-0 | Fax: +49 511 450670-960 | info@dzhw.eu

Table of Contents

Table of Figures

List of Tables

I Introduction

The DZHW PhD Panel with the project title, 'Careers of PhD Holders', is a panel study of the conditions during the PhD/doctorate phase, the transition to the workplace after graduation and the further professional careers of the graduates.[1] It is conducted by the German Centre for Higher Education Research and Science Studies (DZHW)[2], financed by the Federal Ministry of Education and Research (BMBF) and serves – in addition to the official university statistics – the national educational monitoring.

Within the scope of the BMBF funded project on the development of a research data centre for higher education research and science studies at the DZHW (RDC-DZHW), the data are subsequently edited and documented for the purpose of later use in cooperation with the project staff of the PhD Panel 2014. Using various methods of access, they are made available as *Scientific Use Files* (SUF) for secondary scientific use and as *Campus Use Files* (CUF) for teaching and training purposes. Along with the survey datasets, documentation material on the datasets and the carrying out of the studies are provided.[3]

This data and methods report is part of the documentation for the first and second survey waves of the PhD Panel 2014 (doi: 10.21249/DZHW:phd2014:2.0.0).[4] Further documentation material on the study (dataset reports, questionnaires, question flow diagrams etc.) can be freely downloaded from the RDC-DZHW website (https://metadata.fdz.dzhw.eu). Section II of the report presents an overview of the key data from PhD Panel 2014. The key information on the use of the data in this study follows in section II. Chapter 1 presents content and structure of the PhD Panel in general. The additional structure of the report focuses essentially on the research process. The survey instruments employed are described in chapter 2 and the survey process (sampling procedure, survey procedure, response, data processing) in chapters 3 to 6. A description of the weighting and anonymisation performed follows in chapters 7 and 8.

[1] Latest information on the DZHW PhD Panel can be found by visiting the project website (www.promoviertenpanel.de).

[2] The German Centre for Higher Education Research and Science Studies (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH [DZHW, http://www.dzhw.eu]) was formed in August 2013 through a spin-off of the company HIS Hochschul-Informations-System GmbH.

[3] Information on the available datasets and documentations is provided on the RDC-DZHW website (https://fdz.dzhw.eu).

[4] The data from the third wave, which was carried out in 2017, as well as the data from the two further planned waves cannot be published until a later date.

II Overview of the DZHW PhD Panel 2014

Survey	DZHW PhD Panel
Cohort	PhD graduates 2013/2014
Surveying Institution	German Centre for Higher Education Research and Science Studies (DZHW)
Funding	Federal Ministry of Education and Research (BMBF)
Project Contributors (Project Leader)	Gesche Brandt, Kolja Briedis, Susanne de Vogel, Steffen Jaksztat, Carola Teichmann
Themes	Educational and professional biography prior to the PhD Career path since PhD graduation Doctorate conditions Doctorate results Scientific and professional further qualification Academic activities Mobility experiences Aims, motives and personality Social background
Survey Design	Panel Design
Population	Persons that have completed a PhD in the winter semester of 2013-14 or in the summer semester of 2014 at a university with the right to award doctorates in the Federal Republic of Germany
Samples	Full census
Survey Method	1st wave: standardised self-administered survey 2nd wave: standardised online survey
Survey Time Period	1st wave: December 15, 2014 to February 17, 2016 2nd wave: March 15, 2016 to April 18, 2016
Number of Cases (Data Set)	1. wave: n = 5.410 2. wave: n =3.184
Response Rate	1. wave: 27,2 % 2. wave: 66,1 %
Data Products and Mode of Access	CUF: Download SUF: Download, Remote-Desktop, On-Site
Data Set Structures	Individual data in wide format Spell data in long format

DOI	10.21249/DZHW:phd2014:2.0.0
Further Information	https://fdz.dzhw.eu

Project Publikations*

Brandt, G., de Vogel, S., & Jaksztat, S. (2016). Entwicklung und Testung eines Instruments zur Erfassung der Lernumwelt in der Promotionsphase. Ergebnisse der Entwicklungsstudie. Werkstattbericht. DZHW: Hannover.

* All project publiations are available for download on the project website (www.dzhw.eu/promovierte).

III Data Use Instructions

[Data Use Requirements] Data from the PhD Panel 2014 are anonymised and made available by the RDC in accordance with Federal Data Protection Law (cf. § 40 paras. 1 and 2 BDSG) exclusively for scientific research purposes.[5] The RDC provides *Scientific Use Files* (SUF) for scientific secondary use and a *Campus Use Files* (CUF) for teaching and exercise purposes.

Requirements for the use of a SUF are an employment at a scientific institution and the conclusion of a data use agreement. Before the conclusion of a data use agreement, the RDC verifies the presence of a scientific use purpose. Students or doctoral students without an employment at a scientific institution must be able to prove cooperation with a supervisory employee of a scientific institution. A form for the data use contract can be downloaded from the RDC website. In order to use a CUF a registration with the name and the purpose of use has to be undertaken. Afterwards the CUF will be transmitted by the RDC.

[Data Access] The CUF of the PhD Panel 2014 can be used at the local computer. The SUF is provided using three modes of access, which differ in their restrictions with respect to storage location, the opportunity for autonomous access to external data and RDC control options for restrictive data. These methods include:

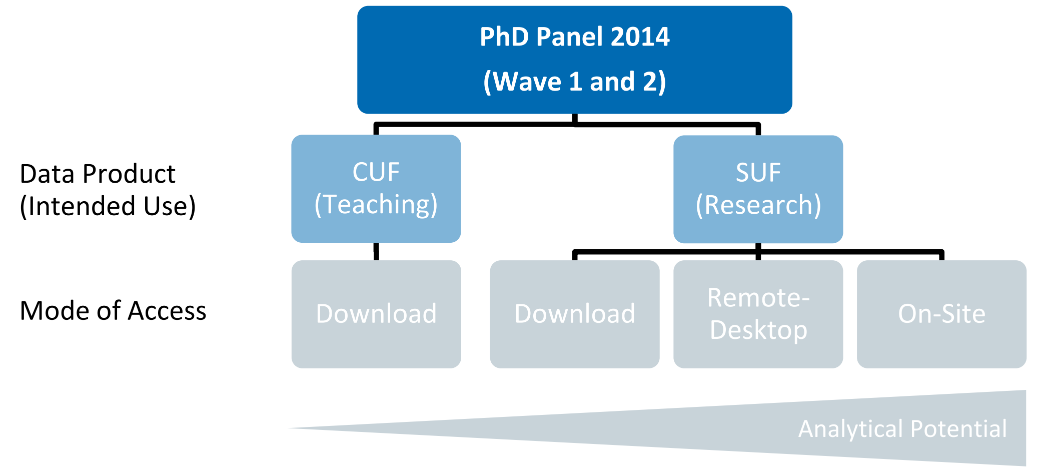
- **Download:** Data will be sent via a secure email connection or are available for download from the RDC website. Users can save the data on their local computer to link with data from external sources as well as perform analysis using their own software.
- **Remote Desktop:** Data are available on a RDC terminal server. Using a secure connection between the user's local computer and the RDC terminal server, the data can be analysed using the software on the terminal server. The transfer of data to the local computer is not possible. Analysis results are made available only after a data protection clearance test by the RDC.
- **On-Site:** Data are made available for analysis at a secure computer on RDC premises and in a controlled environment. As with remote desktop access the analysis results are made available only after a data protection clearance test by the RDC.

The extent of information access from the data made available differs according to the mode of access, which further impacts analytical potential (cf. Figure 1). More detailed information is made available for data users in accordance with the degree of restrictions governing the user's data access through technical and organisational measures.[6] Such procedures ensure the highest degree of usability, and simultaneously, the best possible data protection.

[5] The RDC's data protection policy is based on the portfolio approach of Lane et al. 2008, pp. 6, on upon which the Leibniz Institute for Educational Trajectories (LIfBi) (cf. Koberg 2016, pp. 699) and the RDC of the Federal Employment Agency at the Institute for Employment Research (cf. Hochfellner et al. 2012, p. 9) have oriented themselves. The RDC has adapted the portfolio approach to the requirements of its own data files and uses four categories of measures in securing data protection, which are combined in various ways: legal-institutional measures, informational measures, technical measures and statistical measures.

[6] Cf. Chapter 8 on the various levels of anonymisation and analytical potential of the CUF and the differing SUF variants.

Figure 1: Modes of Access and Analytical Potential



[Datenprodukte] With the *Digital Object Identifier* (DOI) 10.21249/DZHW:phd2014:2.0.0, central information on the study, further documentation materials and an overview of available data products from the study can be found on the website.

The available data of the PhD Panel 2014 are saved in two data sets. There is an individual data set in in wide format and a spell data set in long format (cf. Chapter 6.5). For the SUF and CUF[7] for each mode of access both data sets are available respectively with analytical potential specific to the mode of access (cf. Figure 1).

Download-SUF and Download-CUF are available respectively in Stata and SPSS format. For Remote Desktop and On-Site modes of access, by default data sets are available in Stata format.

[Charges for Data Access] Currently SUF and CUF are available free of charge (effective June 2017). The present fees regulation can be found on the RDC website (https://fdz.dzhw.eu).

[Responsibilities of Data Users] Data users are obliged to observe the following rules[8]:

- **Scientific Use:** Data must be used exclusively for scientific research purposes. Commercial use is forbidden.
- **De-anonymisation forbidden:** Any attempt of re-identification for the units of analysis (e.g. persons, households, institutions) is prohibited.
- **Duty to report security loopholes:** If data users become aware of security loopholes with respect to data protection or data security, the RDC should be informed immediately.
- **No data disclosure:** SUF may only be used by persons who have made a data use contract. CUF may only be disclosed in the context of specified teaching activities.
- **Duty to delete:** SUF downloads must be deleted after expiry of the agreed period of use (as a rule three years) from all computers, servers and data storage devices. Likewise all backup copies, modified data sets (e.g. work-, excerpt- or help-data) as well as printouts must be destroyed.

[7] For reasons of anonymisation, however, only data from a sub-sample are available (cf. Chapter 8).

[8] The data use contract regulates terms and conditions of use in detail.

- **Notification/Provision of Publications:** The RDC has to be notified of all types of publications that are produced using data of the RDC. An electronic version of the publication shall be provided immediately. A list of existing publications based on the data can be found in the Metadata Search Portal.[9]
- **Citation rules**: The data used must be cited according to the following requirements in publications, other essays (e.g. final dissertations) and presentations.

[Citation]

- **Data Set:**
  Brandt, G., Briedis, K., de Vogel, S., Jaksztat, S. & Teichmann, C. (2016). *DZHW Phd Panel 2014*. Edited by Lange, K., Scheller, P. & Vietgen, S., doi: 10.21249/DZHW:phd2014:2.0.0, DATA SET NAME[10], released 2018. Hannover: RDC-DZHW.
- **Data and Methods Report:**
  Brandt, G., de Vogel, S., Jaksztat, S., Teichmann, C., Lange, K., Scheller, P. & Vietgen, S. (2018). DZHW PhD Panel *2014. Data and Methods Report on the PhD Panel 2014 (1st and 2nd Survey Waves).* Hannover: RDC-DZHW.

In addition, the data used must be acknowledged in the text using the following formulation:

> *"This scientific work uses data of the DZHW PhD Panel 2014, conducted by the German Centre for Higher Education Research and Science Studies (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung; DZHW). The data were published by the Research Data Centre of the DZHW, doi: 10.21249/DZHW:phd2014:2.0.0."*

[9] https://metadata.fdz.dzhw.eu/#!/en

[10] Please insert the exact name of the version of the data set that has been used, e.g. phd2014_p_d_2-0-0.dta for the download SUF of the individual data set of graduates of the PhD Panel 2014.

1 Content and Design of the Study

The DZHW PhD Panel started in 2013 as part of the funding line, ‘Research on young academics’ (FoWiN), from the Federal Ministry of Education and Research. The study looks at what influences the formal doctoral contexts and the specific learning and development conditions the graduates encountered during their PhD/doctorate phase, at the transition to the workplace after graduation and at the further professional career both within and outside of academia.

The survey population consists of 28,147 people (Federal Statistical Office, 2015) who gained a PhD/doctorate at a German institute of higher education authorised to award PhD-level qualifications in the 2014 academic year. The survey is designed as a census in order to obtain a sufficient number of cases. In other words, no random sampling was taken, but rather every graduate from the cohort was invited to take part in the first survey.

The initial survey took place in 2015; a paper-and-pencil questionnaire was posted out for this purpose around six to eighteen months after gaining the PhD/doctorate. The second survey took place approximately one year later in the form of an online survey. Three further waves are planned.

Along with basic data used for the creation of the educational report, the dataset contains detailed information on the learning and development conditions during the PhD/doctorate phase as well as on the courses of life taken by the graduates following their PhD/doctorate. In addition, the dataset contains a series of personality traits (*Big Five*, self-efficacy, internal/external locuses of control) as well as socio- and educational-biographic background information. This provides a large and hitherto unavailable analysis potential for higher education and scientific research. The panel design and the collection of month-by-month historical data enable causal analyses at an individual level over the course of time (e.g. in the form of event data and sequence pattern analyses).

2 Survey Instruments

In the first survey wave of the Graduate Panel 2014, a standardised paper questionnaire in German was used as a survey instrument.[11] The second survey was completed using a standardised online questionnaire in German.[12] The survey instruments include, on the one hand, core elements which were asked anew in each wave. This includes a table for the month-by-month collection of the key periods of employment, the variables of the professional and academic activities and information on the family situation. On the other hand, the particular survey waves contain individual specialisations, such as the retrospective recording of the doctoral phase in the first wave.

2.1 Contents of the Survey Instruments

The first survey focussed on the learning environment and the general framework of the PhD/doctorate as well as the academic activities and practical work experience during the doctoral phase (see Table 1).

Table 1: Thematic structure of the first survey

Topic	Question numbers
General framework /Information on PhD/doctorate phase	1.1 to 1.21
Financing	2.1 to 2.3
Mentoring and support	3.1 to 3.10
Academic activities	4.1 to 4.10
Practical experience	5.1 to 5.6
Personal characteristics, aims and objectives/Personality variables/Attitudes	6.1 to 6.6
Professional development	7.1 to 7.5
Employment and occupational activity	2.3; 7; 5; 8.1 to 8.13
Sociodemographic variables	9.1 to 10.4
Previous education/Access to higher education	9.9 to 9.11
Social background	10.1 to 10.4

A theoretically sound and empirically tested model was developed for the recording of data on the learning environment during the doctoral phase. The theoretical concept of learning environments is based on the SSCO Model (*Structure - Support - Challenge - Orientation*) from Bäumer, Preis, Roßbach, Stecher & Klieme, 2011. Another core element of the first survey is the table of occupation (question 2.3), which was used to record the job activities since the start of the PhD/doctorate. Respondents were asked to provide information about the start and end of the work, the occupational status, the working hours, the term of employment and, where appropriate, about changes of employer and the academic relevance of the job. In addition, re-

[11] Interviewees could also request a PDF questionnaire by e-mail in German or English if desired.

[12] The questionnaires can be downloaded from the RDC website. There is also a question flow diagram showing the filtering procedure for both questionnaires.

spondents were asked in questions 8.1 to 8.13 to give further details on their first job since gaining their PhD/doctorate and on current employment (e.g. profession, industry sector, adequacy). Furthermore, psychometric scales were included in the questionnaire for measuring general self-efficacy (ASKU) (Beierlein, Kovaleva, Kemper & Rammstedt, 2014), the 'Big Five Personality Traits (BFI-10) (Rammstedt, Kemper, Klein, Beierlein & Kovaleva, 2014) as well as internal and external locuses of control (IE-4) (Kovaleva, Beierlein, Kemper & Rammstedt, 2014) based on the templates and guidelines of GESIS. Using a slightly amended version of the scale 'General openness towards geographical mobility (modified version)' (Otto, Glaser & Dalbert, 2004), the willingness of respondents to relocate within Germany on account of their job or to spend a limited period of time abroad was recorded. Individual items (some of them in slightly modified form) were taken from different surveys on the subject of PhD/doctorates and highly qualified people.[13]

Some of the questions from the first survey were used in modified form in the second survey. This was necessary for several reasons: firstly, there was a change of survey mode and not all questions were equally suitable for paper-and-pencil and online surveys. Secondly, clarification of the question was required for some of the items. Thirdly, the time reference was modified for many of the questions in order to facilitate the updating of data in subsequent panel waves.

In the second survey, a range of information was also gathered retrospectively on the educational path which, due to lack of space, couldn't be collected in the first survey (Table 2).

Table 2: Thematic structure of the second survey

Topic	Question numbers
PhD/doctorate result	3.1 to 3.4; 6.3 to 6.4, 6.16 to 6.17
Academic and further professional qualifications	5.10 to 5.12; 6.10
Academic activities	4.2 to 4.7; 5.1 to 5.10; 5.13 to 5.19; 6.1 to 6.3; 6.10
Educational and occupational biography before commencing PhD/doctorate	1.1 to 1.2; 2.1 to 2.4
Employment history since finishing PhD/doctorate	4.1; 4.8 to 4.9; 6.7 to 6.20
Mobility	1.3; 1.6; 2.5 to 2.6; 6.21 to 6.24
Personal characteristics/private life situation	6.25 to 6.34

The questionnaire for the second survey contained two instruments for collecting the work histories following the PhD/doctorate: As in the first wave, an occupation tableau was used for recording the key periods of employment on a month-by-month basis and to illustrate detailed variables for each period. The tableau was supplemented with the question regarding place of work. In addition, an occupation calendar was incorporated into the questionnaire in the second

[13] For example, Federal Statistical Office [BFS] (2011) (Wave 1, Question 2.1), Jungbauer-Gans und Gross (2013) (Wave 1, Question 3.3), Grühn, Hecht, Rubelt und Schmidt (2009) (Wave 1, Question 4.1), Auriol, Felix und Schaaper (2012) (Wave 1, Question 5.4), Blickle, Kuhnert und Rieck (2003) (Wave 1, Wave 7.1), Egeln, Gottschalk, Rammer und Spielkamp (2003) (Wave 2, Question 6.3), Federal Statistical Office (2013) (Wave 2, Question 6.4). The Institute for Entrepreneurship and Innovation at the University of Potsdam [BIEM-CEIP] (2010) (Wave 1, Question 4.10) and the Centre for Research on Higher Education and Work [University of Kassel] (2009) (Wave 1, Question 7.2).

wave, which was used to record the professional and non-professional activities since the PhD/doctorate on a monthly basis (e.g. further training courses, scholarships, parental leave, family work, unemployment).

2.2 Pre-test

Before the first survey, initially a cognitive and later a quantitative pre-test were carried out. The main aim here was to test the newly created items for surveying the learning environment during the PhD/doctoral phase and to test their usefulness for questioning doctoral candidates from different subject areas and types of PhD/doctorate courses. The objective of the pre-tests was to determine the understanding of the questions and the response behaviour of the doctoral candidates, to establish the length of response time and to uncover possible sequence effects. This is documented in detail in a workshop report (Brandt, Vogel & Jaksztat, 2016).

The questionnaire for the second survey was also tested in advance as part of a cognitive pre-test with doctoral candidates from various types of formal PhD/doctorate courses and subject areas being tested. The programming of the questionnaire and the carrying out of the survey were both performed using the DZHW online software 'Zofar'.

3 Population and Contacting

In order to be able to draw as representative a picture as possible and meaningful conclusions about the PhD/doctorate conditions, the professional and private development of the doctoral candidates and to enable subject and context-specific analyses, the survey was designed as a census. The population consists of all doctoral candidates who gained a PhD/doctorate at a German higher education institution authorised to award PhD/doctorate-level qualifications in the 2014 academic year (winter semester 2013/14 and summer semester 2014). The official statistics show 28,147 doctoral candidates nationwide in Germany for the relevant examination year of 2014 (Federal Statistical Office, 2015).

For data protection reasons, the initial contact with the doctoral candidates and the sending of the survey documents had to take place via the higher education institutions (addressing procedure). The administration departments of all 146 higher education institutions authorised to award PhD/doctorates were informed in advance of the planned survey and asked to give their support to the research project. Provided the university administration department expressed no objection, requests were made to the departments responsible for managing the doctoral records within the higher education institutions.

4 Implementation of the Surveys

[Maintenance of Contacts and Addresses] For data protection reasons, the sending of the questionnaires for the first survey took place via the deaneries and the examination offices responsible for the PhD/doctorate courses. In most cases the responsibility lay within the individual deaneries. At a few of the higher education institutions, central or decentralised examination offices or graduate academies were identified as having responsibility.

To enable direct contact through the DZHW with the people who continued to be willing participants in the second survey wave, their address details were collected in the questionnaire from the first wave. Upon receipt of a questionnaire at the DZHW, a unique identification number was stamped on both the questionnaire and the address section of the questionnaire using a numbering stamp and a reference list created from address sections from the identification number to the associated address in each case.[14] In order that people be taken into account who had changed address in the meantime, the address lists were examined between the respective waves and updated where appropriate.

Contact was made between the first survey and the second survey wave. All respondents who had provided an e-mail address were informed about the project status by e-mail and thanked for their participation to date. Furthermore, the upcoming second survey was mentioned in the correspondence. The second contact enabled the list of e-mail addresses to be checked for up-to-dateness and potential recording errors. If there was no (valid) e-mail address then contact was made by post in the interim. The 194 panel participants without (valid) e-mail addresses were asked to provide their current e-mail and postal addresses using an enclosed postcard. The e-mail addresses of 68 respondents were obtained in this way. A further e-mail address was successfully obtained through a name correction following an address revision by Deutsche Post. A total of 4,822 address data were updated in the panel population.[15]

The second survey was conducted as an online survey. For this purpose, it was possible to access the updated address details of 4,816 respondents from the first survey who had provided their contact details and had agreed to take part in subsequent questionnaires.

[Survey Instruments] The survey documents for each person to be surveyed in the first survey wave consisted of a letter (incl. data protection information), the paper questionnaire, an accompanying letter from the Federal Ministry of Education and Research and a prepaid envelope addressed to the DZHW for returning the completed questionnaire. Two reminder letters were also sent. The survey for the second wave was programmed using the DZHW's online survey software 'Zofar'. The three reminder letters were purposely only sent to those specific individuals who had not yet taken part in the second survey.

[Fieldwork Phase] The survey period for the first survey wave lasted from 15 December 2014 until 17 February 2016. The two reminder letters were sent around four and eight weeks respectively after the start of the fieldwork phase.[16] The reason for this relatively long time frame lies in the fact that the administration departments of the higher education institutions

[14] To ensure data protection, the address section was detached from the questionnaire and the reference list was stored separately from the survey data on a protected server.

[15] Out of these, four respondents in the first survey gave no particulars and a further six respondents refused in advance to participate in the second survey.

[16] From December 2014 to January 2015, it took over four weeks in some cases between the sending of the survey documents and the sending of the first reminder due to delivery delays. A major reason for this was a change in postage costs at short notice, which meant that post-franking with special stamps was required.

responded to the request to participate in the project with varying degrees of promptness. It was not possible to reach all the administrative offices for the sending of the reminders.[17] However, it was possible to send the survey documents and the two reminder letters to the majority of the administrative offices as planned. Due to the method of contact used via the examination offices, the DZHW was unable to have any direct influence on the precise delivery time of the survey documents.

For the second wave, it was possible to set a specific date for the survey invitations using the address list made available to the DZHW from the first wave. The second wave survey was accessible to respondents from 15 March 2016 to 18 April 2016.

[Measures to Increase Response] The measures for increasing the response rate were aimed on the one hand at encouraging the higher education institutions to provide organisational support for the survey, and at the individual motivation of respondents on the other. As a thank you for their support, the higher education institutions received exclusive preliminary the results from the survey. On request, the faculties were able to obtain separate evaluations for their own particular institution, provided there were sufficient cases available for this purpose. The measures for increasing the response rate aimed at the individual motivation of respondents included, in the first instance, highlighting the subjective and social relevance of the subject in the letter to respondents. Furthermore, the accompanying letter from the Federal Ministry of Education and Research (BMBF) presented the political relevance of the study for higher education institutions and called on their support. The sending of reminder letters took place in addition to this.[18] The confirmation of selected results from the study [19] after the first wave was intended, on the one hand, to improve the bond between study participants and the panel in order to motivate them to take part in subsequent waves. On the other hand, the sending of results served towards a renewed updating of the address list. In addition, a laptop, three iPads and several book vouchers were raffled among all participants. A smartphone and several travel vouchers were raffled among all the participants of the second wave.

[17] The reasons for this were, for example, too much time and too high personnel costs, longer absences among responsible staff or personnel changes within the relevant administrative offices.

[18] Based on the contact method applied by the examination offices, reminder letters were sent to all people from the random sampling in the first survey – including those who had already taken part in the survey – as the exam officials were unaware which persons had already returned a questionnaire to the DZHW. Non-participants in the second wave were sent three reminders.

[19] A brief overview of the key findings was available in an online flyer. More detailed information according to subject was also made available.

5 Response Rate

[Response Rate] The survey of doctoral candidates from the 2014 examination year took place at a total of 112 higher education institutions with authorisation to award PhD/doctorates. For the most part, the higher education institutions supported the survey through the central participation of all faculties (80 institutions), 32 institutions took part with individual faculties, 15 institutions didn't take part (see Table 3). No survey took place at 19 higher education institutions authorised to award PhD/doctorates, as none of the doctorate courses at these would have been completed within the time frame under investigation.

Table 3: Participation of Higher Education Institutions

Participants	Number
Higher education institutions authorised to award PhD/doctorates	146
No PhD/doctorates completed in the 2014 examination year	19
Participated	112
participated in full	80
participated in part	32
did not participate	15

Table 4 shows the number of questionnaires sent and the response rates. While in the first wave around 27 per cent of the doctoral candidates invited took part, around 66 per cent of willing participants from the first wave were successfully recruited for the second wave.

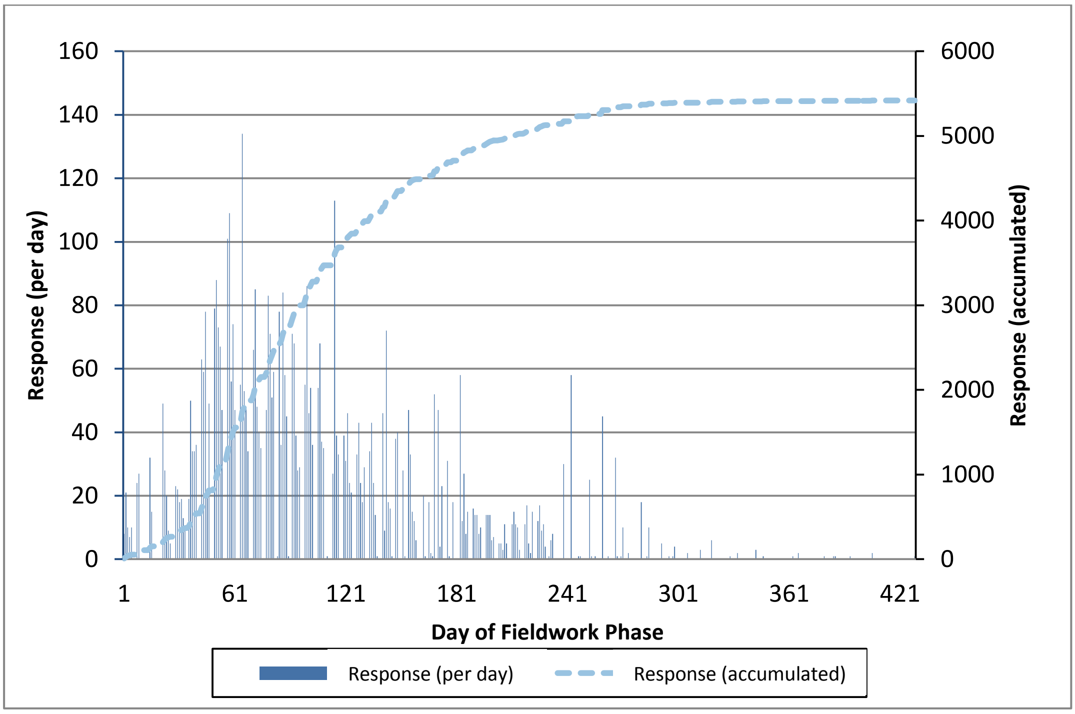
Table 4: Gross and Net Response Rates from the DZHW PhD Panel 2014

	1st Wave	2nd Wave
Survey documents sent	19.916	4.816
Sending of survey documents confirmed[20]	19.900	4.816
Survey documents response rate	5.423	3.188
Usable survey documents	5.410	3.184
Response rate (usable/sending confirmed)	27,2 %	66,1 %
Response rate (usable/sent)	27,2 %	66,1 %
Percentage Wave 2 (gross) of Wave 1 (gross)		24,2 %
Percentage Wave 2 (net) of Wave 1 (net)		58,8 %
Percentage Wave 2 (net) of Wave 1 (gross)	-	16,0 %

[20] As the direct sending of the survey documents was not possible for data protection reasons, the actual sending had to be estimated. The response form from the examination offices, on which the actual number of questionnaires sent was marked, served as a basis for the estimation.

In the first wave, 19,916 questionnaires were sent to 127 higher education institutions. Of these, 5,423 questionnaires were returned by the respondents (see Table 4).[21] Excluding empty, duplicated and illegible questionnaires, it was possible to electronically record 5,410 questionnaires (see chapter 6.3). Figure 2 shows the responses to the questionnaires over time in the fieldwork phase of the first survey wave. A large proportion of the completed questionnaires reached the DZHW in the first half of the fieldwork phase, during which the reminder postcards were also sent out. At the same time, it should also be noted that some further questionnaires were returned at later dates, even long after the second reminder had been sent.

Figure 2: Response Rate of the DZHW PhD Panel 2014 over time, 1st Wave

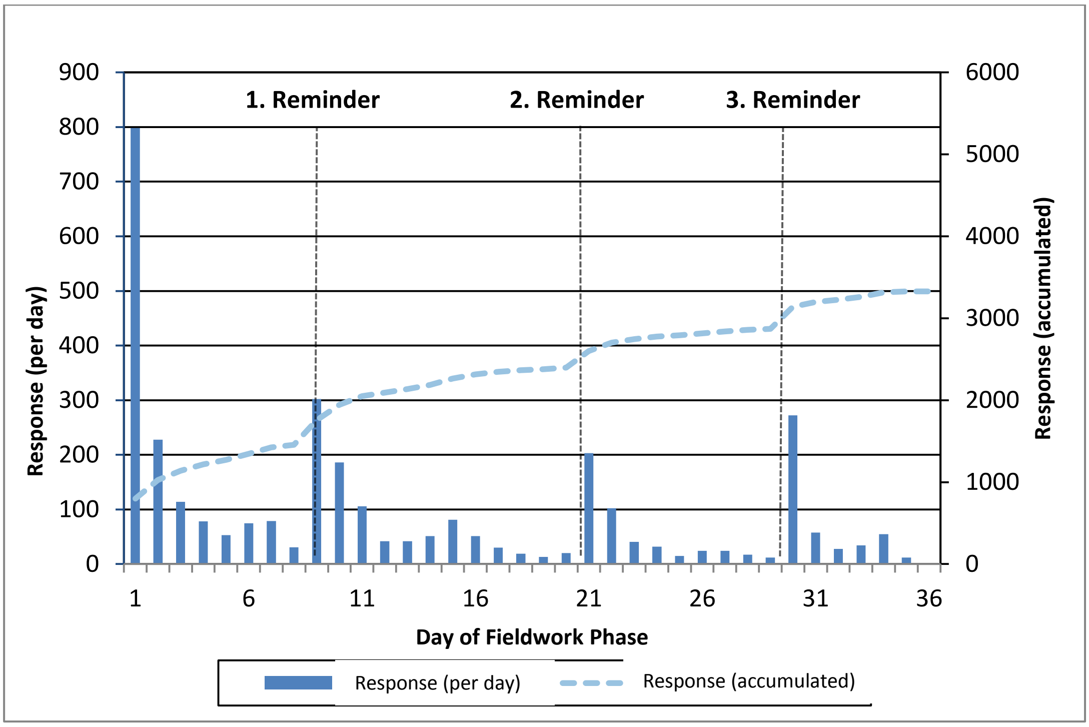


4,816 people, i.e. around 89 per cent of the 5,410 participants in the first wave, agreed to be contacted for further surveys. A total of 3,188 graduates took part in the second survey, from which 3,184 surveys were evaluable.[22] The response curve in Figure 3 shows the response rate since the beginning of the survey and the dates of the three reminder activities. On 15.03.2016, 4,816 respondents were invited (less eleven non-deliverable invitations) to participate in the next survey (second wave). On the first day of the survey, 24 per cent of the respondents took part and 35 per cent in the first three days. This clearly shows the effects of the reminders, which resulted in a noticeable increase in the response rate in each case.

[21] Three questionnaires were received after the data entry and were therefore no longer acceptable. Five questionnaires were received uncompleted, while three further ones had to be rejected due to highly implausible data. A questionnaire was also deleted from one person, which had been completed twice. One person requested the deletion of his survey data.

[22] A person was classified as having taken part whenever he/she responded to at least one question.

Figure 3: Response Rate of the DZHW PhD Panel 2014 over time, 2nd Wave



[Panel Attrition] The PhD Panel 2014 was affected by panel-typical attrition processes[23]. These include the basic refusal to participate in subsequent surveys (non-provision of address details for contacting in the second wave) and the non-participation in the second survey wave after (attempted) contact was made.

A look at the response rate over time shows that the gross sample in the second wave only includes 24 per cent of the survey documents sent out in the first wave. Of the 5,410 evaluable cases from the first wave, it was possible to survey 59 per cent of the respondents again in the second survey (see Table 4). Furthermore, only 16 per cent of the graduates contacted in the first survey took part in both survey waves.

[23] For panel-typical attrition processes see Schnell, Hill & Esser, 2005, p. 241.

6 Data Preparation

The steps described below for data preparation in the first and second survey waves were partly carried out analogously, and in some cases they were carried out differently due to the survey mode. The procedures described in Chapters 6.1 to 6.3 had already been conducted by the primary research project. The generation of variables (Chapter 6.4) was carried out by the primary project as well as the RDC during data preparation. Procedures described in Chapters 6.5 to 6.7 were carried out by the RDC building on the work of the primary research project. Additional procedures (e.g. weighting and anonymisation) are explained separately in Chapters 7 and 8.

6.1 Data Transfer

In the first wave, the respondents' data were transferred from the paper questionnaires on the basis of a code plan for further processing in a computer-readable format (the second survey took place online). Previously, numerical codes were marked on the paper questionnaires for most of the open responses (see chapter 6.2) and preliminary manual corrections performed to facilitate the data transfer (see chapter 6.3).

[Production of a Code Plan] A code plan was produced based on the survey questionnaire. It recorded to which question or sub-question a variable is assigned, the name of the variable and the numerical codes used for the standardised answers of the respondents. To establish the order of data entry, the variables were additionally numbered.[24]

[Data Entry] For data transfer, the code plan, further instructions on data entry and the prepared paper questionnaires were given to an external service provider. Their typists manually performed the compilation of the data.

6.2 Coding of Open Responses

Before the data transfer, the (semi-) open responses were coded. Using a coding list, numerical codes were assigned to them. For each variable, various code lists were used. This was done using classification keys for official statistics (e.g. German Classification of Occupations, key lists of student and examination statistics etc.) or keys already used in prior graduate panels. For some variables, new code lists were developed on the basis of the entries from the PhD Panel 2014. For some semi-open questions, no new variables with numerical codes were created. Instead, entries were only assigned to the existing (closed) response categories. Some of the open questions were not encoded as they were mainly collected as context information for the encoding of other open data.[25] Coding choices by the primary research project were not modified.

Coded topics and respective code lists are presented in Table 5. The data set contains exclusively the coded numerical variables. The open entries themselves are not contained in the data

[24] Data were generated in a simple, column-oriented text format without a heading containing the variable names. The code plan therefore established in which order the data were to be generated so that the data belonging to a variable could be entered in the correct column.

[25] This applies in both waves to the fields of activity and jobs, which were recorded along with the occupational title in questions 8.4 and 8.5 of the first wave and in question 6.12 of the second wave. The data served merely for the collection of additional information for the encoding of the occupational title, which was also gathered as an open response.

set. The values of the variables are documented in the Data Set Report as well as in the Metadata Search Portal[26].

Table 5: Coded Topics and Code Lists in the DZHW PhD Panel 2014

Topics	Code List Resource
PhD awarding higher education institution, higher education institution of graduation	German institution of higher education: Destatis Key List of Student and Examination Statistics (Winter Semester 2013/2014 and Summer Semester 2014) Foreign institution of higher education: DZHW encoding on the basis of country codes
PhD/doctorate subject, subject/field of study	Destatis Key List of Student and Examination Statistics (Winter Semester 2015/2016 and Summer Semester 2016)
Profession, parents' professions	German Classification of Occupations 2010
Place of professional activity	Allocation of postal codes
Country of professional activity	Nationality and Region Classification 2014
Place of birth, place where university entrance qualification was gained	Allocation of postal codes
Country of birth, Country where university entrance qualification was gained	Nationality and Region Classification 2014
Graduation	DZHW coding based on the Key List of Student and Examination Statistics (Winter Semester 2013/2014 and Summer Semester 2014)
other open responses	Allocation to specific categories or project coding

6.3 Data Checking and Data Cleansing

[Preliminary Manual Correction] In the first wave, a manual inspection and, if necessary, an amendment of the respondents' data was performed on the paper questionnaires before the data transfer.[27] This was intended to facilitate the capturing of data. The form of the existing data was amended for this purpose. For example, hardly legible data or crossings out made by the respondents were highlighted, numerical data entered right-aligned in the designated boxes and verbal entries translated from grades to figures (e.g. 'good' = 2.0).

On the other hand, the manual inspection was also aimed at correcting initial errors or inconsistencies in the respondents' data before the software-supported correction (see below). For example, the response 'none' had to be scored out if one or more articles in a row were marked 'total number'. Moreover, checks were made as to whether the data regarding occupational activities in the employment tableau concurred with the corresponding data on gainful employment (questions 2.1 and 8.1-8.13). Any identified inconsistencies were, if possible, eliminated by

[26] https://metadata.fdz.dzhw.eu/#!/en

[27] The number of corrections was not recorded centrally, but simply on the paper questionnaires, and can therefore no longer be systematically reconstructed.

the comparison with other responses in the questionnaire or alternatively by assigning a corresponding missing code (see chapter 6.7).

[Software-Assisted Correction] Following the data transfer in the first wave, a comprehensive review and correction of the data took place with the aid of the DZHW's own in-house software, and in the second wave with the assistance of the statistical software program Stata. On the one hand, the aim here was to identify any errors from the previous preliminary manual correction and data transfer, while on the other hand, further inconsistencies in the respondents' data that were unable to be checked in the pre-correction could also be identified. Following this, valid value ranges and response combinations were defined and checked based on formal rules. The following types of tests were carried out:

- *Test of Value Ranges:* It was tested whether the response lay in the value range defined of the respective recorded variable.
- *Test of Adherence to Filter Procedures:* Based on the defined filter procedure of the questionnaire, it was tested whether responses that would have been expected from the respondent were not (i.e. completeness test) and whether responses were made that should not have been (i.e. filter errors).[28]
- *Test of Variable Consistency:* The consistency of responses within a questionnaire as well as between survey waves was tested. In addition to combinations of characteristics, which were already tested in the preliminary manual correction, more complex feature combinations could also be tested here.

Missing, incorrect or implausible values were first tested using the paper questionnaire to determine whether the corresponding value had been falsely (or not at all) transferred. Then the correct value was inferred using other responses in the questionnaire. In case of doubt, a specific missing code was assigned (cf. Chapter 6.7). Corrections of mistakes were documented[29] and checked by at least one further person.

[Deletion of Cases] Cases were removed from the dataset in all three waves. A case was deleted if it had been entered twice (one case), if less than one question had been answered (five cases) or if there were too many inconsistencies (three cases). Eight cases in total were deleted in the first survey wave and three cases in the second survey wave.

6.4 Generation of Variables

In addition to the variables containing the coded answers of the respondents, the PhD Panel 2014 also generates variables. One the one hand, this includes variables that were numerically coded from the originally open entries (cf. Chapter 6.2). On the other hand, variables were changed due to data protection reasons (cf. Chapter 8) and more frequently required variables were generated from the values of one or more source variables (e.g. merging course subjects in to areas of study and subject groups or deriving the location and type of the higher education institution from the higher education institution variables). The newly generated variable is identified in the data by the suffix "_g#". An overview of all generated variables for the PhD Panel

[28] The input filter of the variables assigned to the individual questions is documented in the Data Set Report as well as in the Metadata Search Portal (https://metadata.fdz.dzhw.eu/#!/en). They define which surveyed group should answer a question for a respective variable.

[29] Documentation of the correction of mistakes was performed manually on the paper questionnaires and thus cannot be systematically reconstructed.

2014 as well as detailed documentation of the individual variables with information on their respective characteristics and calculation rules can be found in the data set report as well as the Metadata Search Portal.[30]

6.5 Generation of the Data Sets

[Merging of the Waves] The data from the first and second waves were merged. The matching of cases took place using the respondents' identification numbers, which were assigned as part of the fieldwork phase (see chapter 4).

[Generation of Individual and Spell Data Set] The merged data were stored in two separate data sets. The *Individual Data Set* contains a large part of the survey data as well as the additionally generated variables. For this format, there is a data record for each respondent (wide format). The sequence of the variables is oriented to the sequence of related questions in the questionnaire. The *Spell Data Set* contains only the answers from the calendars (Question 4.8 of the 2nd wave). For each respondent, one or more spells are recorded. A spell is thus defined as a time period distinguished by a specific occupation (e.g. employment or training) or other status (e.g. parental leave or unemployment. Each spell of one respondent corresponds to one data row (long format). The structure corresponds to the standard structure for spell data (cf. Scherer, Brüderl 2010, p. 1042). The spells were sorted by case, i.e. all spells of the same respondent follow each other directly. Different types of occupation in the same time period were coded as independent spells. If activities of the same type immediately followed each other, or were practised simultaneously, they were summarised as one spell. Thus it cannot be discerned from the spell data whether a spell comprised one or more activities of the same type. However, detailed information is contained in the corresponding variables of the individual data set regarding employment activity and academic qualification (Wave 1 Question 2.3, Wave 2 Question 4.9). . The data from these variables can be connected with the spell data. Individual and spell data sets can be merged using the respondent's identification number (variable: *pid*).

[Data Format] All data sets are available in Stata as well as SPSS format (cf. Section III).

6.6 Assignment of Variable Names, Variable Labels and Value Labels

[Variable and Value Label Assignment] For variable and value label assignment, formulations from the questionnaire were used, or in some instances, concise formulations were chosen. As a rule, the variable labels are based on the corresponding question. Depending on the type of question, value label assignments are based on the response options or a combination of the question and response options. For generated variables based on definite classifications, value labels were adopted verbatim from the classification keys. Variable and value labels are available in German and English. In the SPSS format, there is a separate data set for each language. In the Stata format, bilingual labels were created in the same data set.

[Naming Variables in the Individual Data Set] A consistent naming system was created at the RDC for the naming of variables. With the exception of the identifier variable (pid) as well as the wave variable (wave),[31] variable names in the individual data set were formed according to a prefix-root-suffix scheme that facilitates automated processing. In addition, the variable names

[30] https://metadata.fdz.dzhw.eu/#!/en

[31] This contains information on case participation in both waves (participation only in the first or in both waves).

provide meta-information on the corresponding variable. The prefix of the variable contains the wave identification in one letter. The root of the variable contains the thematic area to which the variable is assigned and is denoted by a three-letter English abbreviation. Table 6 presents an overview of the various thematic areas of the PhD Panel 2014 as well as the related abbreviations for the root of the variable name. The suffix, separated from the root by an underscore, contains various additional information so as to identify generated variables as well as various modes of data access.

Detailed information on variable assignment for the PhD Panel 2014 can be found in the Data Set Report.

Table 6: Thematic Areas and Abbreviations used for Variables Names of the DZHW PhD Panel 2014

Thematic Area Abbreviation	Thematic Area (English)	Thematic Area (German)
sys	system variables	Systemvariablen
stu	studies	Studium
occ	occupation	Beschäftigung
ski	skills	Fähigkeiten
fvt	further vocational training	Berufliche Fort- und Weiterbildung
dem	demographic information	demographische Informationen
abr	(experiences) abroad	Auslandserfahrung
fin	financing	Finanzierung
fut	future prospects	Zukunftsaussichten
goa	goals (occupational, life)	(Berufs- und) Lebensziele
inc	income	Einkommen
job	job	Jobs
mot	motives	(Tätigkeits-)Motive (für Studium / Ausbildung / Promotion / Erwerbstätigkeit)
mov	move	Regionale Mobilität/Umzüge
net	network	Netzwerk
par	Information about partner	Informationen über Partner
sat	satisfaction	(Berufs-)Zufriedenheit
sch	school	Schulzeit
sci	scientific experiences	wissenschaftliche Aktivität
voc	Vocational training/education	(Berufs-)Ausbildung
con	conditions of doing a phd	Rahmenbedingungen der Promotion
phd	to do a phd	Promotion
scc	Structure-Support-Challenge-Orientation-Scale (SSCO)	Dimension zur Erfassung von Lernumwelten
sel	self-efficacy scale	allgemeine Selbstwirksamkeit Kurzskala
bfi	Big-Five-Inventory (short)	Kurzskala zur Erfassung der fünf wichtigsten Persönlichkeitsdimensionen
wgt	weights	Gewichtungsvariablen

[Variable Labels in the Spell Data Set] Variables in the spell data set include the respondent's identification number (pid), the identification number of the respective spell (eid), activity (status) as well as the beginning and end dates of the spell time period. The latter is coded using four variables (Month: begin_m and end_m; Year: begin_y; end_y).

6.7 Coding of Missing Values

For coding missing values, a comprehensive system was created in the RDC, in order to guarantee unified coding for missing values across various data sets of the DZHW. Missing responses were coded using three-figure negative values. Table 7 presents an overview of the system for coding missing values. The coding for missing values used in the PhD Panel 2014 is highlighted.

Missing values can be assigned to four different groups. First, missing values may arise if the respondent does not answer the survey questions (i.e. non-response). Second, missing values may be assigned due to the filter procedure, i.e. if questions are not relevant to the respondent (not applicable). The third group contains missing values assigned through the primary research project or the RDC in the course of the data preparation (i.e. edited missing value). This group also includes the encoding for missing values, which was assigned for particular variables[32] due to anonymisation measures (see Chapter 8).

Table 7: System of the RDC-DZHW for Missing Values

Range of Values	Code	Value Label
-999 bis -990: Non-response	**-999**	**don't know**
	-998	**no answer**
	-997	no answer (response category)
	-996	**interview break-off**
	-995	**not participated (panel)**
	-994	refused
-989 bis -970: Not applicable	**-989**	**filtered**
	-988	does not apply
	-987	missing by design (questionnaire split)
	-986	missing by design (wave)[a]
	-985	missing by design (cohort)[b]
-969 bis -950: Edited missing values	-969	unknown missing[c]
	-968	**implausible value[d]**
	-967	anonymised
	-966	**not determinable[e]**
	-965	**invalid multiple answer**
-949 bis -930: Item-specific missing values	*(not assigned)*	
-929 bis -920: Other missing values	-929	loss of data

[a] This value is only assigned for data sets in long format.
[b] This value is only assigned for pooled data sets.
[c] This value is assigned when no cause can be reconstructed.

[32] A fourth group includes special codes for missing values, which were only assigned for particular items as part of the creation of a concrete dataset.

[d] Responses which are classified as implausible due to various factors in the coding phase receive this value. An exact reconstruction may no longer be possible.
[e] This category is assigned when clear coding is not possible, e.g. open response which could not be coded because it is illegible.

7 Weighting

7.1 Procedure and Instructions for Use

As part of the project, two separate weighting steps were carried out. In the first step, the master sample of the first wave was adjusted to the population as ideally as possible. The realized sample of the first wave was compared to the population of PhD holders of the examination year 2014. The comparison between the realized sample and the population showed that the master sample presented an accurate depiction of the target population. Differences between sample and population could be compensated by calculating cross-sectional weights (redressment) (see Section 7.2). In the scope of attrition analyses on the level of higher education institutions a possible connection between the participation probability of a higher education institution and specific characteristics of the higher education institution was additionally checked. The considered characteristics were the type of higher education institution, the number of doctorates in the examination year 2014 and the federal state. No systematic attritions could be determined. A check of attritions on the level of administrative offices could not be carried out due to missing information concerning possibly relevant characteristics.

For the second wave panel weights were additionally calculated in order to compensate possible selection mechanisms concerning the participation in the follow-up survey. In contrast to the non-participants of the first wave, for non-participants of the second wave comprehensive information from the initial survey was available, which could be used for modelling the participation probability.

Table 8: Weighting Variables in the Scientific Use File

Variable name	Description
wgt_t1	Cross-sectional weight of the first survey wave (redressment weight)
wgt_t1t2	Longitudinal weight of the 2-wave-panel (trimmed)

The provided weights can be incorporated into Stata with the aid of .ado-specific options.[33] The weight wgt_t1 is intended for evaluation of the first wave, the weight wgt_t1t2 for evaluation of the two-wave-panel.

7.2 Weighting of the Data

For the calculation of the cross-sectional weights of the first wave the distribution of PhD holders was contrasted by field of study x region (east Germany/west Germany) x gender in the sample and in the population. The reference data for the population was collected by the Federal Statistics Office.[34] The combination of the three weighting factors resulted in a contingency table with 59 x 2 x 2 = 236 cells, for which a target-actual-comparison was performed. Each combination of characteristics was assigned a weight resulting from the quotient of the propor-

[33] See also the Stata guide (command: *help weights*).

[34] Statistisches Bundesamt, Hauptberichte. Auswertung aus der ICE-Datenbank der Länderministerien (ICE = Information, Controlling, Entscheidung). Bestand: 50001

tion of the respective combination of characteristics in the population and its proportion in the sample. Individual cells had to be combined for the calculation of the weights.[35]

Additional, wave-specific attrition processes had to be modelled for the second wave. For the nonresponse analyses survey data of the first wave was used. The basis for the panel weighting was a logistical regression model, which predicted the probability of a person participating in the second wave. As part of an exploratory approach, individual variables were identified which help illustrate the probability of participation. The final estimation model included the variables gender, age, doctoral conditions, doctoral result, place of birth, subject group, current/last gross salary, the industry sector in which the person was working in during wave 1 and an item for self-evaluation of the personality from the Big Five inventory ("I am easy-going, prone to laziness"). To enable an estimation of the probability of participation for each person, missing values for all variables were included in the model as an additional category.[36] The conditional probability of participation could be ascertained from this model, whose reciprocal value represents the nonresponse weight for the second wave:[37]

$$NR_{gew_{t+1_i}} = P\left(Res_{t+1_i} \middle| \sigma_{t_i}\right)^{-1}$$

The overall weight for the two-wave panel (wgt_t1t2) of the dataset is the result of the product of the cross-sectional weight (wgt_t1) and the longitudinal weight ($NR_{gew_{t+1_i}}$):

$$wgt_{i_{t1t2}} = wgt_{i_{t1}} \times NR_{i_{gew_{t2}}}$$

The initially calculated weights exhibit a small proportion of outlying weighting factors. In order to remove them, all weights were subjected to a trimming according to Potter 1990 (see also Valliant et al. 2013, pp. 388). The procedure is based on the assumption that the weights conform to a probability distribution (beta distribution). All those weights that lie above the 99 percent quantile are truncated to this limit. Excess on the other side of the truncation is distributed among the remaining weights.

[35] This was the case for cells whose combination of characteristics occurred only very rarely in the population and which could therefore not be represented by the sample.

[36] See Appendix 1 for the estimation model.

[37] The process corresponds logically to *Propensity Score Matching*, which goes back to Rosenbaum and Rubin (1983) (see also Blumenstiel and Gummer (2015)).

8 Anonymisation

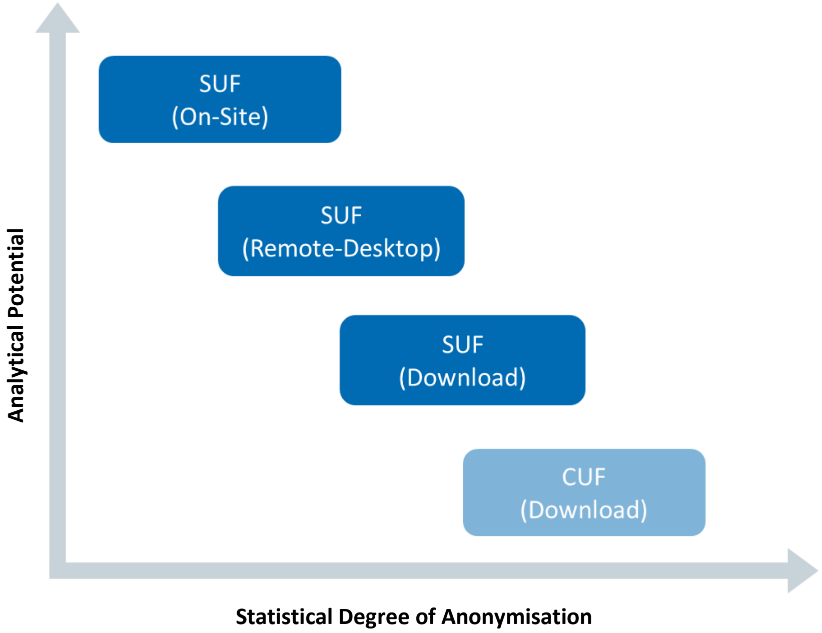
[Data Protection Legal Framework] The Federal Data Protection Act (BDSG) applies to personal data that the DZHW collected through volunteer surveys.[38] Accordingly, personal data that are collected during scientific research may be processed or used exclusively for the purposes of scientific research (cf. §40 para. 1 BDSG). Moreover, personal data must be anonymised (cf. §40 para. 2 BDSG) in order to protect respondents. According to the BDSG, the procedure of anonymisation is defined as “the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual” (§3 para. 6 BDSG). Regarding the disclosure of data from scientific research projects to third parties, the data must either be *absolutely anonymised* so that no reference to the person can any longer be produced, or at least *de facto anonymised* so that the construction of a reference to a person would mean a disproportionally high expenditure and thus the likelihood of re-identification of a person is minimal.

[Data Access, Level of Anonymisation and Analytical Potential] For the PhD Panel 2014, the RDC makes two types of data files available. Whereas SUF for scientific secondary use are de-facto anonymised, CUF for teaching and exercise purposes are absolutely anonymised. The anonymity of the surveyed persons is thus protected by a combination of statistical measures and technical access barriers. The more strongly data access is technically controlled, the lower is the risk of de-anonymisation of the data, the less the data must be limited in terms of information by statistical measures and the greater their analytical potential remains.

While the CUF is directly transmitted by the RDC after registration, the SUF is provided using three different modes of access: download, remote desktop and on-site (for further information cf. Section III). For each mode of access a different SUF variant is made available, which is varyingly strongly anonymised and correspondingly contains less or more information. Figure 4 gives an overview of the respective level of statistical anonymisation and the related analytical potential. In the following the statistical anonymisation measures performed are explained according to data product (SUF/CUF) and mode of access.

[38] The BDSG is applicable since the DZHW GmbH is legally a public body of the federal government (cf. § 2 para. 3 BDSG). The federal government possesses an absolute majority of the shares in DZHW GmbH and the institute performs duties of public administration of the federal government in the broadest sense. For interpretation of individual legal aspects the European Data Protection Guidelines can be used as a complement.

Figure 4: Modes of Access, Statistical Degree of Anonymisation and Analytical Potential of the Data of the DZHW PhD Panel 2014



[Statistical Degree of Anonymisation] In the course of anonymisation, all information that directly allows individuals or institutions to be identified is deleted. These so-called *direct identifiers,* such as names, addresses and email addresses, were placed in a separate data set (cf. Chapter 4, Footnote 14)) during the field phase of the PhD Panel 2014 and are neither contained in the CUF nor in the various SUF variants. To further prevent any re-accessing of this information, the original identification number was removed and replaced with a new randomly assigned identification number.

Then the *quasi-identifiers* were specified, i.e. information which, in combination with or by the allusion to external information, is designed to identify a person indirectly.[39] The following quasi-identifiers were identified for the PhD Panel 2014, which are present in both external data sources[40] and the doctorate data: Higher education institution, PhD/doctorate subject, subject/field of study, type of degree, cost of PhD/doctorate, job details, regional information (on higher education institution, place where university entrance qualification was gained, place of work or stays abroad) and personal data (e.g. year of birth, details of own children, nationality and country of birth). In order to prevent a definite allocation of the doctorate data, these key variables were aggregated or deleted according to data product or method of access (see Table 9). For example, in the variable 'place of birth' in the SUF for on-site use, the first three digits of the postal code, the first two digits of the postal code in the remote desktop SUF, the download SUF and the download CUF were assigned to aggregated federal states. Open data are also

[39] It's worth noting that the identification of a person is already made more difficult by the non-participation of other people, as there is some uncertainty as to whether or not an interviewed person demonstrates a unique combination of variables within the population.

[40] E.g. student and examination statistics from the Federal Statistical Office, university alumni networks or even professional networks.

quasi-identifiers (see Ebel, 2015, p. 3) and were either encoded or deleted as part of the anonymisation.

Health information was collected as part of the PhD Panel 2014, for which no additional consent to secondary use was obtained from the respondents. These responses were therefore pooled with the category ‘other’ in the CUF and all SUF versions. In order to ensure the absolute anonymisation of the CUF data, a random sampling of the data was collected (20 per cent of the graduates interviewed).

Finally it was checked whether the data contained *sensitive information*, e.g. on health, sexual orientation or political views. This information, although not suited for re-identification of individuals or institutions, can be used in case of de-anonymisation (cf. Koberg 2016, p. 694). Therefore, its protection is particularly important (cf. §3 para. 9 BDSG, Art. 8 para. 1 and 2a Data Protection Directive [EG-DSRL]). Health information was collected as part of the PhD Panel 2014, for which no additional consent to secondary use was obtained from the respondents. These responses were therefore pooled with the category ‘other’ in the CUF and all SUF versions. In order to ensure the absolute anonymisation of the CUF data, a random sampling of the data was collected (20 per cent of the graduates interviewed).

Table 9: Statistical Anonymisation Measures for the Data of the DZHW PhD Panel By Mode of Access[41]

Characteristic	On-Site SUF	Remote Desktop SUF	Download SUF	Download CUF (Sub-sample)
Direct identifiers	Deletion and assignment of random ID	Deletion and assignment of random ID	Deletion and assignment of random ID	Deletion and assignment of random ID
Questionnaire receipt	Available	Deletion	Deletion	Deletion
Study/doctoral subject	Aggregation to areas of study[a]	Aggregation to areas of study[a]	Aggregation to subject area[a]	Aggregation to subject area[a]
Higher education institution	Aggregation to type of higher education institution[b]	Aggregation to type of higher education institution[b]	Deletion	Deletion
Location of higher education institution	Aggregation to federal states	Aggregation to groups of federal states	Aggregation to groups of federal states	Aggregation to groups of federal states
Award for dissertation	Available	Available	Deletion	Deletion
Further academic qualification (type of degree)	Available	Aggregation to master‘s, state examination, graduate diploma, Bachelor, Master, other	Aggregation to master‘s, state examination, graduate diploma, Bachelor, Master, other	Aggregation to master‘s, state examination, graduate diploma, Bachelor, Master, other
Place of work and place where course entry qualification was	Germany: postcode (digitis 1 to 3) Abroad: availa-	Germany: postcode (digits 1 and 2)	Germany: four federal states shown individually; otherwise aggrega-	Germany: four federal states shown individually; otherwise aggrega-

[41] Detailed information on the anonymised variables can be found in the Data Set Report and the Metadata Search Portal (https://metadata.fdz.dzhw.eu/#!/en).

Characteristic	On-Site SUF	Remote Desktop SUF	Download SUF	Download CUF (Sub-sample)
gained (postcode)	ble	Abroad: Aggregation to world regions[c]	tion to five groups of federal states Abroad: aggregation to world regions[c]	tion to five groups of federal states Abroad: aggregation to world regions[c]
Stays abroad (country)	Available	Available	Aggregation to world regions[c]	Aggregation to world regions[c]
Occupation	Aggregation to occupational main groups[d]	Aggregation to occupational main groups[d]	Aggregation to occupational main groups[d]	Aggregation to occupational main groups[d]
Personnel categories	Available	Available	Aggregation: Scientific research assistant to other; PD to lecturer; lecturer to senior lecturer	Aggregation: Scientific research assistant to other; PD to lecturer; lecturer to senior lecturer
Company size	Available	Available	Aggregation: Wave 1: 1-20; 21-249; 250-1000; 1001 and more Wave 2: 1-19; 20-249; 250-999; 1000 and more	Aggregation: Wave 1: 1-20; 21-249; 250-1000; 1001 and more Wave 2: 1-19; 20-249; 250-999; 1000 and more
Nationality (foreign)	Available	Available	Aggregation to world regions[c]	Aggregation to world regions[c]
German nationality since	Available	Available	Aggregation: until 1989; 1990-1999; 2000-2009; since 2010	Aggregation: until 1989; 1990-1999; 2000-2009; since 2010
Year of immigration	Available	Available	Aggregation: until 1989; 1990-1999; since 2000	Aggregation: until 1989; 1990-1999; since 2000
Year of birth	1961 to 1988 shown individually, otherwise aggregation: until 1949; 1950-1954; 1955-1960; 1989 and younger	1961 to 1988 shown individually, otherwise aggregation: until 1949; 1950-1954; 1955-1960; 1989 and younger	Aggregation: until 1959; 1960-1969; 1970-1979; 1980-1981; 1982-1983; 1984-1985; 1986-1987; since 1988	Aggregation: until 1959; 1960-1969; 1970-1979; 1980-1981; 1982-1983; 1984-1985; 1986-1987; since 1988
Number of children	Available	Available	Top-Coding[e]	Top-Coding[e]
Year and month of birth of children	Available	Year of birth: aggregation: until 1997; 1998-2003; 2004-2009; 2010-2012; since 2013 Month of birth: deletion	Year of birth: (only for the four youngest children) aggregation: until 1997; 1998-2009; since 2010 Month of birth: deletion	Year of birth: (only for the four youngest children) aggregation: until 1997; 1998-2009; since 2010 Month of birth: deletion
Data on children (own child, resident in house-	Available	Available	Only for the four youngest children	Only for the four youngest children

Characteristic	On-Site SUF	Remote Desktop SUF	Download SUF	Download CUF (Sub-sample)
hold)				
Occupation of parents	Aggregation to occupational sub-groups[d]	Aggregation to occupational groups[d]	Aggregation to occupational main groups[d]	Aggregation to occupational main groups[d]
Response categories for health	Pooling with the category 'other reasons'	Pooling with the category 'other reasons'	Pooling with the category 'other reasons'	Pooling with the category 'other reasons'
Other open responses	Deletion	Deletion	Deletion	Deletion

[a] According to the Key List of Student and Examination Statistics Winter Semester 2015/2016 and Summer Semester 2016 from the Federal Statistics Office.

[b] In federal states where the type of higher education institution appears only seldom (< 3 times), the type of institution is anonymised.

[c] According to the Destatis Nationality and Region Classification 2014.

[d] According to German Classification of Occupations from 2010 from the Federal Statistics Office.

[e] Data on four or more children were pooled into a single category.

9 Bibliography

Auriol, L., Felix, B. & Schaaper, M. (2012). *Mapping Careers and Mobility of Doctorate Holders* (OECD Science, Technology and Industry Working Papers, 2012/07). doi:10.1787/5k4dnq2h4n5c-en

Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L. & Klieme, E. (2011). *6 Education processes in life-course-specific learning environments* (2. Aufl.) (Nr. 14). doi:10.1007/s11618-011-0183-6

Beierlein, C., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2014). *Allgemeine Selbstwirksamkeit Kurzskala (ASKU). Zusammenstellung sozialwissenschaftlicher Items und Skalen* (GESIS – Leibniz-Institut für Sozialwissenschaften, Hrsg.). doi:10.6102/zis35

Blickle, G., Kuhnert, B. & Rieck, S. (2003). Laufbahnförderung durch ein Unterstützungsnetzwerk. *Zeitschrift für Personalpsychologie, 2* (3), 118–128. doi:10.1026//1617-6391.2.3.118

Blumenstiel, J. E. & Gummer, T. (2015). Prävention, Korrektur oder beides? Drei Wege zur Reduzierung von Nonresponse Bias mit Propensity Scores. In J. Schupp & C. Wolf (Hrsg.), *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (S. 13–44). Wiesbaden: Springer Fachmedien Wiesbaden. doi:10.1007/978-3-658-10459-7

Brandt, G., Vogel, S. de & Jaksztat, S. (2016). *Entwicklung und Testung eines Instruments zur Erfassung der Lernumwelt in der Promotionsphase. Ergebnisse der Entwicklungsstudie*. : Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW).

Bundesamt für Statistik (Hrsg.). (2011). *Von der Hochschule ins Berufsleben. Erste Ergebnisse der Hochschulabsolventenbefragung 2009* (Statistik der Schweiz: Bildung und Wissenschaft), Neuchâtel

Das Institut für Gründung und Innovation der Universität Potsdam. (2010). *Onlinebefragung des BIEM-CEIP zur Karriereentwicklung von Wissenschaftlerinnen und Wissenschaftlern in Forschungsteams*.

Ebel, T. (2015). *Empfehlungen zur Anonymisierung quantitativer Daten*. Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften.

Egeln, J., Gottschalk, S., Rammer, C. & Spielkamp, A. (2003). *Spinoff-Gründungen aus der öffentlichen Forschung in Deutschland* (Dokumentation Nr. 03-02). Mannheim: Zentrum für Europäische Wirtschaftsforschung GmbH (ZEW).

Grühn, D., Hecht, H., Rubelt, J. & Schmidt, B. (2009). *Der wissenschaftliche „Mittelbau" an deutschen Hochschulen. Zwischen Karriereaussichten und Abbruchtendenzen*. Berlin: ver.di - Vereinte Dienstleistungsgewerkschaft.

Hochfellner, D., Müller, D., Schmucker, A. & Roß, E. (2012). *FDZ-Methodenreport. Datenschutz am Forschungsdatenzentrum* (Nr. 06). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung (IAB).

Jungbauer-Gans, M. & Gross, C. (2013). Determinants of Success in University Careers. Findings from the German Academic Labor Market / Erfolgsfaktoren in der Wissenschaft – Ergebnisse aus einer Habilitiertenbefragung an deutschen Universitäten. *Zeitschrift für Soziologie, 42* (1). doi:10.1515/zfsoz-2013-0106

Koberg, T. (2016). Disclosing the National Educational Panel Study. In H.-P. Blossfeld, J. v. Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys. The example of the National Educational Panel Study* (S. 691–708). Wiesbaden: Springer VS. doi:10.1007/978-3-658-11994-2

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal, 14* (1), 22–59.

Kovaleva, A., Beierlein, C., Kemper, C. J. & Rammstedt, B. (2014). *Internale-Externale-Kontrollüberzeugung-4 (IE-4). Zusammenstellung sozialwissenschaftlicher Items und Skalen* (GESIS - Leibniz-Institut für Sozialwissenschaften, Hrsg.). doi:10.6102/zis184

Lane, J., Heus, P. & Mulcahy, T. (2008). Data access in a cyber world: Making use of cyberinfrastructure. *Transactions on Data Privacy, 1* (1), 2–16.

Otto, K., Glaser, D. & Dalbert, C. (2004). *Hallesche Berichte zur Pädagogischen Psychologie. Skalendokumentation " Geografische und berufliche Mobilitätsbereitschaft"* (Dalbert, C., Hrsg.) (Nr. 8).

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Survey Research Methods Section,* 225–230.

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C. & Kovaleva, A. (2014). *Big Five Inventory (BFI-10). Zusammenstellung sozialwissenschaftlicher Items und Skalen* (GESIS – Leibniz-Institut für Sozialwissenschaften, Hrsg.). doi:10.6102/zis76

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* (1), 41–55. doi:10.2307/2335942

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63* (2), 581–592.

Scherer, S. & Brüderl, J. (2010). Sequenzdatenanalyse. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 1031–1051). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnell, R., Hill, P. B. & Esser, E. (2005). *Methoden der empirischen Sozialforschung* (7. Aufl.). München: Oldenbourg.

Statistisches Bundesamt. (2013). *Hochqualifizierte in Deutschland. Erhebung zu Karriereverläufen und internationaler Mobilität von Hochqualifizierten*, Wiesbaden

Statistisches Bundesamt. (2015). *Bildung und Kultur. Prüfungen an Hochschulen 2014* (Statistisches Bundesamt, Hrsg.). Wiesbaden: Statistisches Bundesamt.

Valliant, R., Dever, J. A. & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York (NY): Springer New York. doi:10.1007/978-1-4614-6449-5

Wissenschaftliches Zentrum für Berufs- und Hochschulforschung. (2009). *Neue Ausbildungsformen – andere Werdegänge? Eine Untersuchung zu Ausbildungs- und Berufsverläufen ehemaliger Doktoranden*. Fragebogen.

Appendix 1: Probit-Regression for Generating the Panel Attrition Weight in Wave 2

	Odds Ratio	Std. Err	z	P>|z|
Gender				
Male	Ref.			
Female	1,06	0,07	0,91	0,365
Not specified	0,80	0,36	-0,49	0,625
Age				
23-30	1,13	0,08	1,58	0,115
31-34	Ref.			
35-39	1,03	0,08	0,36	0,719
40-44	1,21	0,17	1,33	0,185
45-49	2,18	0,44	3,81	0,000
50-60	1,80	0,40	2,67	0,008
61-79	1,14	0,44	0,33	0,744
Not specified	0,14	0,08	-3,55	0,000
Doctoral conditions				
Research Assistant (budget)	0,98	0,08	-0,29	0,775
Research Assistant (external funds)	Ref.			
Structured PhD program	0,77	0,77	-2,18	0,029
Stipend program	0,83	0,09	-1,68	0,092
Individual PhD studies	1,06	0,10	0,62	0,534
Not specified	0,46	0,19	-1,84	0,066
Doctoral result				
summa cum laude	1,19	0,12	1,78	0,074
magna cum laude	1,14	0,09	1,67	0,094
cum laude	Ref.			
satis bene	0,97	0,37	-0,07	0,946
rite	1,00	0,21	0,00	0,997
Not specified	1,11	0,35	0,33	0,742
Place of birth				
Germany	Ref.			
Other country	0,78	0,08	-2,56	0,010
Not specified	0,69	0,16	-1,58	0,113
"I am easy-going, prone to laziness"				
Does not apply at all	0,79	0,79	-2,94	0,003
2	0,87	0,07	-1,77	0,077
3	Ref.			
4	1,27	0,16	1,97	0,048
Strongly applies	0,98	0,23	-0,07	0,946
Not specified	0,28	0,15	-2,32	0,020
Subject group				
Humanities	1,06	0,14	0,42	0,677
Sports	0,82	0,29	-0,56	0,577

Law, Economics, Social Sciences	0,76	0,07	-2,87	0,004
Mathematics, Natural Science	Ref.			
Medicine/Public Health	0,63	0,08	-3,7	0,000
Agricult. Sc., Forestry, Nutrition, Veterinary med.	1,05	0,16	0,31	0,756
Engineering	0,68	0,07	-3,59	0,000
Arts	1,28	0,36	0,86	0,392
Not specified	0,69	0,08	-3,13	0,002
Current/last gross salary				
Up to 3200€	1,05	0,08	0,66	0,507
3201€ to 5000€	Ref.			
More than 5000€	1,08	0,09	0,94	0,348
No income	1,40	0,26	1,84	0,065
Not specified	0,61	0,07	-4,37	0,000
Industry sector				
Agriculture, Fishery, Energy, Water management	0,91	0,21	-0,42	0,675
Chemical industry	0,62	0,09	-3,14	0,002
Mechanical engineering, Vehicle construction	0,60	0,10	-3,03	0,002
Electrical engineering, Electronics, IT equipment	0,69	0,14	-1,85	0,065
Other manufacturing industries	0,69	0,14	-1,77	0,077
Trade, Banks, Insurance companies	0,57	0,12	-2,67	0,007
Software development	1,04	0,22	0,19	0,848
Legal, Business, Human resources consulting	0,57	0,09	-3,38	0,001
Media, Publishing, Advertising	0,47	0,12	-2,89	0,004
Healthcare	0,79	0,09	-2,06	0,040
Other services	0,89	0,14	-0,75	0,452
Schools	0,85	0,20	-0,72	0,474
Higher education institutions	Ref			
Research institutions	1,03	0,13	0,19	0,849
Other education, research, culture	1,25	0,35	0,79	0,429
Public administration	0,77	0,77	-1,57	0,116
Other non-profit organizations	0,75	0,16	-1,35	0,177
Not specified	0,55	0,09	-3,74	0,000
Constant	2,16	0,29	5,77	0,000

N=5412; Pseudo R^2=0,0344