FDZ-Arbeitspapier
Nr.8

Rainer Lenz
Hans-Peter Hafner
Daniel Schmidt

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Scientific analyses using
the Continuing Vocational
Training Survey 2000

2006

FDZ-Arbeitspapier
Nr.8

Rainer Lenz
 Hans-Peter Hafner
 Daniel Schmidt

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Scientific analyses using
the Continuing Vocational
Training Survey 2000

2006

Scientific analyses using the
Continuing Vocational Training Survey 2000

Rainer Lenz, Hans-Peter Hafner and Daniel Schmidt[1]

1. Introduction

Since the end of last year, scientists may use official statistics data on continuing vocational training at enterprises for their own analyses. In a joint project, the Federation's and the federal states' statistical offices anonymised those basic data of the Second European Continuing Vocational Training Survey (CVTS 2) in 2000 for 1999 as the reference year to make them meet the strict confidentiality requirements stipulated by law, on the one hand, ensuring that these data were still offering sufficient potential for scientific analysis, on the other hand.

This data set, which is referred to as "scientific use file", containing data on some 3200 German enterprises, has been compiled at short notice in reply to a request recently expressed by scientists, who wished to get access to data on continuing vocational training. The efforts to compile scientific use files, such as e.g. data of the Microcensus, the Income and Consumption Sample Survey and the Time Budget Survey, are supported by the Federal Ministry for Education and Research in the context of its own efforts to improve the informational infrastructure between science and statistics in Germany. The scientific use file on CVTS 2 finds its first application in what is called a competition of expertise, initiated

---
[1] Rainer Lenz, Federal Statistical Office of Germany, Research Data Centre, Gustav-Stresemann-Ring 11, 65180 Wiesbaden (rainer.lenz@destatis.de); Hans-Peter Hafner, Statistical Office of Hesse, Research Data Centre, Rheinstraße 35-37, 65175 Wiesbaden (hhafner@statistik-hessen.de); Daniel Schmidt, Federal Statistical Office of Germany, Education, Research and Development, Culture, Justice, Gustav-Stresemann-Ring 11, 65180 Wiesbaden (daniel.schmidt@destatis.de)

and already launched by the Council for Social and Economic Data on the subject of "Education in Working Life".[2]

In 1987, article 16, paragraph 6 of the Federal Statistics Law3 gave science a privileged access to official statistics microdata. Accordingly, individual data transmission to science is allowed, if these data can be re-identified only with an unreasonable amount of time, cost and labour force (factual anonymity). Here, "unreasonable" means that the expense involved in re-identifying the data exceeds their utility. It implies that in a factually anonymous set of data there is no need to exclude a possible deanonymisation risk of individual data with absolute safety, as it would be unattractive for a potential data intruder to make an attempt of disclosing the data. This paper presents a sufficiently anonymised file for scientific purposes (a so-called scientific use file), generated from CVTS 2 data (i.e. data from the Second Continuing Vocational Training Survey in Europe) undertaken in 2000 with 1999 as the reference year. This product has evolved from a cooperation project between Hessisches Statistisches Landesamt (Statistical Office of Hesse) and the Federal Statistical Office of Germany.

2. Basic material

The survey collected data from 3184 enterprises with more than 10 employees in the economic sectors C-K and O of the NACE rev.1 on their employees' participation in continuing vocational training measures in 1999.

The data contain information on the various forms offered in terms of continuing vocational training, about participants, hours of instruction and the cost involved (in relation to tuition classes) as well as qualitative data about the conceptual approach to continuing vocational training and the importance that the respective enterprise attaches to such training. We

---

[2] You will find more information on the internet at http://www.ratswd.de/wettbew.htm.
[3] Law on Statistics for Federal Needs (Federal Statistics Law – BStatG) of 22 January 1987 (Federal Law Bulletin I pp. 462, 565) as amended for the last time by article 16 of the law of 21 August 2002 (Federal Law Bulletin I p. 3322).

were, in particular, successful in our efforts to make sure that the data, which had been anonymised, remained suitable for a scientific treatment of relevant questions concerning economic sectors and employee size classes. Further information on the basic material and a full list of variables can be found in (Egner, 2002).
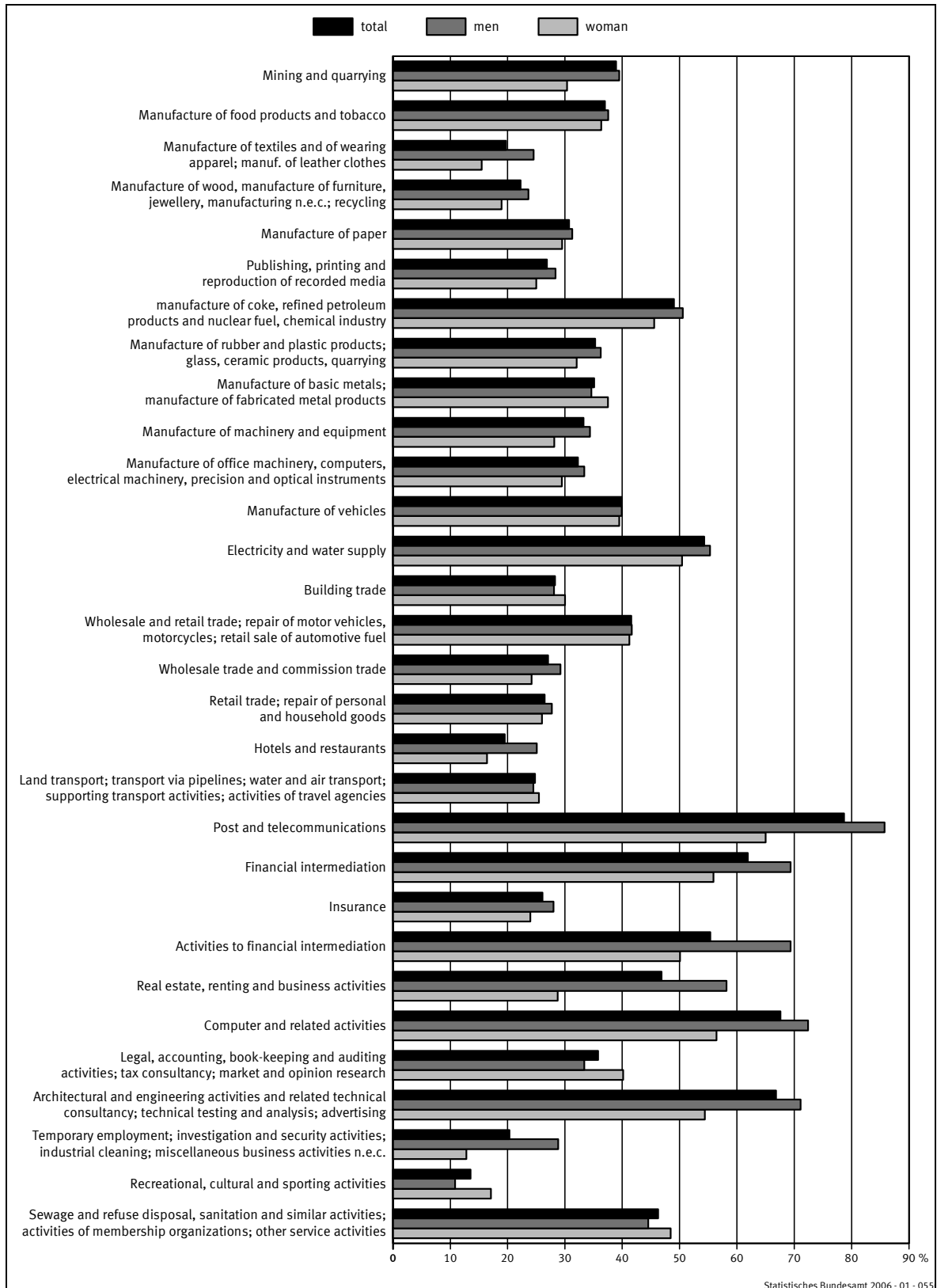
In 1999, tuition courses of continuing vocational training were attended by employees of 67% of the enterprises covered by the survey. All enterprises of the credit, insurance and computer industry, which were covered by the survey, were offering such courses, whereas in the textile and clothing industry that share was as low as 42%, in mining 46% and in transport and traffic 50%.

36.2% of people employed by the enterprises interviewed were participating in tuition courses (in relation to enterprises offering such courses). Men's attendance rate was 38.2%, clearly exceeding that of women (32.7%). But here again there were very strong discrepancies between the various industries. A total attendance rate of as little as 13.5% or so was found in fields such as culture, sports and entertainment (however, since only a few enterprises have responded to the survey, this value need not necessarily be representative), whereas in telecommunications it was up to about 78.6%. In the metal-working industry, women's and men's attendance rates were almost contrary to those of all other enterprises (women: 37.5%, men: 34.6%) and also in some other industries, such as the building trade and in legal, tax and business consultancy, in market and opinion research, the attendance rate of women was above

Figure 1: Enterprises providing continuing vocational training by branches of economic activity

Statistische Ämter des Bundes und der Länder, Forschungsdatenzentren, Arbeitspapier Nr. 8
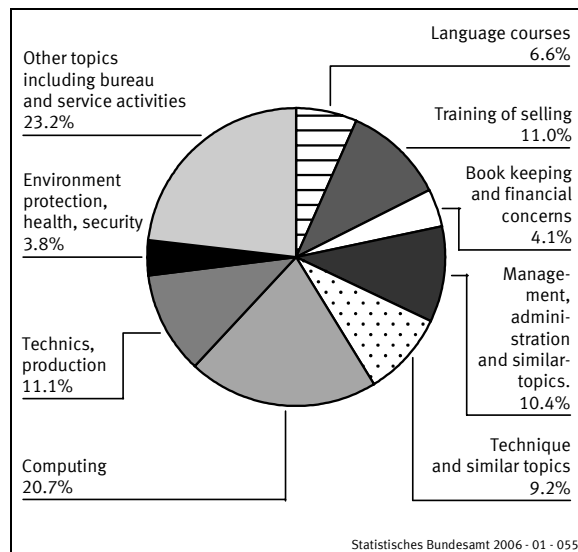
Figure 2: Participation at continuing vocational training by gender and economic activity

that of men, whereas in the temporary employment of personnel, investigation and security services, industrial cleaning and miscellaneous services for companies the share of men attending tuition courses was more than twice as high as the respective share of women (28.9% to 12.8%).

Figure 3: Length of participation at continuing vocational training by topics

Statistisches Bundesamt 2006 - 01 - 055

Figure 4: Methods of continuing vocational training applied in enterprises

Statistisches Bundesamt 2006 - 01 - 055

Measured in terms of the proportion of attendance hours in tuition courses by theme, computing were clearly on top with 20.7%, followed by technology and production with 11.1%, sales training with 11.0%, and management, administration etc. with 10.4%.
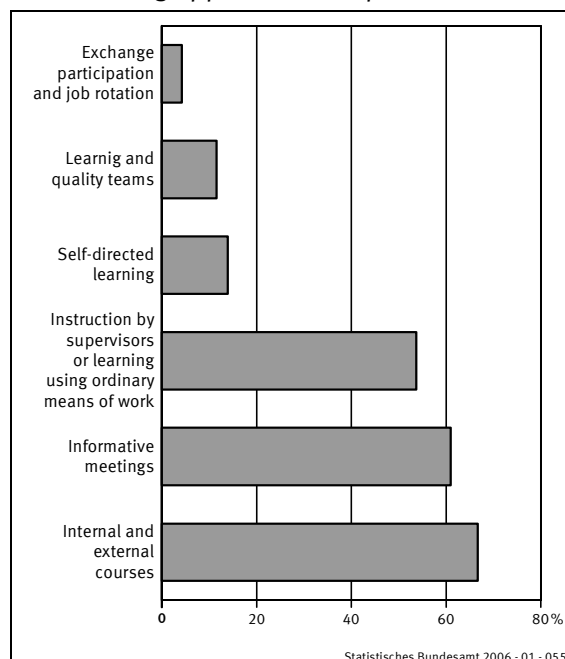
Environmental protection, health and security taken together had a proportion of merely 3.8%; however, in mining and quarrying these themes accounted for 18.4% of the time spent on training. It is little surprising that the share of continuing vocational training on computers was highest in the data processing industry with 65.9% and it is not entirely unexpected either that sales training accounted for 42.2% of the time spent on continuing training courses in the insurance trade.

Widespread methods of continuing vocational training other than tuition courses included information meetings (at 61.0% of enterprises) and instruction given by superiors or normal learning on the job (53.7%). However, self-controlled learning (13.9%), learning and quality circles (11.5%) as well as exchange programmes and job rotation (4.2%) played a minor role.

3. Risk potential analysis

Although some protection is derived form the fact that CVTS 2 is a sample survey, there is a need for further safeguards in compiling scientific use files, which enable scientists to access data outside the protected rooms of official statistics. This is equally important from a position of maintaining the faith of participating enterprises in official statistics and of motivating them for participation in future surveys.

As extra knowledge about continuing vocational training courses of enterprises is not accessible in a systematic form to a potential data intruder, the utmost imaginable risk in respect of these variables seems to consist in individual attacks, i.e. individual investigations about a specific enterprise. However, this would require a lot of inside information about that enterprise and it would hardly make new kinds of information available. The one really critical overlapping variable to commercial databases is the number of employees. However, since there are large discrepancies between data from different sources on this variable and since the lower and medium-sized employment categories include very many enterprises, the employment figures of which lie in a relatively close proximity to each other, the risk will be biggest actually for the largest enterprises.

4. Anonymisation methods

The methods described below attach particular importance to coarsening those categorical variables that a potential data intruder might use as overlaps with confidential CVTS 2 data.

Regional breakdown

No regional information is made available. Regional data is particularly susceptible to re-identification. That is why the presence of such variables in a scientific use file would raise a big problem in terms of anonymisation. Moreover, the low case number in the new federal states (585 enterprises) would not allow achieving reliable results separately for old and new federal states, so that the omission of the regional variable does not imply a major restriction of the potential for analysis.

Branch of activity classification

The point of departure is given by the 30 economic sectors of the NACE 30 classification, which were used for drawing the sample for the survey.

On the basis of expansion factors some economic sectors were revealed as particularly vulnerable. Therefore, they had to be pooled together with other economic sectors as follows:

1. Activities linked to credit and insurance trade were pooled together with insurance.
2. Mining and quarrying were pooled together with coking plant, refined petroleum processing and chemical industry.
3. Telecommunications were pooled together with transport.
4. Trade and repair of motor vehicles were pooled together with retail trade and repair of consumer goods.
5. The industry of wood, manufacture of furniture, jewellery, musical instruments, toys, sports goods and other goods was pooled together with the industry of paper.

Expansion factors for pooled economic areas were newly calculated. They were obtained as quotients of the number of enterprises in the pooled sector of the total population and of the number of enterprises in the pooled sector of the survey population. The sample of the services sector generally included a very low number of enterprises. As there was no risk of

re-identification in that case, the decision of whether or not it would be necessary to pool some items together for analysis could be left to the scientists.

Table 1 below shows the list of industries included in the anonymised file (basically the so-called European NACE 30 classification), together with the respective case numbers:

Table 1: Distribution of CVTS 2 enterprises by economic sector

Economic classification	NACE 30 specification	Number	Population
Manufacture of food products and tobacco	02	184	16 886
Manufacture of textiles and of wearing apparel; dressing and dyeing of fur; manuf. of leather clothes	03	193	3 549
Publishing, printing and reproduction of recorded media	06	125	5 737
Manufacture of rubber and plastic products; glass, ceramic products, quarrying	08	148	8 207
Manufacture of basic metals; manufacture of fabricated metal products	09	188	16 107
Manufacture of machinery and equipment	10	96	10 449
Manufacture of office machinery and computers; manuf. of electrical machinery and apparatus; manuf. of precision and optical instruments	11	77	10 658
Manufacture of vehicles	12	163	1 806
Electricity and water supply	13	228	1 363
Building trade	14	202	61 083
Wholesale trade and commission trade	16	115	20 727
Hotels and restaurants	18	125	16 543
Financial intermediation	21	198	2 788
Real estate, renting and business activities; R&D	24	22	27 940
Computer and related activities	25	22	4 169
Legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling; business and management consultancy; holdings	26	25	15 167
Architectural and engineering activities and related technical consultancy; Technical testing and analysis; Advertising	27	26	9 559
Labour recruitment and provision of personnel; Investigation and security activities; Industrial cleaning; Miscellaneous business activities n.e.c.	28	17	7 939
Recreational, cultural and sporting activities	29	16	8 650
Sewage and refuse disposal, sanitation and similar activities; activities of other membership organizations; other service activities	30	30	15 765

Mining and quarrying; manufacture of coke, refined petroleum products and nuclear fuel, chemical industry	01 and 07	254	2 746
Manufacture of wood, manufacture of furniture, jewellery, manufacturing n.e.c.; recycling; manufacture of paper	04 and 05	253	9 710
Wholesale and retail trade; repair of motor vehicles, motorcycles; retail sale of automotive fuel; Retail trade, except of motor vehicles and motorcycles; except of retail sale of automotive fuel); repair of personal and household goods	15 and 17	221	48 343
Land transport; transport via pipelines; water and air transport; supporting transport activities; activities of travel agencies; post and telecommunications	19 and 20	204	16 138
Insurance; activities auxiliary to financial intermediation	22 and 23	52	905

Employees of the enterprise

The absolute number of employees was not shown for enterprises with more than 2000 employees as of 31 December 1999, instead, a remark was made that the enterprise had more than 2000 employees. Furthermore, the numbers of male and female employees of those enterprises were given merely in terms of percentages of the total number of employees and instead of the number of employees at the end of 1998 we gave the percentage change from 1998 to 1999. In the case of variables that depended on the number of employees (hours worked, personnel costs, expenses, hours attended, participants) the original values were recalculated into per-capita values (division by the number of employees as of end of 1999) or into values per participant. 83 enterprises or nearly 3 % of all cases were affected by these changes. The threshold chosen was 2000 employees, taking into account that for enterprises with a maximum of 2000 employees the distance from the respective next bigger or next smaller enterprise was always less than 3 % in terms of the number of employees. A difference, which is so small, involves a very high risk of possibly wrong identification. In addition, there were often clear discrepancies from comparable data in commercial databases, not at least because of the fact that apprentices and trainees were not counted as employees in the CVTS 2 survey.

Further anonymisation methods

Some of the variables were removed from the data set or modified. This concerned, in particular, the following variables:

Annual average employees 1999: These figures were to be submitted by enterprises with strong seasonal oscillations of employment only. Merely 76 enterprises were supposed to indicate these values. Instead, a "season" variable is shown, with season = 1, if there are seasonal oscillations of employment, and season = 0 in other cases.

Percentage of indirect cost in total personnel costs: Removed because there are only isolated assured figures.

Number of persons fully or partially employed with tuition courses: Removed because there are only isolated assured figures.

Funds to which contributions have been paid for continuing vocational training: These variables were omitted, since the number of "yes" replies was as low as 77 for regional funds, 7 for national funds and 47 for other funds.

Receipts, sources of receipts, balance: These variables were omitted, as the data set contained just 57 enterprises, which reported receipts from training courses. If these variables are known to a potential data intruder, such knowledge, in combination with other overlapping variables (industry and employment data), might identify specific cases unambiguously. Therefore, we calculated a new variable, which adopted the value 1, if an enterprise had receipts from training courses, and 0 in other cases.

All in all, after the application of anonymisation methods some 180 variables have been retained in the data set.

5 Assessing the effect of disclosure protection

To measure the effect of disclosure protection, the Federal Statistical Office has developed Destatis-Anonymeter, a simulation software tool to implement a so-called database cross match scenario. The theoretical description and functionality of the software can be found in (Lenz, 2005). In a database cross match scenario a data intruder tries to identify as many units as possible in an external database, which match the target data (confidential anonymised data). Simulations were made by using MARKUS, a commercially available database, as a possible external database to be used by a potential data intruder. What we found out, first of all, were the re-identification effect inherent in data on an enterprise's location (regional information), mentioned in the preceding chapter, and a necessity to pool some industries together. To assess the protective effect of the anonymisation measures, we made use of the distribution obtained as a result of simulation, showing the risk of disclosure with regard to the industries described above and seven employee size classes[4], taking into account the underlying sample, thus implying that in many cases a potential data intruder was supposed to be aware of a specific enterprise's participation in the survey. All of the data attack simulations made with data anonymised on a trial basis have shown that a disclosure of units could only be achieved at an unreasonably high expense and that it would be affected by a lot of insecurity for the data intruder. That means that the data, converted into a scientific use file, can be made available to scientists.

6 Final remarks

Since the Federal Statistics Law entered into force in 1987, scientists have enjoyed a privileged position among the users of federal statistics data, the so-called science privilege. It consists in a provision, allowing the transfer of microdata rendered de facto anonymous by the Federal Statistical Office and statistical offices of the federal states to universities or other institutions, which are entrusted with independent scientific research,

---

[4]The following categories were used: 10-19 employees, 20-49, 50-99, 100-249, 250-499, 500-999, 1000 and more employees.

for the purpose of conducting scientific work. Such use of the data is very advantageous to scientists, as they receive the micro data for further evaluation at their own workplaces, e.g. see (Zühlke, Zwick, Scharnhorst, Wende, 2004). Another advantage of a Scientific Use File is the guarantee that the data material is one and the same even in the case of different users. This makes it easier for scientists to build research networks and strengthens the principle that scientific findings should be verifiable.

At present, the Research Data Centre of the statistical offices of the Federation and the federal states is working on a so-called Campus File of the CVTS 2 survey, i.e. a set of data for broad use, which will be available to everybody. The Scientific Use File presented in this essay costs 65 Euros and is available for scientific purposes via the Research Data Centres of the statistical offices of the Federation and the federal states. The applications required for that purpose are available as downloads on the internet at www.forschungsdatenzentren.de.

References

Egner, U. (2002), "Continuing vocational training at enterprises CVTS 2" (german), Federal Statistical Office of Germany, project report.

Lenz, R. (2005), "Measuring the disclosure protection of micro aggregated business microdata – An analysis taking the example of German Structure of Costs Survey", to appear in: *Journal of Official Statistics*, Sweden.

Zühlke, S., Zwick, M., Scharnhorst, S., Wende, T. (2004), "The research data centres of the Federal Statistical Office and the Statistical Offices of the Länder", *Journal of Applied Social Science Studies (Schmoller's Jahrbuch)*, 124, pp. 567-578.