Content

methods, data, analyses | Vol. 8(2), 2014, pp. 125-136 DOI: 10.12758/mda.2014.005

PIAAC and its Methodological Challenges

Beatrice Rammstedt[1] *& Débora B. Maehler*[1,2]

[1] *GESIS – Leibniz Institute for the Social Sciences*
[2] *College for Interdisciplinary Education Research (CIDER)*

Abstract

This article gives an overview of the *Programme for the International Assessment of Adult Competencies* (PIAAC) and introduces the methodological challenges in implementing the survey – especially those encountered in Germany. Adherence to high methodological standards is a prerequisite to participation in PIAAC and to inclusion of the national data of the respective participating countries in the international dataset (OECD, 2010). Depending on the standard in question, and on national circumstances, compliance is a challenging undertaking. This Special Issue discusses methodological challenges at different levels, and steps taken to implement PIAAC standards in Germany. The aspects addressed include sample design, survey instruments, field work preparation, data collection, and estimation standards. In this introductory article, we outline the central elements of the PIAAC design and the methodological challenges of the survey, and we present the other six articles in this Special Issue.

Keywords: PIAAC, methodological standards, competencies, basic skills, Germany

The Programme for the International Assessment of Adult Competencies (PIAAC)

The Programme for the International Assessment of Adult Competencies (PIAAC) aims to assess basic skills of the adult population in an internationally comparable way. The skills assessed – literacy, numeracy, and problem solving in technology-rich environments – are considered to be essential for successful participation in modern society and to be a foundation for developing numerous other, more specific, skills and competencies (OECD, 2013a). PIAAC provides information about the extent to which the adult population in the respective participating countries differs in terms of the basic skills assessed. Moreover, it examines factors associated with the acquisition, retention, and maintenance of these skills, and sheds light on their effects on social and, in particular, economic participation.

The PIAAC Design

PIAAC was initiated by the Organisation for Economic Co-operation and Development (OECD) and is steered by the PIAAC Board of Participating Countries. Twenty-four countries, including Germany, participated in the first round of PIAAC, which started in 2008. Results were published in 2013. In Germany, PIAAC was implemented by GESIS – Leibniz Institute for the Social Sciences and funded by the Federal Ministry of Education and Research (BMBF) with support from the Federal Ministry of Labor and Social Affairs (BMAS). GESIS was also part of the international consortium commissioned by the OECD to design PIAAC and supervise its implementation in the participating countries. As a PIAAC Consortium partner, GESIS supported the development of the PIAAC background questionnaire. The institute was also responsible for validating the background questionnaire and developing guidelines for its translation.

PIAAC is designed to be repeated at regular intervals. The currently published round, PIAAC 2012, marked the starting point of this multi-cycle program. Further cycles are planned at ten-year intervals, which will enable future changes in adult skills to be monitored and analyzed. As mentioned above, twenty-four countries participated in Round I of the first cycle of PIAAC. A second round, which started in 2012, includes nine additional countries. First results for these Round II countries are expected to be published in 2016. Just this year (2014), the OECD initiated a third round of the first cycle of PIAAC with presumably another five additional

Direct correspondence to
Beatrice Rammstedt, GESIS – Leibniz Institute for the Social Sciences,
PO Box 12 21 55, 68072 Mannheim, Germany
E-mail: beatrice.rammstedt@gesis.org

Figure 1 Participating countries in the three rounds of the first cycle of PIAAC

countries. Figure 1 shows the geographic distribution of the participating countries in the three rounds.

In PIAAC Round I, at least 5,000 randomly selected respondents between the ages of 16 and 65 were interviewed and assessed in each participating country. In Germany, approximately 5,400 interviews were conducted using a two-stage register-based sampling technique. The survey was carried out as a personal interview (background questionnaire) with a subsequent skills assessment. Together, the background questionnaire and the assessment of skills took between 1 1/2 to 2 hours to administer. After answering the background questions, respondents independently completed a computer- or paper-based version of the assessment in the presence of the interviewer (see Zabal et al., 2014).

The Basic Skills Assessed in PIAAC

PIAAC focuses on the assessment of three central basic skills, namely literacy, numeracy, and problem solving in technology-rich environments. Literacy is defined as the ability to understand, use, and interpret written texts. Hence, it is a prerequisite to developing one's knowledge and potential and successfully participating in modern society (Jones et al. 2009; OECD, 2013a; Zabal et al., 2013). The literacy domain in PIAAC includes tasks such as reading and understanding a medication package insert or a brief newspaper article. In addition, there are tasks that involve digital media, for example reading an online job posting. Numeracy

refers to the ability to access, use, and interpret everyday mathematical information in order to manage the mathematical demands of adult daily life (Gal et al., 2009; Zabal et al., 2013). This is measured, for example, with items involving the evaluation of a special offer or the interpretation of numerical information in figures and tables.

PIAAC marks the first time that problem solving in technology-rich environments has been assessed in an international survey (OECD, 2013a). Problem solving in technology-rich environments is defined as the ability to successfully use digital technologies, communication tools, and networks to search for, communicate, and interpret information (Rouet et al., 2009; Zabal et al., 2013). In the first cycle of PIAAC, this domain focuses on the ability to access and make use of information in a computer-based environment. Tasks include sorting and sending e-mails, filling out digital forms, and evaluating the informational content and the credibility of a number of different websites.

The construct definition and item development of each of the three competence domains was based on a theoretical framework developed by renowned experts in each field. The quality and appropriateness of the items was thoroughly tested before the PIAAC Main Survey. For all three domains, results are presented in the form of proficiency scales based on Item Response Theory models (OECD, 2013b). To facilitate the interpretation of the resulting scale scores, each scale was divided into skill proficiency levels with 50-point intervals (similar to other scales with 50-point intervals used in studies such as PISA). This results in five skill proficiency levels for both the literacy and numeracy domains and three skill proficiency levels for the problem solving in technology-rich environments domain. In addition, the area below the lowest level is classified as "Below Level I" (OECD, 2013b; Rammstedt, 2013).

The PIAAC Background Questionnaire

The background questionnaire used in PIAAC was developed by the PIAAC Consortium in cooperation with a Background Questionnaire Expert Group. Based on a framework specifying the analytical underpinnings (OECD, 2011), the development of the background questionnaire was guided by three additional criteria: first, it should possess analytical utility, especially in combination with the competence measures; second, it should provide internationally comparable data; and third, completion time should not exceed 45 minutes, on average. The Consortium developed a source version of the background questionnaire in English, which had to be adapted and translated by each country. An initial – longer – version of, the background questionnaire was tested in the PIAAC Field Test. Based on the empirical findings of the Field Test, the extent to which the aforementioned criteria were met

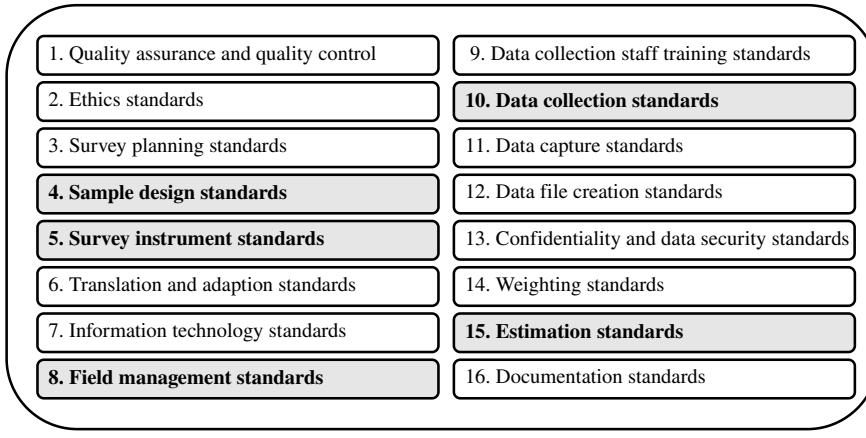
was investigated for each item. The resulting background questionnaire used in the Main Survey covers areas such as socio-demographic information, education and training, and questions relating to the respondent's work and background (Zabal et al., 2014).

Two of the most crucial pieces of information provided by the background questionnaire are the highest level of educational attainment and the current educational engagement of the respondent. Not surprisingly, given their postulated learnability, the competencies assessed in PIAAC are highly dependent on education. In Germany, for example, respondents with the highest level of education possess literacy skills that are, on average, 1.5 proficiency levels higher than those without any formal educational qualification (Maehler et al., 2013).

One of the several innovative aspects of PIAAC is the assessment of the job requirements – that is, the level of skills people need to carry out their everyday work. Based on this information, a central intended outcome of PIAAC was an estimate of the match, or mismatch, between the personal skills assessed in PIAAC and the skills used in the workplace. However, the originally intended measure for skill mismatch, which was also used by the OECD in its initial report on the PIAAC data (OECD, 2013), proved to be problematic (see Klaukien et al., 2013). Therefore, in their article in this Special Issue, Perry, Wiederhold, and Ackermann-Piek seek an alternative, more valid, measure for skill mismatch based on the PIAAC data. They also systematically compare existing and newly developed skill-mismatch measures in a Mincer regression (Mincer, 1974) and investigate the importance of skill mismatch for individual earnings.

PIAAC's Methodological Standards

In order to provide high quality data that allow policymakers and scientists to draw reliable conclusions, PIAAC aims to meet the highest quality standards. This is especially true of the sample design and the survey operations implemented in the various countries. Therefore, detailed Technical Standards and Guidelines (TSG; OECD, 2010), which span almost 200 pages, were developed for the implementation of PIAAC. An overview of the main aspects addressed in these standards and guidelines is given in Figure 2. When developing the TSG, the PIAAC Consortium closely adhered to existing, scientifically recognized best practices and gold standards. One of the main sources was the set of standards developed for the European Social Survey (ESS; European Social Survey, 2012). For example, in accordance with these standards, the target response rate for PIAAC was set at 70%, and the minimum response rate at 50%. Diverging from the ESS procedure, however, the inclusion of countries in the international data set is directly dependent on compliance with these criteria. Countries reaching response rates below 50% in PIAAC



Notes. The highlighted fields indicate standards whose implementation in PIAAC in Germany posed methodological challenges that will be addressed in this Special Issue. Data source: OECD 2010.

Figure 2 Overview of the methodological standards of PIAAC 2012

are included in the data set only if their national data have a low nonresponse bias (OECD, 2013b).

As already described in the response-rate example, each participating country is required to follow all standards and guidelines formulated in the TSG and to document any deviation caused by factors such as national requirements or circumstances. For example, PIAAC could not be fielded in the region of Fukushima in Japan, as the area was highly contaminated with radiation at the time. This resulted in higher undercoverage in Japan than the allowed maximum of 5%. Strict adherence to the guideline whereby cases for interview validation should be randomly preselected, including cases finalized as nonresponse (Guideline 10.9.3A in OECD, 2010, p. 159), was extremely challenging for Germany because re-contacting adamant refusers is not allowed under German law.

Before data release and the publication of the international PIAAC results, the quality of the data of each participating country is investigated and assessed. When the first results from PIAAC were published in 2013, compliance with the quality standards had been certified and confirmed for 23 of the 24 countries that participated in PIAAC Round I. Only at a later point in time did the OECD confirm that the Russian Federation had met the quality standards, despite the fact that some data abnormalities had been identified (cf. OECD, 2013b). These abnormalities led, for example, to the exclusion of the Moscow municipal area from the Russian data.

PIAAC's Methodological Challenges

As described in the last section, PIAAC aims to meet very high methodological standards. Adherence to these standards is crucial to each country's inclusion in the data set and the comparative analyses, thereby enabling it to justify its participation in PIAAC. Depending on the standard in question, and on the national circumstances, meeting these standards is a challenging undertaking. In Germany, too, traditional methods of field work preparation, organization, implementation, and monitoring had to be rethought against the background of the PIAAC TSG. Zabal's article in this Special Issue describes important fieldwork measures and procedures for the PIAAC Main Survey in Germany, and describes how some of these required adaptations with regard to the PIAAC standards. Based on the experience with the PIAAC fieldwork in Germany, the author reflects on the limitations and possibilities posed by international survey operation standards in national implementation.

One standard that proved surprisingly challenging for all countries was the technical requirements for the competence assessment (see Standard 7.1.1 in OECD, 2010). The competence assessment is computer-based by default. Only if the respondent is unable or unwilling to complete the assessment on the computer, is a paper-based assessment administered (OECD, 2013b). However, the items and the virtual machine that displays them were developed for a laptop screen format (4:3) that was already outdated by the time PIAAC was fielded. To meet the standards and to guarantee sufficient resolution and size of the displayed items, 17-inch laptops had to be purchased for all interviewers in Germany. As the laptops, together with all additional material (extra battery, testlets etc.), were comparatively heavy, interviewers in Germany were equipped with wheeled suitcases.

As described above, new and challenging procedures for controlling interviewer performance must be followed.[1] For example, PIAAC TSG (Standard 10.9.5 in OECD, 2010) requires participating countries to review tape recordings of each interviewer's work. If the review reveals performance problems, intervention- and interviewer-retraining measures must be implemented. In their article in this Special Issue, Ackermann-Piek and Massing report on the use of these audio-recorded interviews, and describe interviewers' actual behavior with regard to standardized interviewing techniques and correlations between this behavior and interviewer characteristics.

From a German point of view, the greatest challenge posed by the PIAAC TSG was to reach the minimum response rate of 50% (see Guideline 4.7.4B in OECD, 2010). This is due to the fact that, for years now, response rates in such register-based face-to-face surveys have been dramatically decreasing in Germany – they are usually around 40%, or even lower (cf. European Social Survey, 2012; Wasmer,

1 For an overview of the requirements with regard to interview validation see Massing, Ackermann, Martin, Zabal, Rammstedt, 2013.

Scholz, & Blohm, 2010; Zabal et al., 2014). In order to achieve this challenging goal, and to thereby ensure the inclusion of the German data in the international data set, numerous measures were taken, including, for example, the payment of an attractive incentive to the respondents. In their article in this Special Issue, Martin, Helmschrott, and Rammstedt describe the incentive experiment conducted within the framework of the German PIAAC Field Test to determine the optimum amount of the incentive to be used in the Main Study.

The various measures taken when fielding PIAAC in Germany proved to be successful. In the end, a response rate of 55% was achieved – a figure that had not been reached in such surveys in Germany for years, or even decades. However, the PIAAC TSG (Standard 4.7.6 in OECD, 2010) requires all participating countries with response rates below 70% to conduct extensive nonresponse-bias analyses to prove that this bias was of an acceptable size. In their article in this issue, Helmschrott and Martin report selected results of these nonresponse analyses from the PIAAC Main Study with a special focus on the identification of the main factors influencing survey participation in PIAAC Germany.

Besides the challenges posed by the PIAAC methodological standards, the implementation of PIAAC in Germany faced another major challenge as an error occurred during sampling. Due to this error, people no longer had the same probability of inclusion in the sample (for details, see Zabal et al., 2014). In order to estimate the selection probability of each element of the sampling frame post hoc, an innovative simulation approach was developed and implemented by Gabler, Häder, and Kolb. This approach is described in detail in their article in this Special Issue.

In addition to all these methodological issues, the biggest challenge that countries faced when conducting PIAAC was the very tight timeline. Even though the deadline was extended by a further six months, the time allocated to perform the various tasks was hardly enough. For example, when preparing the national report of the PIAAC data, which was published on the internally fixed date – October 8th 2013 – we received the data of one of the 23 countries only one week before sending the manuscript to the printers. The tight timeline (see Figure 3) was most probably due to the fact that PIAAC is a newly developed and methodologically innovative study. The international design was developed and implemented in parallel with the preparation of the national implementation of PIAAC. In the light of this situation and the constraints and challenges it caused, it is impressive that all countries were able to adhere to this timeline and to meet the methodological requirements.

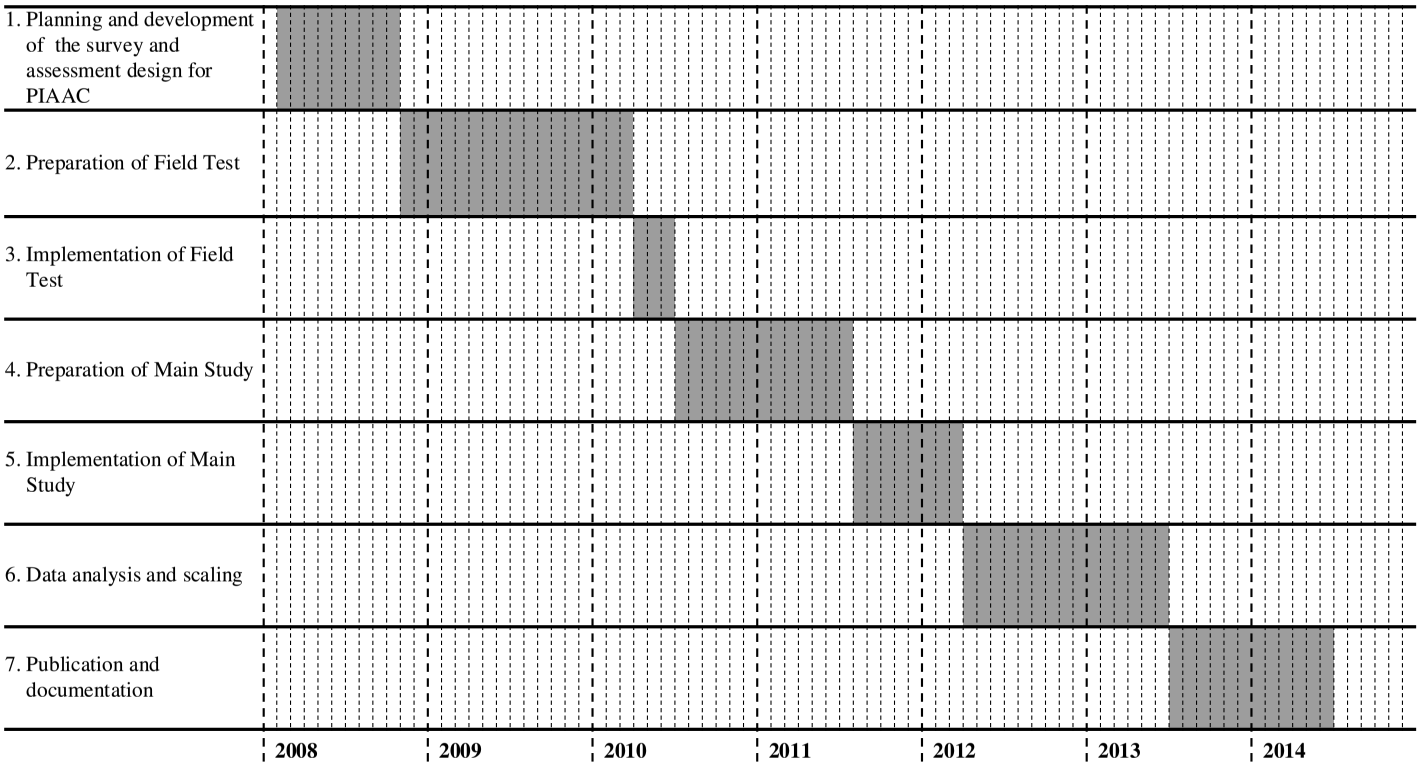


Figure 3 Rough timeline for the preparation and implementation of PIAAC 2012

References

Boudard, E., & Jones, S. (2003). The IALS approach to defining and measuring literacy skills. *International Journal of Educational Research*, 39(3), 191-204.

European Social Survey. (2012). *ESS5-2010 documentation report. The ESS Data Archive*. (3.0 ed.). Bergen: Norwegian Social Science Data Services. Retrieved March 2014, from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_ documentation_report_e03_0.pdf

Gal, I., Alatorre, S., Close, S., Evans, J., Johansen, L., Maguire, T., … Tout, D. (2009). PIAAC numeracy: A conceptual framework. *OECD Education Working Papers* No. 35. Paris: OECD Publishing.

Jones S, Gabrielsen E, Hagston J, Linnakylä P, Megherbi H, Sabatini J., … Vidal-Abarca E. (2009). PIAAC literacy: A conceptual framework. *OECD Education Working Papers* No. 34. Paris: OECD Publishing.

Klaukien, A., Ackermann, D., Helmschrott, S., Rammstedt, B., Solga, H., & Wößmann, L. (2013). Grundlegende Kompetenzen auf dem Arbeitsmarkt. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012* (pp. 127-166). Münster: Waxmann.

Maehler, D. B., Massing, N., Helmschrott, S., Rammstedt, B., Staudinger, U. M., & Wolf, C. (2013). Grundlegende Kompetenzen in verschiedenen Bevölkerungsgruppen. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012* (pp. 77-124). Münster: Waxmann.

Massing, N., Ackermann, D., Martin, S., Zabal, A., & Rammstedt, B. (2013). Controlling interviewers' work in PIAAC – the Programme for the International Assessment of Adult Competencies. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in survey impact, reasons, detection and prevention* (pp. 117-130). Frankfurt am Main: Peter Lang.

Mincer, J. (1974). *Schooling, experience and earnings*. New York, NY: National Bureau of Economic Research.

OECD. (2010). *PIAAC technical standards and guidelines* (December 2010). Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm

OECD. (2011). *PIAAC conceptual framework of the background questionnaire main survey*. Paris: OECD.

OECD. (2013a). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. (2013b). *Technical report of the Survey of Adult Skills (PIAAC)*. Paris: OECD.

Rammstedt, B. (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Rouet, J.-F., Bétrancourt, M., Britt, M. A., Bromme, R., Graesser, A. C., Kulikowich, J. M., … van Oostendorp, H. (2009). PIAAC problem solving in technology-rich environments: A conceptual framework. *OECD Education Working Papers* No. 36. Paris: OECD Publishing.

Wasmer, M., Scholz, E., & Blohm, M. (2010). *Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2008*. GESIS Technical Reports. Bonn: GESIS – Leibniz Institute for the Social Sciences.

Zabal, A., Martin, S., Klaukien, A., Rammstedt, B., Baumert, J., & Klieme, E. (2013). Grundlegende Kompetenzen der erwachsenen Bevölkerung in Deutschland im internatio-

nalen Vergleich. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012* (pp. 31-76). Münster: Waxmann.

Zabal, A. Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann Verlag.

methods, data, analyses | Vol. 8(2), 2014, pp. 137-174 DOI: 10.12758/mda.2014.006

How Can Skill Mismatch be Measured? New Approaches with PIAAC

Anja Perry[1], *Simon Wiederhold*[2] & *Daniela Ackermann-Piek*[1,3]

[1] *GESIS – Leibniz Institute for the Social Sciences*
[2] *Ifo Institute – Leibniz Institute for Economic Research at the University of Munich*
[3] *University of Mannheim*

Abstract

Measuring skill mismatch is problematic, because objective data on an individual skill level are often not available. Recently published data from the *Program for the International Assessment of Adult Competencies* (PIAAC) provide a unique opportunity for gauging the importance of skill mismatch in modern labor markets. This paper systematically compares existing measures of skill mismatch in terms of their implications for labor market outcomes. We also provide a new measure that addresses an important limitation of existing measures, namely, assigning a single competency score to individuals. We find that the importance of skill mismatch for individual earnings differs greatly, depending on the measure of mismatch used.

Keywords: skill mismatch, skill use, labor market, PIAAC, Job Requirement Approach

How Can Skill Mismatch be Measured? New Approaches with PIAAC

Skills are the new "global currency of 21st-century economies" (OECD, 2012, p. 10). However, skills must be put to effective use in order to facilitate economic growth and personal labor market success. When skills are not used effectively, we think of them as being mismatched. Skill mismatch occurs when skills possessed by the workers exceed or do not meet the skills required at their workplace. It can lead to skill depreciation and slower adaptation to technological progress, from a macroeconomic perspective (OECD, 2012), and impacts workers' earnings and job satisfaction, from a microeconomic perspective (e.g., Allen & van der Velden, 2001). Recently, the issue of skill mismatch has gained importance in the policy sphere. For instance, the European Union's *Agenda for New Skills and Jobs* (European Commission, 2010) identifies skill mismatch as one of the core challenges faced by today's labor markets. Similarly, the OECD stresses the importance of understanding the causes and consequences of skill mismatch (OECD, 2012).

However, measuring skill mismatch is problematic, because objective data on skills at the individual level are often not available (Leuven & Oosterbeek, 2011, Allen & van der Velden, 2001). The *Programme of the International Assessment of Adult Competencies* (PIAAC), which is an internationally harmonized test of cognitive skills, offers new opportunities to measure skill mismatch. However, there is no widely accepted skill mismatch measure to date. Instead, a number of different approaches to measure skill mismatch have been suggested. Because the variety of existing skills measures imply different shares of mismatched workers in the population and lead to different conclusions regarding the relationship between skill mismatch and labor market outcomes, they also entail different political implications.

This paper is the first one that systematically compares skill mismatch measures, based on the PIAAC data, and assesses their validity by comparing the various measures in a Mincer regression (Mincer, 1974), thus demonstrating the importance of skills for individual earnings. We also introduce a new direct measure of skill mismatch that improves existing measures (discussed in this paper) across

Direct correspondence to
Anja Perry, GESIS – Leibniz Institute for the Social Sciences,
PO Box 12 21 55, 68072 Mannheim, Germany
E-mail: anja.perry@gesis.org

Acknowledgment: The authors wish to thank Matthias v. Davier, Eugenio J. Gonzales, and Jan Paul Heisig for helpful advice on the methodological implementation of measures developed or extended in this paper. The comments of an anonymous referee also helped to improve the paper. Simon Wiederhold gratefully acknowledges financial support from the European Union's FP7 through the LLLight'in'Europe project.

several dimensions. Finally, we perform an analysis for three countries (Austria, Germany, and the United States) to investigate whether both the occurrence and consequences of skill mismatch are affected by differences in labor and product market regulations.

The paper proceeds as follows. In the next section, we highlight the importance of analyzing skill mismatch. We then briefly discuss general approaches to measure skill mismatch in Section 3. In Section 4, we present several skill mismatch measures, using the PIAAC data. In Section 5 we explain the method used to compare and validate those measures; in Section 6, we compare the measures regarding their explanatory power in a Mincerian earnings regression. Finally, we critically discuss the results of our analyses and conclude.

Theoretical Background

Skills form the human capital of an economy. They can be cognitive (such as literacy or numeracy skills) and non-cognitive (such as physical or soft skills). Cognitive skills have been found to correlate positively with individuals' success in the labor market, participation in society, and economic growth (Hanushek, Schwerdt, Wiederhold, & Woessmann, 2014; Hanushek & Woessmann, 2008; OECD, 2013a; Rammstedt, 2013). Indeed, several studies indicate that the above correlations reflect a causal effect of skills (see, for instance, Hanushek & Woessmann, 2012; Oreopoulos & Salvanes, 2011; Riddell & Song, 2011). At the individual level, developing skills enables workers to understand and perform better, and improve economic processes. This productivity-enhancing effect of skills increases a person's wages or allows him or her to escape unemployment and find a job in the first place (e.g., Hanushek & Woessmann, 2014). At the macroeconomic level, better skills lead to faster technological progress and facilitate technology adoption (e.g., Benhabib & Spiegel, 2002; Ciccone & Papaioannou, 2009; Nelson & Phelps, 1966).

Skills, however, must be put to effective use. Only when the workforce uses its skills effectively can individuals generate adequate earnings, which, in turn, foster economic growth (OECD, 2012). We refer to skill mismatch when skills possessed by workers are lower or higher than the level of skills required at the workplace. Thus, workers can either be over-skilled, hence possessing more skills than actually needed on the job (skill surplus), or under-skilled, possessing less skills than needed on the job (skill deficit, e.g., Quintini, 2011b).

Skill mismatch can arise from structural changes in the economy. Innovation and technological change are typically skill-biased, thus increasing the demand for certain types of skills (e.g., Tinbergen, 1974, 1975). Individuals who possess skills that allow fast adaptation to such changes have better chances to stay

employed or to find new employment once they are laid off. Individuals lacking those skills become unemployed or have to accept jobs that do not match their skill portfolios (Acemoglu & Autor, 2011). Several studies suggest that this depends on whether skills are general in nature, that is, whether they are productive in various occupations and therefore transferrable (Hanushek, Schwerdt, Woessmann, & Zhang, 2014), or whether they are occupation-specific (Acemoglu & Autor, 2011; Gathmann & Schönberg, 2010; Nedelkoska, Neffke, & Wiederhold, 2014; Poletaev & Robinson, 2008).

In addition, skill mismatch is related to certain socio-demographic factors. It is likely that a mismatch occurs early in a professional career (Jovanovic, 1979). Inexperienced workers are often found in temporary and entry-level jobs; here, skill requirements are often lower than workers' skills. As workers gain more experience – and are better able to signal their skills by referring to past work experience – it becomes easier for them to move into jobs in which they can adequately apply their skills (Desjardins & Rubenson, 2011; OECD, 2013a). Moreover, women may be more under-skilled than men at the workplace if they are subject to discrimination in the labor market (Desjardins & Rubenson, 2011), or if taking care of children or older family members forces them to work in part-time jobs that typically require fewer skills (OECD, 2013a). Skill mismatch is also a common phenomenon among immigrants whose qualifications can often not be adequately assessed and recognized when they apply for jobs in the host country (Quintini, 2011b).

Previous research calls for a nuanced picture when assessing the consequences of skill mismatch for the economy. On the one hand, a skill surplus can serve as a skill reserve that can be activated once more advanced technologies are introduced at the workplace. On the other hand, skills that are not used may depreciate. Hence, a skill surplus can eventually lead to a loss of skills and thus to a waste of resources that were used to build up existing skills (Krahn & Lowe, 1998; Schooler, 1984) and to lower enterprise productivity as employee turnover increases (Allen & van der Velden, 2001; OECD, 2012). In addition, a skill deficit can challenge existing skills or help to build them up (Schooler, 1984). However, it can also slow down economic growth, because workers possessing too few skills are less able to adapt to technological changes.

Finally, apart from its macroeconomic effects, skills mismatch also influences outcomes at the individual level. First, mismatch affects workers' wages. Typically, over-skilled workers must expect a wage penalty, compared to workers who possess the same skills and match the requirements of their jobs. This is because only skills actually required at a job are rewarded through wages (Tinbergen, 1956). Under-skilled workers are rewarded for applying a large portion of their skills in the job (a proportion presumably larger than someone who is well-matched) and, thus, receive a wage premium. In addition, skill mismatch has an impact on job

satisfaction and the likelihood of workers actively searching for a better match in a new job (Allen & van der Velden, 2001).

However, despite the recent upsurge in interest in skill mismatch, one key challenge remains: How do we adequately measure skill mismatch? The international PIAAC data contain direct measures of adult cognitive skills in various domains, thus providing a unique opportunity to assess skill mismatch in the labor market. In the following section, we present various approaches to measuring skill mismatch, using PIAAC.

Measuring Skill Mismatch

There are essentially two ways to measure skill mismatch: self-reported skill mismatch and direct, objective measures of skill mismatch. Both approaches are predominantly based on methods typically used to measure educational mismatch. Leuven and Oosterbeek (2011) provide a survey of various educational mismatch measures and Quintini (2011a) summarizes skill mismatch measures.

Self-Reported Versus Direct Measures of Skill Mismatch

Most often, self-reports are used to measure skill mismatch. Information on self-reported skill mismatch is obtained by asking workers to what extent their skills correspond to the tasks performed at work (e.g., Allen & van der Velden, 2001; Green & McIntosh, 2007; Mavromaras, McGuinness, & Fok, 2009; Mavromaras, McGuinness, O'Leary, Sloane, & Fok, 2007).[1] Self-report measures have the advantage of being easily implementable in a survey; thus, up-to-date information on skill mismatch can be obtained. However, self-reports are prone to biases. Respondents may have the tendency to overstate the requirements of their workplace and upgrade their position at work (see Hartog, 2000, for education mismatch).

Skill mismatch can also be measured directly, which provides a more objective measure. In all direct skill mismatch measures, workers' skills are compared to skills required at their workplace. For instance, required skills can be measured using the Job Requirement Approach (JRA: Felstead, Gallie, Green, & Zhou, 2007). However, biases can also arise from this approach if respondents overstate their skill use at work. Alternatively, required skills can be measured by obtaining a general, occupation-specific skill level (e.g., Pellizzari & Fichen, 2013), similar to the "Realized Matches" approach applied in education mismatch research (Hartog, 2000; Leuven & Oosterbeek, 2008). Both direct approaches for measuring skill

1 In a similar vein, measures of educational mismatch typically refer to a match between educational qualifications obtained in the past and education required for the job.

mismatch require data on skills actually possessed by the workers. These are typically available in large-scale assessments, such as the International Adult Literacy Survey (IALS), the Adult Literacy and Lifeskills (ALL) Survey, or, most recently, PIAAC. National competency assessments, such as the German National Education Panel Study (NEPS), also provide such information. However, the implementation of large-scale competency assessments is costly. Data on workers' skills are therefore scarce and only available for a limited number of countries and time periods. Nevertheless, direct skill data provide a compelling avenue for measuring skill mismatch.

The PIAAC Data

Overview. Developed by the OECD and implemented between August 2011 and March 2012, PIAAC provides internationally comparable data about skills of the adult population in 24 countries.[2] PIAAC was designed to provide representative measures of cognitive skills possessed by adults aged 16 to 65 years.

Together with information on cognitive skills, PIAAC also offers extensive information on respondents' individual and workplace characteristics, for instance, occupation and skill use at work. This information is derived from a background questionnaire completed by the PIAAC respondents prior to the skills assessment. Using the PIAAC data, we can derive a direct measure of skill mismatch, rather than relying on self-reports, which are prone to biases. Moreover, because PIAAC also contains a measure of self-reported skill mismatch, we can compare direct and self-reported mismatch measures.

Cognitive skills. PIAAC provides measures of cognitive skills in three domains: literacy, numeracy, and problem solving in technology-rich environments. These skills were measured on an infinite scale. By default, respondents had to work on the assessment tasks by using a computer. Respondents without sufficient computer experience were assessed in pencil-and-paper mode.[3] This paper focuses on numeracy mismatch. The average numeracy skill in the three countries at the

2 Countries that participated in PIAAC are Australia, Austria, Belgium (Flanders), Canada, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, the Russian Federation, the Slovak Republic, Spain, Sweden, the United Kingdom (England and Northern Ireland), and the United States.

3 Problem solving in technology-rich environments was measured only in a computer-based mode and was an international option. Cyprus, France, Italy, and Spain did not implement the problem-solving domain.

focus of this paper (Germany, Austria, and the United States) is 267 points, with a standard deviation of 53 points.[4]

The role of plausible values. In PIAAC, skills are a latent variable that is estimated using item-response-theory models (IRT). Because IRT was applied, not all respondents worked on the same set of assessment items and did not receive items covering every skill domain in PIAAC (Kirsch & Yamamoto, 2013). To derive skill information for each respondent and every competency domain, the remaining competency scores for each individual are imputed. To account for possible errors due to imputation, 10 plausible values, instead of only one individual proficiency score, are derived for each respondent and each skill domain. Hence, competency scores in PIAAC represent a competency distribution rather than an individual score (von Davier, Gonzalez, & Mislevy, 2009).

Whereas using the average of the 10 plausible values generally provides an unbiased estimate of a person's skills, the associated standard errors are underestimated, because the uncertainty in skills is not accounted for. Another approach often applied is to use only one plausible value, typically the first one. This also leads to underestimated standard errors, though to a lesser extent. However, the resulting estimates may differ, depending on the plausible value used in the analysis (Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

Existing skill mismatch measures (with the exception of the self-report) neglect the fact that no single proficiency score – neither the first plausible value nor the average of all 10 plausible values – can be assigned to a specific respondent. Allen, Levels, and van der Velden (2013), for instance, use only the first plausible value to compare individual skills with the skills used at the workplace. As we will show in Section 6, replacing the first with another plausible value changes the magnitude of the coefficients on skill mismatch in a Mincer regression. An improved measure of skill mismatch should therefore account for all 10 plausible values, because individual proficiency scores do not adequately represent the individual skill level.[5]

Job Requirement Approach. In addition to the assessment of cognitive skills, PIAAC surveys skills required at the job. To measure job requirements, respondents are asked which skills they use(d) at their current or last workplace and to which extent they use(d) them. This Job Requirement Approach is based on

4 This is very close to the mean (standard deviation) of numeracy skills for all countries that participated in PIAAC: 268 points (53 points). We excluded only the Russian Federation in these calculations because the Russian data are preliminary and may still be subject to change. Additionally, they are not representative of the entire Russian population because they do not include the population of the Moscow municipal area (OECD, 2013b).

5 In Hanushek, Schwerdt, Wiederhold et al. (2014), where the authors measure returns to cognitive skills, using either only the first plausible value or all of them did not affect the results. They thus used only the first plausible value, which greatly reduced the computational burden.

previous work by Felstead et al. (2007). Information on skill use can be compared to the assessed skill level, to decide whether skills possessed by the workers match the skills required at their workplace.

Additional variables. The extensive background questionnaire in PIAAC offers additional information about respondents. It covers education, labor market status, information on the current or most recent job, skills used at the workplace and at home, as well as personal background information. When testing the relationship between skill mismatch and individual earnings (see Section 5), we use years of schooling, gender, and years of work experience as control variables.

Skill Mismatch Measures in PIAAC

As outlined above, PIAAC offers the opportunity to derive direct and objective measures of skill mismatch. However, the PIAAC background questionnaire also includes a skill mismatch self-report, which we additionally examine and include in our analyses. Direct skill mismatch measures discussed here include those derived by Quintini (2012), Allen et al. (2013), the OECD (2013a), and Pellizzari and Fichen (2013), as well as a new measure developed by the authors of this paper.

Whereas direct skill mismatch measures can, technically, be derived for all three proficiency domains in PIAAC, we focus only on numeracy mismatch. We do this because numeracy skills are most likely to be comparable across countries. Moreover, previous research has demonstrated the high relevance of numeracy for wages (e.g., Hanushek, Schwerdt, & Wiederhold et al., 2014; Klaukien et al., 2013). The measures presented here can easily be applied to literacy skills as well. However, greater care must be taken when analyzing skill mismatch related to problem solving in technology-rich environments.[6]

The skill mismatch measures presented in this section are summarized in Table 1.

6 The sample of PIAAC respondents who took part in the problem-solving assessment may be subject to selection effects. In addition, when comparing assessed skills with skill use at work (see Section 3), it is important to remember that the corresponding skill-use index covers only a narrow aspect of this domain (OECD, 2013a).

Table 1 Characteristics of different measures of numeracy mismatch in PIAAC

Measure		Computation	Variables	Consideration of PVs	Pro	Contra
Self-reported measure	Self-report in PIAAC	Categories (well-matched, under-skilled, over-skilled) based on answers to two skill mismatch questions in PIAAC BQ	Skill mismatch self-report (F_Q07a, F_Q07b)	n/a	Can be easily administered in other surveys; refers to general mismatch and not to a specific proficiency domain	Based on self-reported information, which can be biased (e.g. Hartog, 2000); fourth category resulting from combination of both questions "under-skilled as well as over-skilled" is not interpretable; category "well-matched" rather small (e.g., 3.1 % in Germany)
Skill-use-based measures	Quintini (2011), following Krahn and Lowe (1998)	Level of numeracy skill use compared to proficiency level: proficiency level equals numeracy skill use level: well-matched; proficiency level lower than numeracy skill use level: under-skilled, proficiency level higher than numeracy skill use level: over-skilled	Numeracy skill use (G_Q03b G_Q03c G_Q03d G_Q03f G_Q03g G_Q03h); numeracy (PVNUM)	Proficiency level included, not specified whether derived from one PV or from average of all 10 PVs	Can be easily computed	Proficiency and skill use are measured on different scales and should not be compared without standardization; one proficiency level assigned to individuals instead of 10; skill use at work is likely to be overrated by employees (Hartog, 2000); arbitrary cut-off points (one skill level); mismatch restricted to relevant proficiency domain (e.g., numeracy)

Table 1 Characteristics of different measures of numeracy mismatch in PIAAC (cont.)

Measure		Computation	Variables	Consideration of PVs	Pro	Contra
Skill-use-based measures	Allen, Levels, and v. d. Velden (2013)	Three steps 1) PVNUM1 and mean of numeracy skill use standardized to compare different scales 2) Standardized skill use level subtracted from standardized skill level 3) Individuals with resulting value lower than 1.5 points above or below 0: "well-matched", individuals with value less than -1.5: "under-skilled", individuals with value greater than 1.5: "over-skilled"	Numeracy skill use (G_Q03b G_Q03c G_Q03d G_Q03f G_Q03g G_Q03h); numeracy (PVNUM)	PVNUM1	Can be easily computed; numeracy skill use and skill level are standardized to compare the different scales	Only one PV used instead of 10; skill use at work is likely to be overrated by employee (Hartog, 2000); arbitrary cut-off points (1.5 SD); mismatch restricted to relevant proficiency domain (e.g., numeracy)
Realized-matches	OECD (2013a)	Three steps 1) Respondents classified as well-matched based on self-report in PIAAC BQ (see above) 2) Proficiency range for well-matched defined for each country based on self-reported well-matched respondents per occupation	Skill mismatch self-report (F_Q07a, F_Q07b); One-digit ISCO (ISCO1C); numeracy (PVNUM)	Average of ten plausible PVs	Theory-driven approach to define skill mismatch based on workers who are well-matched	Large computational effort; neglects heterogeneity within occupations; base population derived using self-report, which can be biased, resulting in a small N (see above); average of PVs instead of 10 PVs;

Table 1 Characteristics of different measures of numeracy mismatch in PIAAC (cont.)

Measure		Computation	Variables	Consideration of PVs	Pro	Contra
		3) Respondents re-assigned to categories (well-matched, under-skilled, over-skilled) according to defined band-width				Respondents reassigned into mismatch categories according to proficiency range, irrespective to their self-reported information; mismatch restricted to relevant proficiency domain (e.g., numeracy)
Realized-matches	Alternative measure	Four steps 1) Average skill level and SDs computed in each country per occupation 2) Cut-off points for match and mismatch defined for each occupation as 1.5 SD from mean 3) Skill mismatch defined based on cut-off points for each PV for each person (results in 10 skill mismatch variables per person) 4) Average of estimates resulting from 10 skill mismatch variables included in analysis	Two-digit ISCO (ISCO2C); numeracy (PVNUM)	PVNUM1-10	Includes all PVs according to IRT; does not rely on self-reported information, which can be biased	Large computational effort; neglects heterogeneity within occupations; arbitrary cut-off points (1.5 SD); mismatch restricted to relevant proficiency domain (e.g., numeracy)

Notes. BQ = background questionnaire; IRT = item response theory; PV = plausible value; PVNUM = plausible value for numeracy; SD = standard deviation.

Table 2 Self-reported skill mismatch in the PIAAC background questionnaire

		Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job?	
		Yes	No
Do you feel that you need further training in order to cope well with your present duties?	Yes	Over-skilled as well as under-skilled	Under-skilled
	No	Over-skilled	Well-matched

Note. Variables in the PIAAC background questionnaire are: F_Q07a and F_Q07b.

Self-reported Skill Mismatch in PIAAC

The self-report on skill mismatch in PIAAC consists of two questions in the PIAAC background questionnaire (OECD, 2013b):

- Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job?
- Do you feel that you need further training in order to cope well with your present duties?

Each of the questions had to be answered with "yes" or "no" and the combination of both answers provides the self-reported skill mismatch of the respondent (see Table 2).

As shown in Table 2, the combination of both questions leads to four categories, where only the three categories under-skilled, well-matched, and over-skilled are meaningful. It is not entirely clear how we should interpret the remaining category "over-skilled as well as under-skilled". This category may refer to different sets of skills. For example, respondents could consider their mathematical skills when asked whether they have the skills to cope with more demanding tasks at work and confirm. When asked whether they needed further training to cope with their duties, they may have considered their negotiation skills. Furthermore, respondents might feel that they are able to generally cope with more demanding work tasks, but at the same time feel the need for continuously maintaining and developing their skills through training. This is, in particular, the case for highly educated workers who generally have a positive attitude towards education.

Because the answers to these two questions can be interpreted in different ways, we must assume that this measure cannot adequately reflect the construct of skill mismatch. The self-reported measure in PIAAC should therefore *not* be used for measuring skill mismatch.

Skill Mismatch According to Quintini (2012)

Quintini (2012) suggests a PIAAC-based measure of skill mismatch that combines information on skills used at the workplace, using the JRA (Felstead et al., 2007), and competencies assessed in PIAAC. This measure is developed following a previous approach developed by Krahn and Lowe (1998) with data from IALS.

To derive this measure, Quintini grouped skill use and the respective skill proficiency measure into four categories each (level 1 through 4/5). If the levels of skill use and possessed skills are identical, the respondent is well-matched in his or her job. Respondents are under-skilled when their level of skill use is higher than their personal skill level and over-skilled when their skill-use level is lower than their personal skill level.[7]

Krahn and Lowe (1998) assess the validity of their measure and find that using any deviation of skill use from the worker's possessed skills to define mismatch is arbitrary. Whereas Quintini (2012) defines a deviation between skill level and skill use by one level as mismatch, a deviation of two levels defines mismatch for Krahn and Lowe (1998). Hence, agreement on the exact definition of mismatch is lacking. Also, in both studies, skill use is measured by self-reports, which are frequently prone to bias (Hartog, 2000). Allen et al. (2013) point out that skill use and skill level in PIAAC are measured in two different ways and a comparison of these two constructs is not meaningful. In addition, a single plausible value is used to define the numeracy skill level, although how this individual score is derived is not specified. However, a single skill score, irrespective of how it is derived, does not entirely reflect an individual's competency level in PIAAC (Rutkowski et al., 2010; von Davier et al., 2009).

Skill Use in Relation to Skill Level by Allen et al. (2013)

Allen et al. (2013) suggest an alternative, and improved, approach to measure skill mismatch, based on the work of Krahn and Lowe (1998) and Quintini (2012). In a first step, they standardize the average of numeracy skill use and the first plausible value of the numeracy domain, to make both measures comparable.[8] Allen et al. (2013) define mismatch as a deviation of skill use and individual skill level by at least 1.5 standard deviations. Thus, if the difference between standardized numeracy skill use and standardized skill score is below 1.5 standard deviations,

7 Krahn and Lowe (1998) and Desjardins and Rubenson (2011) further disaggregate "well-matched" workers. In Quintini (2012), however, the "well-matched" category corresponds to the other measures presented in this paper.

8 Employed respondents rate their numeracy skill use at their workplace on a six-item scale. A five-point rating scale, ranging from "never" to "every day", was used to measure the respondents' assessments. These are averaged across items to derive a single skill-use score for each employed respondent.

the respondent is defined as being under-skilled. If the difference is larger than 1.5 standard deviations, the respondent is over-skilled. Respondents who are neither over- nor under-skilled are defined as being well-matched.

By standardizing the measures of numeracy skill level and skill use before comparing them, Allen et al. (2013) address an important disadvantage of the measures developed by Krahn and Lowe (1998) and Quintini (2012). However, like the previous authors, Allen et al. (2013) assign an individual skill score to the respondent, even though such an individual skill score does not entirely reflect the respondent's actual competency. Furthermore, self-reported skill use can be overestimated by the respondent (Hartog, 2000). In addition, one can argue that using a bandwidth of 1.5 standard deviations to define mismatch is arbitrary and other boundaries should be considered. The authors argue that this definition of mismatch is "fairly extreme" (p. 10). This is to ensure that workers identified as being mismatched possess skill levels that are indeed unusually high or low, compared to workers facing similar job requirements.

Skill Mismatch by the OECD (2013a) and Pellizzari and Fichen (2013)

In its *Skills Outlook*, the OECD (2013a) presents a new direct measure of skill mismatch that is discussed in detail by Pellizzari and Fichen (2013). This measure follows the "Realized Matches" approach (cf. Hartog, 2000; Leuven & Oosterbeek, 2011).

In a first step, the authors look at respondents who are well-matched, according to the self-report in PIAAC (see above). For this group of workers, they derive a competency bandwidth by country and occupation.[9] To account for outliers, respondents in the top and bottom 5 % of the skill distribution in each occupation are excluded when deriving the bandwidth. Moreover, to obtain a sufficient number of respondents in the well-matched category, only occupations at the one-digit ISCO level were used.[10] Individuals whose skill levels are below/above this bandwidth are considered to be under-skilled/over-skilled. Individuals whose skills are within the bandwidth are labeled well-matched. Importantly, all respondents are assigned

9 In PIAAC, the respondents reported their occupation verbally by naming the profession and describing their work tasks in detail. This information was then recoded into the International Standard Classification of Occupations (ISCO-08, International Labour Organization, 2012).

10 ISCO 0 (armed forces) and ISCO 6 (skilled agricultural, forestry, and fishery workers) were eliminated from the analysis and the categories ISCO 1 (managers) and ISCO 2 (professionals) were combined, due to the small number of observations in these categories.

a level of skill mismatch that is based on the average of their 10 plausible values in numeracy.

The results of this skill-mismatch measure should be interpreted with great caution. As stated above, the self-report used in the PIAAC background questionnaire cannot adequately reflect whether or not a respondent's skills match the skills required at his or her workplace. Moreover, only a small proportion of respondents report being well-matched (see Table 3). Thus, even though the definition of bandwidths is based on the one-digit ISCO level and is therefore very broad, the number of observations within one occupation is often still small. For some occupations in some countries, the bandwidth is based on only very few observations.[11] However, Allen et al. (2013) argue that the derived occupation-level 5th to 95th percentile ranges do not differ systematically from those based on the full sample. Thus, the restriction of using only well-matched workers to derive occupation-specific bandwidths could also be neglected. Allen et al. (2013) further criticize the OECD approach to measuring skill mismatch for neglecting heterogeneity within occupations, because the OECD defines one bandwidth for all respondents within an occupation. In addition, the average of all 10 plausible values is used to assign individual proficiency scores. However, as explained above, the average of plausible values does not reflect individual competency and, when used in analyses, underestimates associated standard errors to an even greater extent than if only one of the ten plausible values is used (Rutkowski et al., 2010).

An Alternative Measure to Compute Skill Mismatch

We propose an alternative measure for calculating skill mismatch that also follows the "Realized Matches" approach, improving on the measure by the OECD (2013a) and Pellizzari and Fichen (2013). We also define bandwidths for each occupation according to the average skill level and, thus, avoid using self-reported information about skill use that may be biased. Also, as Allen et al. (2013) argue, skill levels of workers who report being well-matched in PIAAC do not differ substantially from those of workers in general. Thus, we define boundaries between matched and mismatched workers for each occupation, based on the total population of workers in a country. The resulting increase in the number of observations allows us to use the more detailed two-digit ISCO categorization to derive bandwidths within occupations. To reduce measurement error, we eliminated a few occupations to reach a minimum number of observations by country-occupation cell of 30. Like Allen et al. (2013), we calculate the mean proficiency score for each occupation in each

11 The authors base further steps on at least 10 observations per occupation. However, whenever the sample is reduced (as done in this paper, by looking at full-time employees only), the number of observations decreases on the occupation level.

country and add/subtract 1.5 standard deviations to define the corridor of being well-matched. Contrary to other measures discussed here, we take into account all 10 plausible values for each individual by repeating the above procedure for all plausible values. However, as a result of this procedure, respondents can be categorized simultaneously as well-matched *and* mismatched. Therefore, to calculate estimates, for example, percentages of workers who are mismatched as well as regression coefficients, we take the average of the results computed with each plausible value to derive our final estimate. By applying this procedure, we derive more reliable estimates of skill mismatch than previous studies that use the PIAAC data.

When choosing between different measures of skill mismatch, researchers need to know which measure is most suitable and, especially, most valid for their types of analyses. Following Groves, Fowler, Couper, Singer, and Tourangeau (2004), a measure is valid when the operationalization (in our case the skill mismatch measure) corresponds to the construct of interest (in our case existing skill mismatch). To derive recommendations regarding which measure to use when analyzing skill mismatch, we compare them in a Mincer regression on earnings (Mincer, 1974). The next section describes the Mincer regression in more detail.

Empirical Approach

The aim of this paper is to compare various skill-mismatch measures in PIAAC. After having described the measures in the preceding section, we now attempt to judge their validity by looking at differences in outcomes, namely, the proportion of matched and mismatched workers and the relationship between skill mismatch and earnings in a Mincer regression model (Mincer, 1974).

Empirical Model

When examining the relationship between various measures of skill mismatch and earnings, we rely on a Mincer-type regression model. The Mincer regression is probably the most widely used empirical model in economic research.[12]

The regression equation reads as follows:

$$\text{ln } y_i = \beta_0 + \beta_1 C_i + \beta_2 U_i + \beta_3 O_i + \beta_4 S_i + \beta_4 G_i + \beta_6 E_i + \beta_7 E_i^2 + \varepsilon_i \qquad (1)$$

where y_i is the (pre-tax and pre-transfer) hourly wage of individual *i*. To correct for outliers, we trimmed wages in Germany by removing the highest and lowest 1 % of observed earnings. Due to data restrictions, we do not have access to con-

12 See Heckman, Lochnern, & Todd (2006) for a recent review of the literature.

tinuous wage information for Austria and the U.S. Instead, we used information on the median wage of each decile, which allowed us to assign the decile median to each survey participant belonging to the respective decile of the country-specific wage distribution (Hanushek, Schwerdt, & Wiederhold et al., 2014, apply a similar procedure). *C* is the individual's numeracy skills, *U* is a dummy variable for being under-skilled, *O* a dummy variable for being over-skilled, represented by 10 plausible values[13], *S* is the number of years of schooling (average or most usual time that it takes to complete a qualification), *G* is a dummy variable taking the value 1 for female and 0 for male. We also include a quadratic polynomial in work experience, *E*, to account for positive but diminishing returns of experience on earnings.[14] ε is the stochastic error term.

Sample

For each country participating in PIAAC, a sample of at least 5,000 adults[15] was surveyed. We use sampling weights to obtain nationally representative estimates. Moreover, to account for the complex sample design, we use replicate weights in all estimations.[16]

Our analysis only includes persons who were employed full-time at the time of the survey. Like Hanushek, Schwerdt, Wiederhold et al. (2014), we define full-time employees as those who work 30 hours or more per week. We exclude students and apprentices. Students who work while studying are unlikely to have a job that makes proper use of their skills. Apprentices are typically paid lower wages than equivalent workers who have completed their vocational education. In addition, the self-employed are excluded from the sample, because this group typically includes extreme outliers regarding hourly earnings.

Country Selection

Of the 24 countries surveyed in PIAAC, we focus on Austria, Germany, and the U.S. Our main analysis uses the German PIAAC data. However, to check whether our results can be generalized to other country economies, we compare the results

13 Numeracy is the ability to access, use, interpret and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life (Gal et al., 2009).

14 Numeracy skills and work experience squared are divided by 100, to facilitate exposition.

15 In countries that did not implement the skill domain problem solving in technology-rich environments, at least 4,500 adults were assessed (Mohadjer, Krenzke, & Van de Kerchove, 2013a).

16 Detailed information on the sampling processes in PIAAC is presented in Mohadjer, Krenzke, & Van de Kerchove (2013b).

for Germany with those from Austria and the U.S. We chose Austria because its education system is similar to that in Germany, particularly with respect to its emphasis on vocationally oriented education.[17] In the U.S. education system, on the contrary, skills are less specific to a particular occupation but more general in their applicability. This general education arguably provides students with broad knowledge and basic skills in mathematics and communication, which can serve as a foundation for further learning on the job.[18] Moreover, social and labor market institutions differ vastly between Austria/Germany and the U.S.

Results

In this section, we present the results of our analyses. First, we focus on existing measures of skill mismatch, comparing the percentages of well-matched and mismatched workers in Germany, Austria, and the U.S. and the relationship between mismatch and earnings. We then show that the measure developed by Allen et al. (2013) produces quite different results, depending on the plausible value used in the analyses. Finally, we present results for our newly developed skill mismatch measure and compare them with an adjusted version of the Allen et al. (2013) measure that accounts for all 10 plausible values.

Existing Measures: Percentages of Mismatched Workers

The percentages of mismatched workers differ widely between the skill mismatch measures (see Table 3). For example, the percentage of well-matched workers in Germany ranges from below 4 % in the PIAAC self-report to 84 % in the measure reported by the OECD (2013a) and Pellizzari and Fichen (2013). The percentage of under-skilled workers ranges between 4 %, using the self-report measure, and 30 %, using the measure suggested by Quintini (2012). Finally, for over-skilled workers, the percentages for Germany vary between 8 %, according to Allen et al. (2013), and 46 %, according to the self-reports. We observe similar differences in the percentage of mismatched workers in Austria and the U.S. These findings suggest that different skill mismatch measures will also result in quite different distributions of skill mismatch across subgroups; indeed, we observe such differences for gender, age, and education.[19]

17 See Woessmann (2014) for an extensive discussion of the link between education and individual earnings.

18 Using the IALS data, Hanushek, Schwerdt, Woessmann et al. (2014) show that, at entry-age, employment rates are higher for people who gained vocational education. However, this turns around later, when people with a general education degree have substantially higher employment rates.

19 Results available from the authors upon request.

Table 3 Share of mismatched workers by definition of skill mismatch

Country	Mismatch category	Mismatch measures (Numeracy)			
		Self-report	Quintini (2012)	Allen et al. (2013)	OECD (2013a)
Germany	Under-skilled	3.93	30.42	8.36	2.88
		(0.46)	*(0.84)*	*(0.60)*	*(0.35)*
	Well-matched	3.48	33.96	83.70	84.09
		(0.38)	*(0.87)*	*(0.78)*	*(0.71)*
	Over-skilled	45.81	35.61	7.94	13.02
		(1.11)	(1.02)	(0.58)	(0.69)
Austria	Under-skilled	2.96	23.83	8.65	1.80
		(0.36)	*(0.95)*	*(0.55)*	*(0.29)*
	Well-matched	4.03	34.55	83.03	86.62
		(0.42)	*(0.90)*	*(0.68)*	*(0.74)*
	Over-skilled	53.39	41.61	8.32	11.57
		(0.97)	*(0.98)*	*(0.50)*	*(0.68)*
USA	Under-skilled	2.33	44.71	9.65	4.54
		(0.30)	*(1.09)*	*(0.55)*	*(0.42)*
	Well-matched	5.35	31.63	81.24	86.51
		(0.47)	*(0.98)*	*(0.85)*	*(0.67)*
	Over-skilled	71.84	23.66	9.11	8.95
		(1.09)	*(0.91)*	*(0.72)*	*(0.62)*

Notes. Full-time employees between 16 and 65 years of age, excluding students and apprentices. Standard error in parentheses. Percentages in self-reported measure do not add up to 100 % due the fourth category "under-skilled and over-skilled" that is not reported here. The OECD measure excludes members of the armed forces (ISCO 0) and skilled agricultural, forestry, and fishery workers (ISCO 6). *Data source:* OECD (2013c) and Rammstedt et al. (2014).

Measures: Relationship Between Numeracy Mismatch and Earnings

We now investigate the relationship between skill mismatch and individual earnings. In Figure 1, the length of each bar represents the coefficient magnitude resulting from an estimation of the Mincer regression in Equation (1) for each measure of skill mismatch[20] in numeracy and country.[21] The exact coefficient and level of significance are displayed next to each bar. Similar to previous findings on education mismatch (Hartog, 2000) and skill mismatch (Allen et al., 2013), workers with a surplus/deficit of skills receive wage penalties/premiums, compared to workers with the same skills who are well-matched. However, the result that over-skilled workers suffer a wage penalty shows up more systematically in our data than the wage premium for under-skilled workers. Moreover, the magnitudes of these relationships vary substantially according to the measure of skill mismatch. Considering the wage premium for being under-skilled, the OECD (2013a) measure provides the largest range: from insignificant in Germany and the U.S. to 16 % in Austria. On the other hand, the wage premiums for the Quintini (2012) measure are the smallest and, in fact, never significant.

The coefficients on over-skilling also differ widely across the measures. We further observe pronounced country differences regarding the mismatch estimates. In Germany and the U.S., we obtain very high wage penalties when using the OECD (2013a) measure, whilst, in Austria, penalties are smallest with this measure. The U.S. stands out as having by far the largest wage penalty for over-skilled workers; the coefficient implies a decrease in earnings of 23 % when a worker is over-skilled, using the OECD mismatch measure. In terms of magnitude, the self-reported mismatch measure always yields the smallest earnings penalty for over-skilling. This result is probably due to the fact that, across all measures, the self-report yields by far the largest percentage of over-skilled workers (see Table 3).

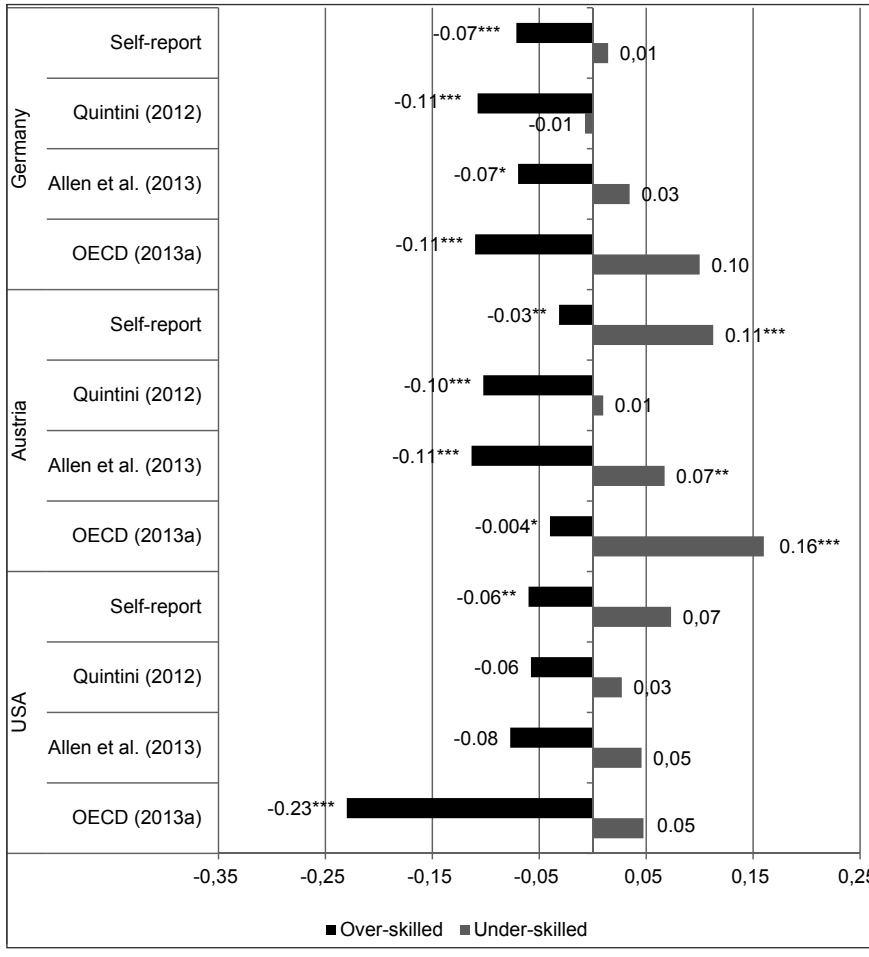
Note that sample sizes differ across the regression models. This is due to omitted cases in professions with a low number of well-matched workers (OECD measure) and to missing values in the background questionnaire (self-reported measure). However, the R^2 do not differ notably across the regression models, when we use a common sample for all measures.[22]

As described above (see Section "*The Role of Plausible Values*"), calculations involving proficiency scores should, ideally, take all 10 plausible values into account. Thus far, however, we performed the Mincer regressions with the original measures that use the average of all plausible values (OECD, 2013a; Pellizzari &

20 We consider the results pertaining to our own mismatch measure in a separate section below.

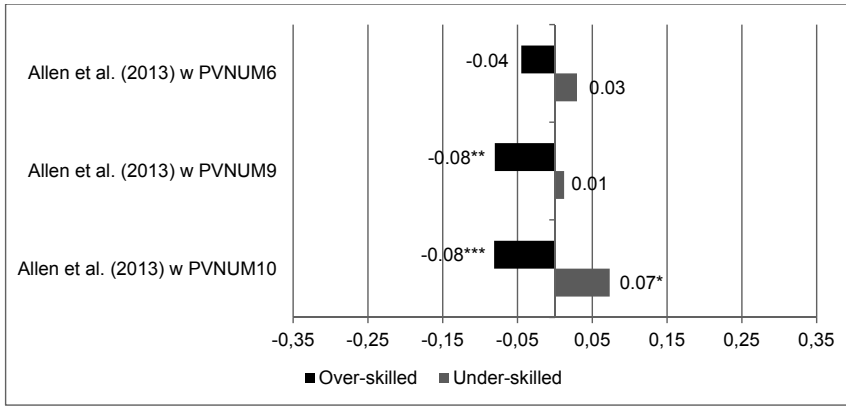
21 See Tables A.1-A.4 for detailed results.

22 Results of this comparison are available upon request from the authors.



Notes. Bars resulting from least squares regressions weighted by sampling weights. Dependent variable: log gross hourly wage. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. The OECD measure excludes members of the armed forces (ISCO 0) and skilled agricultural, forestry, and fishery workers (ISCO 6). See Section "Empirical Approach" for details of the Mincer regression and Tables A.1 to A.4 for regression results. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Figure 1 Coefficients of various skill-mismatch measures in a mincer regression



Notes. Bars resulting from least squares regressions weighted by sampling weights. Dependent variable: log gross hourly wage. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression and Table A.5 for regression results. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Figure 2 Mincer-regression coefficients of skill-mismatch measure of Allen et al. (2013) with three different plausible values for Germany

Fichen, 2013) or only the first plausible value (Allen et al., 2013; Quintini, 2012) to assign individual proficiency scores. To assess the importance of uncertainty in skill scores when analyzing skill mismatch, we calculated the measure suggested by Allen et al. (2013) with the remaining nine plausible values in the same Mincer regression model, as described above. In Figure 2, we present the regression results for plausible values 6, 9, and 10 for Germany.[23] We observe that the results for each alternative plausible value differ to a considerable extent. The increase in earnings if a worker is under-skilled ranges from being insignificant (PVNUM6 and 9) to 7 % (PVNUM10). The earnings decrease for over-skilled workers ranges from being insignificant (PVNUM6) to 8 % (PVNUM9 and PVNUM10).

Refined Measures of Skill Mismatch

Next, we present results from our newly developed skill mismatch measure that takes all 10 plausible values into account. Moreover, as described above, this measure only uses objective skill scores and does not rely on any self-reported information. In Table 4, we present the percentages of well-matched, over-skilled, and

23 See Tables A.5 for detailed results.

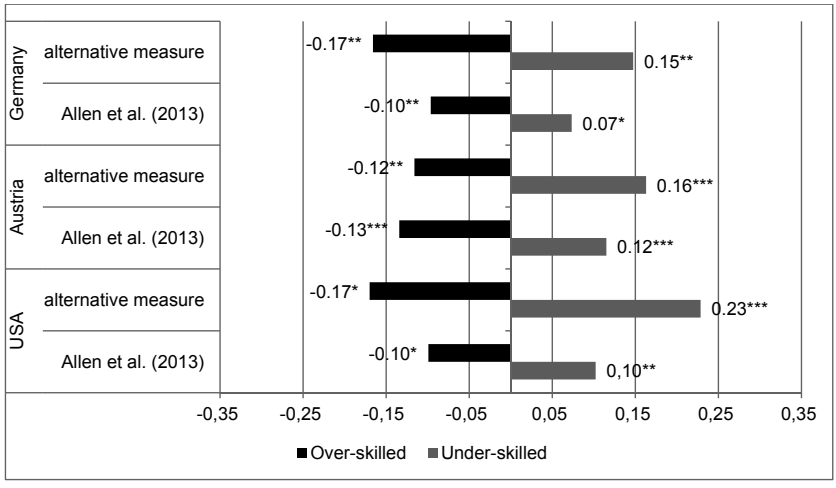
Table 4 Share of mismatched workers by definition of skill mismatch taking all plausible values into account

Country	Mismatch category	Mismatch measures (Numeracy)	
		Allen et al. (2013)	alternative measure
Germany	Under-skilled	8.46	7.39
		(0.66)	*(0.76)*
	Well-matched	83.55	87.23
		(0.93)	*(1.00)*
	Over-skilled	7.99	5.37
		(0.69)	*(0.70)*
Austria	Under-skilled	8.86	6.91
		(0.68)	*(0.62)*
	Well-matched	83.15	87.50
		(0.89)	*(0.86)*
	Over-skilled	7.99	5.59
		(0.59)	*(0.61)*
USA	Under-skilled	9.79	7.65
		(0.66)	*(0.65)*
	Well-matched	80.76	86.70
		(0.94)	*(0.87)*
	Over-skilled	9.45	5.65
		(0.71)	*(0.53)*

Notes. Full-time employees between 16 and 65 years of age, excluding students and apprentices. Standard error in parentheses. The alternative measure excludes workers in professions with less than 30 observations per country (at two-digit ISCO level). *Data source:* OECD (2013c) and Rammstedt et al. (2014).

under-skilled workers according to this measure. For comparison, we also present percentages of workers using the Allen et al. (2013) measure with all plausible values. We focus further analyses on these two measures, because we see both as improvements, compared to previously described skill mismatch measures (i.e., those of OECD, 2013a; Pellizzari & Fichen, 2013; Quintini, 2012).

The percentage of mismatched workers differs only slightly between the two measures, with somewhat large differences regarding the share of over-skilled workers. Especially in the U.S., the percentage of over-skilled workers derived with the adjusted measure of Allen et al. (2013) (9 %) is almost 70 % larger than that derived by the alternative measure (6 %). Generally, the percentage of well-matched workers is lower for the adjusted Allen et al. (2013) measure vis-a-vis our own



Notes. Bars resulting from least squares regressions weighted by sampling weights. Dependent variable: log gross hourly wage. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. The alternative measure excludes workers in professions with less than 30 observations per country (at two-digit ISCO level). See Section "Empirical Approach" for details of the Mincer regression and Tables A.6 and A.7 for regression results. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Figure 3 Mincer-regression coefficients of various skill mismatch measures taking all plausible values into account

measure. Compared to their original measure, the adjusted measure of Allen et al. (2013) leads to slight changes in the percentage of mismatched workers. In particular, the standard errors increase, because uncertainty increases when all plausible values are taken into account.

When using both measures in a Mincer regression, coefficients for being over-skilled and under-skilled again differ (see Figure 3).[24] Considering the wage premium for being under-skilled, our measure consistently produces larger estimates than the refined measure of Allen et al. (2013), ranging from 15 % in Germany (Allen et al.: 7 %) to 23 % in the U.S. (Allen et al.: 10 %). For Germany and the U.S., our measure also shows larger wage penalties for over-skilled workers, namely 17 % (Allen et al.: 10 %), whilst the wage penalty is similar to that yielded by the refined Allen et al. (2013) measure for Austria (12 % vs. 13 %). Importantly, in contrast to the results shown in Figure 1, all coefficients using any of these two skill-mismatch measures are significant at 10 % or better.

24 See Tables A.6 and A.7 for detailed results.

Interestingly, wage premiums for under-skilled workers are smaller or equal to wage penalties of over-skilled workers when the refined measure of Allen et al. (2013) is used. Applying our alternative skill-mismatch measure produces a larger wage premium for under-skilled workers in Austria and the U.S., compared to the wage penalty incurred by over-skilled workers. In Germany, the alternative measure indicates that the wage premium for under-skilled workers is slightly lower than the wage penalty for over-skilled workers.

Again, we report different sample sizes for each measure, because we had to omit cases in professions with less than 30 workers when computing the alternative skill mismatch measure. This results in the reduction of sample sizes by up to 184 cases in Germany. Although the coefficient estimates differ between the two measures, the R^2 are again similar for both measures, when they are compared within the same sample.[25] This implies similar predictive validities of both measures, even though the magnitude of the coefficients differs.

We performed several further checks to test the robustness of these results. For instance, we performed the regression separately for men and women. While the coefficients for skill mismatch become slightly larger in the regression models that contain only male workers, they become insignificant for women, which is due to a smaller sample size. Moreover, we restricted the sample to prime-age workers who, as Hanushek, Schwerdt, Wiederhold et al. (2014), for instance, argue, should be less often mismatched than entry-age workers. Doing so, we, again, find only slight changes compared to our original regression model.[26]

Discussion

Differences in Results Across Skill Mismatch Measures

Although the underlying data were the same in all analyses, the percentages of mismatched workers resulting from different measures vary substantially. While the self-reported measure suggests a very small percentage of well-matched workers, the measures proposed by Allen et al. (2013) and the OECD (2013a) yield a percentage of well-matched workers well above 80%. The higher percentages resulting from the latter two measures seem to be much closer to reality than the self-reported measure, because it is hard to imagine that the majority of workers are mismatched in their jobs. The substantial differences in these results already imply that researchers must carefully consider their choice of skill mismatch measure.

We also compared the relationship between the various skill mismatch measures and earnings in a Mincer regression. Although the results indeed confirm

25 Results of this comparison are available on request from the authors.

26 Results of this comparison are available on request from the authors.

the commonly found relationship between mismatch and earnings (cf. Allen et al., 2013; Hartog, 2000) – namely, under-skilled workers earn a wage premium and over-skilled workers incur a wage penalty – the coefficient magnitudes differ widely between the skill mismatch measures.

One problem with existing skill mismatch measures is that, in assigning a single skill score to each respondent, they neglect important assumptions of IRT. No individual skill score, neither the first of 10 plausible values nor the average of all 10 plausible values, captures the uncertainty in a respondent's skill level in PIAAC. This becomes apparent when, as a simple example, we compare the measure developed by Allen et al. (2013) with three different plausible values.

To overcome this problem, we calculated skill mismatch variables per respondent for all 10 plausible values and took the average of the resulting statistics. While this procedure can, in principle, be applied to all direct measures presented in this paper, we derived results based on this approach only for the measure suggested by Allen et al. (2013), as an improved version of the measure by Quintini (2012), and for the alternative measure we propose in this paper, as an improved version of the OECD measure (OECD, 2013a; Pellizzari & Fichen, 2013).

Comparing our results to the original measure of Allen et al. (2013) reveals differences in Mincer regression coefficients and standard errors. This suggests that whether only one plausible value or whether the mean of all plausible values is used has consequences when the implications of skill mismatch are investigated.

Although results differ for the various skill mismatch measures, the general pattern appears similar: earnings increase when workers are under-skilled and decrease when workers are over-skilled. Previous research finds that wage premiums for being under-skilled are usually smaller than wage penalties for being over-skilled (e.g., Allen et al., 2013; Hartog, 2000). Depending on the extent of skills not used when workers are over-skilled, the drop in earnings can be relatively large. When workers are under-skilled, on the other hand, the skill level they possess limits their productivity and prevents large wage premiums. We are able to replicate these findings using the redefined measure of Allen et al. (2013); however, when using our alternative measure, wage premiums for under-skilled workers are larger than wage penalties for over-skilled workers in Austria and the U.S., but not in Germany. These results resemble previous evidence obtained for education mismatch: there are country-specific differences in the pattern of penalties and rewards related to skill mismatch (cf. Hartog, 2000). Interestingly, we find a large difference between the two measures for under-skilled workers in the U.S. and Germany, but only small differences in Austria. Further research is required to investigate the causes of these differences in parameter estimates. Nevertheless, the predictive validity of both measures (as inferred by the R^2 of the Mincer models) is the same.

The sample size, when applying our measure (as well as the OECD measure), is reduced, compared to the other measures. This is due to omitting cases from the

sample in professions with fewer respondents than the defined threshold. This procedure not only complicates the computation of both measures and is prone to error but it also reduces the representativeness of both measures, because they do not represent the entire population of the analyzed countries. This is especially true for the alternative measure that omits 184 cases for Germany, compared to measures based on comparing skill scores and skill use.

Limitations of the Presented Direct Skill Mismatch Measures

A major disadvantage of all direct skill mismatch measures discussed in this paper is that they focus on only one skill domain, in our case numeracy. Although it is possible to derive additional measures for literacy or problem-solving mismatch, these measures will only shed limited light on actually existing mismatches, because they only cover the cognitive dimension of skills. Ideally, we would like to extend the scope of skill mismatch to other, non-cognitive skills, e.g., extraordinary sales or management talents; however, these are not assessed in PIAAC. We are neither able to measure occupation-specific skills nor any resulting mismatch.[27] In general, looking at only one skill domain – although informative – does not provide a complete picture of skill mismatch.

Conclusions

This paper contributes to existing research on skill mismatch in several ways. First, we review existing measures of skill mismatch and assess their differences in various empirical applications. Second, we discuss the validity of each measure, with a main focus on methodological aspects, such as the wording of the questions in the PIAAC questionnaire of the self-report on skill mismatch and the use of plausible values when considering cognitive skills in the analysis. Third, we develop a new measure of skill mismatch that avoids some weaknesses of existing measures. One major improvement is that all plausible values are taken into account, accurately reflecting the uncertainty in individual skills, as assessed in PIAAC. Moreover, this measure only relies on actually tested skills, neglecting subjective responses on skill use at the workplace, which are prone to misreporting.

Our results indicate that the percentage of mismatched workers in the population, as well as wage implications of being mismatched, differ widely between the measures. Possible sources of these differences may be biases in response behav-

27 See Nedelkoska, Neffke, & Wiederhold (2014) for a discussion of the implication of occupation-specific skill mismatch.

ior, especially when self-reports are used in the calculations, and methodological errors, such as relying on very small samples (i.e., number of respondents by occupations) upon which further computations are based.

Whenever large-scale assessment data are used, one has to carefully consider methodological particularities, such as complex sample design and uncertainty in skill scores expressed through multiple plausible values per individual. Thus, researchers measuring skill mismatch must pay great attention to their choice of measure and its computation. We strongly advise against using the self-report surveyed in the PIAAC background questionnaire because it cannot adequately reflect the respondent's actual perception of match or mismatch. Rather, we recommend the use of direct skill mismatch measures, such as the revised measure of Allen et al. (2013) or our own measure. If an invalid measure of skill mismatch is applied, the resulting policy implications will surely be misleading.

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labour Economics* (Vol. 4b, pp. 1043-1171). Amsterdam: Elsevier B.V.

Allen, J., Levels, M., & van der Velden, R. (2013). *Skill mismatch and skill use in developed countries: Evidence from the PIAAC study.* ROA Research Memorandum. Research Centre for Education and the Labour Market (ROA). Maastricht.

Allen, J., & van der Velden, R. (2001). Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search. *Oxford Economic Papers, 53*(3), 434-452.

Benhabib, J., & Spiegel, M. M. (2002). Human Capital and Technology Diffusion. San Francisco: FRB of San Francisco.

Ciccone, A., & Papaioannou, E. (2009). Human capital, the structure of production, and growth. *Review of Economics and Statistics, 91*(1), 66-82.

Desjardins, R., & Rubenson, K. (2011). An analysis of skill mismatch using direct measures of skills.

European Commission. (2010). New Skills for New Jobs: Action Now: Report by the expert group on New Skills for New Jobs prepared for the European Commission.

Felstead, A., Gallie, D., Green, F., & Zhou, Y. (2007). Skills at work, 1986 to 2006. Cardiff: ESRC Research Centre on Skills, Knowledge and Organizational Performance.

Gal, I., Alatorre, S., Close, S., Evans, J., Johansen, L., Maguire, T. … Tout, D. (2009). *PIAAC Numeracy: A Conceptual Framework*. OECD Education Working Paper No. 35. Paris: OECD Publishing.

Gathmann, C., & Schönberg, U. (2010). How general is human capital? A task-based approach. *Journal of Labor Economics, 28*(1).

Green, F., & McIntosh, S. (2007). Is there a genuine under-utilization of skills amongst the over-qualified? *Applied Economics, 39*, 427-439.

Groves, R., Fowler, F., Couper, M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology.* New York: Wiley.

Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2014). Returns to skills around the world: Evidence from PIAAC. *European Economic Review*, forthcoming.

Hanushek, E. A., Schwerdt, G., Woessmann, L., & Zhang, L. (2014). General education, vocational education, and labor-market outcomes over the life-cycle. Revised version of NBER Working Paper 17504. Stanford University.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature, 46*(3), 607-668.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267-321.

Hanushek, E. A., & Woessmann, L. (2014). *The knowledge capital of nations*. Book manuscript.

Hartog, J. (2000). Over-education and earnings: where are we, where should we go? *Economics of Education Review, 19*(2), 131-147.

Heckman, J. J., Lochnern, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (pp. 307-458). Amsterdam: North-Holland.

IEA Data Processing and Research Center (DPC). (2014). IEA International Database (IDB) Analyzer (version 3.1). Retrieved from http://www.iea.nl/data.html

International Labour Organization. (2012). International standard classification of occupations ISCO-08. Genf: International Labour Organization.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy, 87*(5), 972-990.

Kirsch, I., & Yamamoto, K. (2013). Assessment Design. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing.

Klaukien, A., Ackermann, D., Helmschrott, S., Rammstedt, B., Solga, H., & Woessmann, L. (2013). Grundlegende Kompetenzen auf dem Arbeitsmarkt. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Krahn, H., & Lowe, G. S. (1998). Literacy utilization in Canadian workplaces. *International Adult Literacy Survey monograph series*. Ottawa: Statistics Canada, Human Resources Development Canada (HRDC).

Leuven, E., & Oosterbeek, H. (2008). An alternative approach to estimate the wage returns to private-sector training. *Journal of Applied Econometrics, 23*, 423-434.

Leuven, E., & Oosterbeek, H. (2011). Overeducation and mismatch in the labor market. In E. A. Hanushek, S. Machin & L. Wößmann (Eds.), *Handbook of the economics of education* (Vol. 4, pp. 283-326). Amsterdam: Elsevier B.V.

Mavromaras, K., McGuinness, S., & Fok, Y. (2009). Assessing the incidence and wage effects of over-skilling in the Australian labour market. *Economic Record, 85*, 60-72.

Mavromaras, K., McGuinness, S., O'Leary, N., Sloane, P., & Fok, Y. (2007) The problem of overskilling in Australia and Britain. *IZA Discussion Paper No. 3136*.

Mincer, J. (1974). *Schooling, experience and earnings*. New York, NY: National Bureau of Economic Research.

Mohadjer, L., Krenzke, T., & Van de Kerchove, W. (2013a). Sampling design *Technical report of the Survey of Adult Skills (PIAAC)*. Paris: OECD.

Mohadjer, L., Krenzke, T., & Van de Kerchove, W. (2013b). Indicators of the Quality of the Sample Data *Technical report of the Survey of Adult Skills (PIAAC)*. Paris: OECD.

Nedelkoska, L., Neffke, F., & Wiederhold, S. (2014). *Skill Mismatch and the Costs of Job Displacement*. Unpublished manuscript.

Nelson, R. R., & Phelps, E. S. (1966). Investment in humans, technological diffusion, and economic growth. *The American Economic Review, 56*(1/2), 69-75.

OECD. (2012). Better skills, better jobs, better lives: A strategic approach to skills policies. Paris: OECD.

OECD. (2013a). *OECD skills outlook: First results from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. (2013b). PIAAC Background Questionnaire. Retrieved from http://www.oecd.org/site/piaac/BQ_MASTER.HTM

OECD. (2013c). International Public Use Data Files. Retrieved from http://vs-web-fs-1.oecd.org/piaac/puf-data

Oreopoulos, P., & Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives, 25*(1), 159-184.

Pellizzari, M., & Fichen, A. (2013). *A new measure of skills mismatch: Theory and evidence from the Survey of Adult Skills (PIAAC)*. OECD Social, Employment and Migration Working Papers. OECD Publishing.

Poletaev, M., & Robinson, C. (2008). Human capital specificity: Evidence from the dictionary of occupational titles and displaced worker surveys,1984-2000. *Journal of Labour Economics, 26*(3), 387-420.

Quintini, G. (2011a). Over-qualified or under-skilled: A review of existing literature *OECD Social, Employment and Migration Working Papers*. Paris: OECD.

Quintini, G. (2011b). Right for the job: Over-qualified or under-skilled? Paris: OECD.

Quintini, G. (2012). *The skill proficiency of the labour force and the use of skills in the workplace*. Paper presented at the 10th meeting of the PIAAC BPC. 3-4 May, 2012. Berlin, Germany.

Rammstedt, B. (Ed.). (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Rammstedt, B., Zabal, A., Martin, S., Perry, A., Helmschrott, S., Massing, N., Ackermann, D., . . . Maehler, D. (2014): *Programme for the International Assessment of Adult Competencies (PIAAC), Germany - Reduzierte Version*. GESIS Datenarchiv, Köln. ZA5845 Datenfile Version 1.0.0, doi:10.4232/1.11865

Riddell, W. C., & Song, X. (2011). The impact of education on unemployment incidence and re-employment success: Evidence from the U.S. labour market. *Labour Economics, 18*(4), 453-463.

Rutkowski, L., Gonzalez, E. J., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151.

Schooler, C. (1984). Psychological effects of complex environments during the life span: A review and theory. *Intelligence, 8*, 259-281.

Tinbergen, J. (1956). On the theory of income distribution. *Weltwirtschaftliches Archiv, 77*, 156-175.

Tinbergen, J. (1974). Substitution of graduate by other labor. *Kyklos, 27*, 217-226.

Tinbergen, J. (1975). *Income difference: Recent research*. Amsterdam: North-Holland Publishing Company.

von Davier, M., Gonzalez, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI monograph series: Issues and methodologies in large-scale assessments, 2*, 9-36.

Woessmann, L. (2014). *The economic case for education*. Unpublished manuscript.

Appendix

Table A.1

Mincer regressions with Self-reported skill-mismatch

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.82***	1.03***	1.07***
	(0.08)	*(0.06)*	*(0.07)*
Numeracy/100	0.23***	0.18***	0.18***
	(0.02)	*(0.02)*	*(0.03)*
Over-skilled	-0.07***	-0.03**	-0.06**
	(0.02)	*(0.01)*	*(0.03)*
Under-skilled	0.01	0.11***	0.07
	(0.05)	*(0.03)*	*(0.06)*
Years of education	0.06***	0.06***	0.07***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.12***	-0.11***	-0.15***
	(0.02)	*(0.01)*	*(0.02)*
Work experience	0.03***	0.03***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.35	0.44	0.38
Observations	2368	2330	2063

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. ** $p < 0.05$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.2

Mincer regressions with skill-mismatch according to Quintini (2012)

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.73***	0.98***	0.99***
	(0.08)	*(0.06)*	*(0.08)*
Numeracy/100	0.26***	0.21***	0,20***
	(0.03)	*(0.02)*	*(0.04)*
Over-skilled	-0.11***	-0.10***	-0.06
	(0.02)	*(0.02)*	*(0.04)*
Under-skilled	-0.01	0.01	0.03
	(0.02)	*(0.02)*	*(0.02)*
Years of education	0.06***	0.06***	0.08***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.11***	-0.11***	-0.15***
	(0.02)	*(0.01)*	*(0.02)*
Work experience	0.03***	0.03***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.35	0.45	0.39
Observations	2383	2333	2063

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.3

Mincer regressions with skill-mismatch according to Allen et al. (2013)

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.72***	0.94***	0.99***
	(0.08)	*(0.06)*	*(0.08)*
Numeracy/100	0.25***	0.21***	0.19***
	(0.03)	*(0.02)*	*(0.04)*
Over-skilled	-0.07*	-0.11***	-0.08
	(0.04)	*(0.03)*	*(0.05)*
Under-skilled	0.03	0.07**	0.05
	(0.03)	*(0.03)*	*(0.03)*
Years of education	0.07***	0.06***	0.08***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.12***	-0.10***	-0.15***
	(0.02)	*(0.01)*	*(0.02)*
Work experience	0.03***	0.03***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.34	0.44	0.38
Observations	2383	2333	2063

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.4

Mincer regressions with skill-mismatch according to the OECD (2013a)

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.70***	0.96***	1.01***
	(0.08)	*(0.05)*	*(0.08)*
Numeracy/100	0.29***	0.21***	0.23***
	(0.03)	*(0.02)*	*(0.04)*
Over-skilled	-0.11***	-0.04*	-0.23***
	(0.03)	*(0.02)*	*(0.05)*
Under-skilled	0.10	0.16***	0.05
	(0.08)	(0.05)	(0.06)
Years of education	0.06***	0.06***	0.07***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.12***	-0.11***	-0.17***
	(0.02)	*(0.02)*	*(0.02)*
Work experience	0.03***	0.02***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.35	0.43	0.39
Observations	2332	2262	2039

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. Members of the armed forces (ISCO 0) and skilled agricultural, forestry, and fishery workers (ISCO 6) excluded. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.5

Mincer regressions with skill-mismatch according to Allen et al. (2013) with three different plausible values

Dependent variable: Gross hourly earnings (log)	Allen (2013) with PVNUM6	Allen (2013) with PVNUM9	Allen (2013) with PVNUM10
Constant	0.73***	0.73***	0.70***
	(0.08)	*(0.08)*	*(0.08)*
Numeracy/100	0.25***	0.25***	0.26***
	(0.03)	*(0.03)*	*(0.02)*
Over-skilled	-0.04	-0.08**	-0.08***
	(0.03)	*(0.04)*	*(0.03)*
Under-skilled	0.03	0.01	0.07*
	(0.03)	*(0.04)*	*(0.04)*
Years of education	0.07***	0.07***	0.07***
	(0.00)	*(0.00)*	*(0.00)*
Gender (female)	-0.12***	-0.12***	-0.12***
	(0.02)	*(0.02)*	*(0.02)*
Work experience	0.03***	0.03***	0.03***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.04***	-0.04***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.34	0.35	0.35
Observations	2383	2383	2383

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.6

Mincer regressions with skill-mismatch according to our newly developed skill-mismatch measure

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.60***	0.81***	0.85***
	(0.09)	*(0.07)*	*(0.10)*
Numeracy/100	0.30***	0.27***	0.28***
	(0.04)	*(0.03)*	*(0.04)*
Over-skilled	-0.17***	-0.12**	-0.17*
	(0.05)	*(0.04)*	*(0.08)*
Under-skilled	0.15**	0.16***	0.23***
	(0.06)	*(0.05)*	*(0.06)*
Years of education	0.07***	0.06***	0.07***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.12***	-0.10***	-0.16***
	(0.02)	*(0.01)*	*(0.02)*
Work experience	0.03***	0.03***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.35	0.44	0.39
Observations	2199	2175	1894

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. Workers in professions with less than 30 observations per country (at two-digit ISCO level) excluded. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

Table A.7

Mincer regressions with skill-mismatch according to Allen et al. (2013) and taking all plausible values into account

Dependent variable: Gross hourly earnings (log)	Germany	Austria	USA
Constant	0.69***	0.89***	0.96***
	(0.08)	*(0.06)*	*(0.07)*
Numeracy/100	0.27***	0.24***	0.22***
	(0.03)	*(0.02)*	*(0.03)*
Over-skilled	-0.10**	-0.13***	-0.10*
	(0.04)	*(0.03)*	*(0.05)*
Under-skilled	0.07*	0.12***	0.10**
	(0.04)	*(0.03)*	*(0.05)*
Years of education	0.07***	0.06***	0.07***
	(0.00)	*(0.00)*	*(0.01)*
Gender (female)	-0.12***	-0.10***	-0.15***
	(0.02)	*(0.01)*	*(0.02)*
Work experience	0.03***	0.03***	0.04***
	(0.00)	*(0.00)*	*(0.00)*
Work experience squared/100	-0.04***	-0.03***	-0.06***
	(0.01)	*(0.01)*	*(0.01)*
R^2	0.35	0.44	0.39
Observations	2383	2333	2063

Notes. Least squares regressions weighted by sampling weights. Sample: Full-time employees between 16 and 65 years of age, excluding students and apprentices. See Section "Empirical Approach" for details of the Mincer regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$. ** $p < 0.05$. * $p < 0.10$. *Data source*: OECD (2013c) and Rammstedt et al. (2014).

methods, data, analyses | Vol. 8(2), 2014, pp. 175-198 DOI: 10.12758/mda.2014.007

The Challenge of Meeting International Data Collection Standards within National Constraints: Some Examples from the Fieldwork for PIAAC in Germany

Anouk Zabal

GESIS – Leibniz Institute for the Social Sciences

Abstract

The *Programme for the International Assessment of Adult Competencies* (PIAAC) is an international OECD study that compares key competencies of adults (16-65 years) in the participating countries. In order to obtain high quality data and to ensure equivalence of measurement across countries, the international PIAAC Consortium produced a very detailed and elaborate set of standards and guidelines for all aspects of the national implementations. In Germany, a comprehensive set of measures and procedures was put in place for the PIAAC fieldwork. Some of the international requirements for data collection were not meaningful within the national context and required certain adaptations. This article describes various key fieldwork measures in Germany and discusses how specific measures relate to central international data collection standards. Reflecting on this national experience, some of the possibilities and limitations of national compliance to international standards are discussed.

Keywords: PIAAC Germany, survey standards, data collection, survey operations, fieldwork

1 Introduction

As a part of the *Programme for the International Assessment of Adult Competencies* (PIAAC), which was initiated and developed by the *Organization for Economic Cooperation and Development* (OECD), a first round of the PIAAC survey was carried out in 24 countries between 2008 and 2013 (OECD, 2013a).[1] The PIAAC survey continued, expanded, and refined the foundations established by two previous international large-scale assessments of adult skills: the International Adult Literacy Survey (IALS, 1994-1998; OECD & Statistics Canada, 2000) and the Adult Literacy and Lifeskills survey (ALL, 2002-2008; OECD & Statistics Canada, 2011; Statistics Canada & OECD, 2005). These further developments included extending the coverage of constructs and assessment domains, and improving the survey methodology. PIAAC strove for excellence at all stages of the survey life cycle and set very ambitious goals for the national implementations and the overall data quality. Towards this aim, the international Consortium responsible for the coordination of the PIAAC survey produced a pre-specified design, a strict timetable, and established a comprehensive program for quality assurance and quality control. As a part of the quality assurance system, an elaborate set of technical standards and guidelines (OECD, 2010) was produced to ensure that appropriate methodologies and rigorous standards be followed by all participating countries. In addition, the international Consortium closely monitored countries' work and their adherence to these technical standards and guidelines. Based on the evaluation of countries' compliance to crucial standards, a final assessment of the *fitness for use* (see Juran & Gryna, 1970; Lyberg & Biemer, 2008) of the PIAAC data for their intended purpose was undertaken by the international Consortium, together with the PIAAC Scientific Advisory Board and the Board of Participating Countries.

The countries participating in PIAAC differ significantly with regard to their type of survey organization (e.g., public organizations such as statistical agencies vs. private survey organizations), national survey practices, available sampling frames, funding, legislation, etc. Because PIAAC is a cross-national survey, the challenge thus lies in defining international survey standards that strike the appropriate balance between enforcing an adequate degree of standardization required

1 The PIAAC survey is sometimes also referred to as the Survey of Adult Skills (OECD, 2013a, 2013b).

Direct correspondence to
Anouk Zabal, GESIS – Leibniz Institute for the Social Sciences,
PO Box 12 21 55, 68072 Mannheim, Germany
E-mail: anouk.zabal@gesis.org

Acknowledgment: I would like to thank two anonymous reviewers for their constructive comments and suggestions.

for cross-national comparability, while allowing for enough degrees of freedom to accommodate differences between countries (see Koch, Blom, Stoop, & Kappelhof, 2009).

For the participating countries, translating and adapting international survey operation standards into a smoothly functioning, well-balanced, and coherent set of appropriate national measures was one of the major challenges of fieldwork. When international standards differ considerably from usual national practices, changing from tried-and-tested procedures to new, internationally prescribed ones can be risky. In extreme situations, it may even jeopardize national survey operations. That being said, novel measures and procedures can lead to innovation in national survey methods and impact positively on national survey cultures and best practices.

This article discusses efforts invested in the quality assurance and the quality control of survey operations for the main PIAAC data collection at the international level, and focuses on how some of these international standards were realized and elaborated upon, at the national level, for PIAAC in Germany. I will describe several of the key international data collection standards in PIAAC, provide an overview of the comprehensive set of measures and procedures implemented for the German fieldwork, and consider the possibilities and limitations of national compliance to international standards.

2 The PIAAC Data Collection Standards

Three chapters of the international standards and guidelines for PIAAC pertain to fieldwork survey operations (OECD, 2010): (1) field management standards, (2) data collection staff training standards, and (3) data collection standards. They address the selection, organization, and, in particular, the in-person training of the data collection staff, the data collection itself, including contact procedures, and the monitoring and quality control of fieldwork. The standards generally reflect best practices in survey operations, and the guidelines elaborate on the implementation of these standards. The approximately 65 standards and 120 guidelines, in addition to the further recommendations specified for survey operations in these chapters, go significantly beyond the breadth and depth of standards and procedures established for the precursor surveys IALS (Murray, Kirsch, & Jenkins, 1998; OECD & Statistics Canada, 2000) and ALL (OECD & Statistics Canada, 2011; Statistics Canada & OECD, 2005). The academically based, methodologically rigorous European Social Survey (ESS) also has a comprehensive set of specifications for fieldwork (e.g., European Social Survey, 2011, 2013). Many of the PIAAC standards and the ESS specifications overlap. However, the PIAAC standards and guidelines are even more elaborate than the specifications of the ESS.

Minimizing nonresponse error and increasing the rate of survey participation is at the heart of any quality survey design (see Groves & Couper, 1998). Not surprisingly, one of the most central and challenging standards in the PIAAC standards and guidelines was related to the response rates (OECD, 2010): Countries were required to achieve at least 70% overall response; however, the standards also indicated that a 50% response rate or higher would also be acceptable if the results of subsequent nonresponse bias analyses showed no evidence of significant bias. In addition, the PIAAC standards targeted a maximum non-contact rate of 3%. Both the targeted response rate and the maximum non-contact rate are the same as in the ESS (European Social Survey, 2011, 2013). In PIAAC, a data adjudication process evaluated the quality of each national data set and determined whether any limitations on the release of the data or in the international reporting should be put into effect (OECD, 2013c, Appendix 7). Response rate standards were an important element in this evaluation process. In contrast, in the ESS, although deviations from response rate standards are documented, not achieving the prescribed response rates does not have direct repercussions for the national data releases.

Many of the PIAAC standards and guidelines for fieldwork operations and data collection represent best practice methods and procedures to be implemented as a comprehensive strategy towards reaching this golden goal for response rates and to minimize nonresponse bias. Furthermore, they aim at reducing the measurement error and achieving the overall goal of collecting high-quality, internationally comparable data. Key international PIAAC standards for survey operations specified by OECD (2010) include, for example:

(a) close monitoring of data collection at all stages,

(b) attractive remuneration of interviewers that is independent of the number of completed interviews,

(c) extensive in-person interviewer training,

(d) at least four in-person contact attempts before coding a case as a non-contact,[2]

(e) thorough documentation of contacting attempts and results,

(f) no substitution of selected individuals whatsoever; use of interpreters/translators acceptable for the administration of the background questionnaire (not, however, for the cognitive assessment),

(g) standardized administration of the survey instruments on laptops complying to specific hardware and software specifications,

(h) development of a national best practice strategy to maximize response rates,

(i) implementation of effective refusal conversion strategies, and

2 This is the standard for countries initially contacting the sample persons in person, which is recommended.

(j) verification of at least 10% of each interviewer's work (random selection of all dispositions, including cases of nonresponse).[3]

The PIAAC standards and guidelines, extensive further documentation and material, and in-person training sessions for the National Centers, were crucial elements of the PIAAC quality assurance plan for the data collection. Compliance to key international standards was closely monitored by the international Consortium. Any proposed deviations from these standards required approval by the international Consortium. As a part of the quality control process, countries were required to fill out numerous forms and to provide information at regular intervals to keep the international Consortium updated about all aspects of national implementation and progress.

3 Key Facts about the PIAAC Data Collection

As described in more detail in OECD (2010, 2013a, 2013b, 2013c), the PIAAC interview consisted of a background questionnaire administered as a CAPI (*computer-assisted personal interview*) followed by a cognitive assessment (per default with a computer-based administration, but with the option of a paper-based administration, if required). All participating countries carried out the PIAAC interview face-to-face. In general, the interview took place at the respondent's home and was designed to take approximately 90 minutes. It was administered in the national language(s). For the background questionnaire, it was possible to recruit an interpreter to translate the questions.[4] For the assessment, absolutely no help was allowed. Respondents worked on the cognitive assessment tasks on their own and without any time limitations. The cognitive assessment represented a non-standard requirement for both interviewers and respondents. The target population consisted of adults between 16 and 65 years of age who were non-institutionalized and were living in the country at the time of the data collection period.[5] Countries needed to realize a probability-based sample representative of the target population. Substitutions of sampling units were not permitted at any stage.

Germany participated in the first round of the PIAAC survey, and the national implementation of the PIAAC survey was the responsibility of the German National Center at GESIS – Leibniz Institute for the Social Sciences (Rammstedt, 2013).[6]

3 In addition, 100% validation of any interviewer whose work was suspect was required.

4 The interpreter could be a family member, for example.

5 The target population was defined irrespective of nationality, residential status, or language skills.

6 The German National Center was appointed and funded by the Federal Ministry of Education and Research with the participation of the Federal Ministry of Labor and Social Affairs.

PIAAC in Germany included all domains of the cognitive assessment, i.e., literacy and numeracy, as well as the international options problem solving in technology-rich environments and reading components. Thus, the required sample size consisted of at least 5,000 cases. As indicated in the national technical report (Zabal et al., 2014), which gives a comprehensive account of all aspects of the German implementation, a registry-based sample with a two-stage stratified and clustered sampling design was realized, with 320 sample points (in 277 municipalities) and a gross sample size of 10,240 target persons. The eight-month data collection period started on 1 August 2011 and terminated on 31 March 2012. Following the PIAAC definition for a completed case (OECD, 2010), a realized sample size of 5,465 respondents was achieved in Germany. The official design-weighted final response rate for Germany (according to the PIAAC response rate definition) was 55% (Mohadjer, Krenzke, & Van de Kerckhove, 2013).[7]

4 Overview of the Fieldwork Measures in PIAAC Germany

In Germany, the data collection was subcontracted to TNS Infratest, a renowned survey organization with extensive experience in conducting face-to-face national probability-based surveys to high standards. Careful thought went into specifying a set of best practice standards and procedures for the data collection that would optimize national fieldwork and adhere as closely as possible to the PIAAC standards and guidelines, to ensure comparability and equivalence across the PIAAC countries. In order to enforce compliance with the PIAAC standards and guidelines, these were included as an appendix to the contract with the survey organization, thus emphasizing that the PIAAC data collection would entail departures from routine procedures. However, the implementation of new methods and procedures needed to be feasible within the survey organization's general organizational structure and working framework. Although the PIAAC specifications and recommendations coincided, in many instances, with best practice in Germany and in the survey organization, there were other instances where adaptations, compromises, and innovations in implementation were required.

Figure 1 shows the key elements of the German fieldwork measures. A number of these fieldwork measures are common practice for high-quality national surveys, although in general, not all measures are realized in one survey. The outstanding characteristic of the PIAAC fieldwork in Germany is that it unites a large number of measures, these measures were often undertaken with unusual intensity, and some novel methods were introduced.

---

7 The non-contact rate for Germany was 3.4% and thus only slightly above the required standard (Zabal et al., 2014).

Interviewers ▪ *129 experienced interviewers with excellent track records* ▪ *Five-day interviewer training* ▪ *Assigned exclusively to PIAAC for four weeks* ▪ *Attractive interviewer remuneration, including an add-on for large cities*	**Quality Control** ▪ Thorough fieldwork monitoring by both survey organization and German National Center ▪ *Extended* interview validation ▪ *Monitoring of field performance (e.g. audio tapes)*
Contacting and Gaining Cooperation ▪ First contact in-person ▪ Four contact attempts minimum ▪ Documentation of contact attempts and further information ▪ Refusal conversion ▪ *Tracing respondents who had moved*	**Incentives** ▪ Attractive conditional incentive of *50 €* ▪ Small non-monetary unconditional incentive ▪ *Discretionary at-the-door non-monetary incentives for refusal conversion phase*
Introductory and At-the-Door Study Materials ▪ Advance letter ▪ *Brochure* and flyer ▪ *Endorsement letter and tailored letters for refusal conversion phase* ▪ *Folder with press clippings*	**Public Relations** ▪ Press releases and *targeted PR work* ▪ Study website ▪ Toll-free hotline for respondents

Note. New measures or procedures that, in some way, went beyond standard national practice are shown in italics.

Figure 1 Key elements of the fieldwork measures in PIAAC Germany.

Fieldwork was subdivided into two main working phases and five re-issue phases (see Zabal et al., 2014). Continuous and meticulous monitoring of interviewers' work is an important aspect of survey quality control and crucial to reducing interviewer error.[8] In PIAAC Germany, monitoring took place at various levels: (a) checking that assignments were being worked on as required (e.g., checking individual response, non-contact, and refusal rates), (b) checking the quality of the interview administration (e.g., reviewing the survey data, reviewing audio tapes), (c) validating the interviews (checking for falsifications), and (d) checking the demographic composition of the realized sample and monitoring nonresponse bias. At

8 This requires ongoing collection of information on interviewers' performance, the evaluation of this information, and providing interviewers with prompt feedback (see Fowler & Mangione, 1990).

the German survey organization, eight supervisors were responsible for the day-to-day operational tasks (e.g., case assignments and re-assignments, communications and instructions regarding fieldwork procedures, and feedback) and they closely monitored the interviewers. The survey organization also provided regular and detailed updates to the German National Center, which carried out further quality control and monitoring. The German National Center and the survey organization worked together closely during the entire duration of the fieldwork. In addition, regular monitoring reports were provided to the international Consortium. Issues identified during monitoring were promptly addressed with the required corrective actions.

The next sections will focus on the following subset of the key fieldwork measures listed in Figure 1: Interviewer selection, interviewer remuneration, interviewer training, incentives, contacting and gaining cooperation, and interview validation. Where appropriate, the discussion addresses tensions between the international PIAAC standards and national survey operations. A more comprehensive account of the PIAAC fieldwork, including some of the survey materials and fieldwork results, is provided in Zabal et al. (2014).

Interviewer Selection

Interviewers implement the survey design directly in their contacts with the respondents and are crucial to the quality of the survey data. With regard to the selection of the data collection staff, the PIAAC standards and guidelines recognized that numerous country-specific factors influenced the recruitment and required number of interviewers, and therefore strict standards were not prescribed. Instead, a number of considerations that countries needed to take into account were noted (Montalvan & Lemay, 2013a).

The German survey organization had a large pool of freelance interviewers at its disposal. Only experienced face-to-face interviewers with an excellent track record in the administration of high-quality registry-based CAPI surveys were considered for selection for the PIAAC survey. In addition to their experience with interview administration, interview protocols, record-keeping, and organizing their own work, these interviewers had strong interpersonal and communication skills. Experienced interviewers are more likely to be successful in gaining respondent cooperation (Groves & Couper, 1998). The selection process also took interviewers' availability for training and their availability during the eight-month fieldwork period into account (interviewers had to be able to handle their assigned workload reliably). The geographical location of the interviewers, i.e., their proximity to sample points, was also a selection factor. Only local interviewers were recruited, to maximize the number of call attempts made per case while reducing travel costs. Furthermore, local interviewers who are familiar with the area and the local dialect

and customs may achieve higher response rates than non-local interviewers (Alcser & Clemens, 2011). Several factors were considered in establishing the number of interviewers to be selected for PIAAC. In terms of reducing interviewer effects, a large number of interviewers was desirable. However, given the five-day interviewer training and the special laptop requirements for PIAAC (which necessitated the purchase of new laptops), there were pragmatic restrictions. Thus, a total of 130 freelance interviewers was selected.[9] Most interviewers were over 50 years of age and had more than three years' experience working for the survey organization; almost 30% had more than 10 years' tenure (for more information on interviewer characteristics, see Ackermann-Piek & Massing, in this volume).

Interviewer Remuneration

Interviewer payment schemes can vary significantly across different countries and cultures (Alcser & Clemens, 2011). As a consequence, rigid standards regarding interviewer remuneration in cross-national surveys may be quite challenging, especially since specific survey organizations are unlikely to depart from their firmly established interviewer payment practices (Alcser & Clemens, 2011; Stoop, Billiet, Koch, & Fitzgerald, 2010). There are basically two standard interviewer payment arrangements: one is based on payment per completed interview, the other on an hourly rate. The advantages of a per piece payment scheme are that it is easier to monitor and it facilitates the estimation and control of interviewer costs. Paying an hourly rate is equitable in that interview length can vary substantially. Furthermore, it provides interviewers with an incentive to invest time in chasing target persons who are hard to reach or generally more reluctant to participate in surveys, and also compensates interviewers for time spent on administrative tasks and record-keeping. Finally, it discourages interviewers from speeding through the interview and undermines interviewer satisficing strategies associated with "sloppy" work. The PIAAC standards and guidelines regarding interviewer remuneration prescribed a payment per hour. The payment was to reflect the length and complexity of the PIAAC interview and be attractive in comparison with other national surveys (OECD, 2010).

As mentioned above, interviewers work on a freelance basis for the German PIAAC survey organization, as is generally the case in Germany. Consequently, the established payment for face-to-face surveys is per piece. This is markedly different from the usual practice in the United States (whose best practices in data collection shaped several PIAAC standards and guidelines), where interviewers are generally paid an hourly rate (Rosen, Murphy, Peytchev, Riley, & Lindblad, 2011). Despite

9 Due to one case of interviewer attrition prior to the start of fieldwork, 129 interviewers actually worked on the German PIAAC survey.

the weight carried by the PIAAC standards and the importance of the PIAAC survey, such a fundamental deviation from the survey organization's standard interviewer remuneration was one aspect of fieldwork which could not be changed.

As a consequence, a unique mixed payment scheme was developed for PIAAC in Germany (see Zabal et al. 2014). It consisted of three main components: (a) an attractive base rate for each completed interview, (b) an additional payment for interviews undertaken in large municipalities, and (c) an hourly payment component for interviews that were particularly long.[10] The base rate per completed interview was higher than in other comparable national surveys and took into account the length and the complexity of the PIAAC interview, as well as time demands made by contact documentation tasks. The add-ons for large municipalities were introduced as a compensation for the increased interviewer burden in urban regions. In urban areas, sample persons more frequently live in dwellings with access impediments than in rural areas, and they are also less frequently at home. Thus, the add-ons for large municipalities were intended to achieve a fair, or fairer, payment across interviewers by providing additional compensation for sample points in areas with generally lower response rates and, thus, with higher interviewer burden. The hourly component for long interviews ensured that interviewers would take the time actually needed for the interview and not "rush" through. This was a crucial component, especially given that the cognitive assessment is at the heart of the PIAAC survey, and the assessment was administered without any time restriction whatsoever: Respondents worked on the cognitive tasks at their own pace and could take as long as they liked.

Interviewer Training

In PIAAC, interviewer training was regarded as a crucial feature of cross-national survey operations and as an effective tool for improving the quality of interviewers' work. Due to the complexity of the PIAAC survey, the challenging response rate goals, the importance of the PIAAC protocols both for the administration of the background questionnaire as well as for the administration of the cognitive assessment, and also given that the interview was delivered on a novel technological platform, the PIAAC international Consortium prescribed a five-day interviewer training.[11] Such an extensive interviewer training was a challenging novelty in Germany, where interviewer trainings are typically much shorter, if provided at all (see Zabal et al., 2014).

---

10 In addition, all travel costs were reimbursed.

11 For interviewers with specific profiles (experience in PIAAC field test, experience with other surveys) somewhat reduced training loads were regarded as acceptable (Montalvan & Lemay, 2013a).

At the international level, interviewer training was provided according to a train-the-trainer model (similar to the procedures in the Survey of Health, Ageing and Retirement in Europe, SHARE; see Alcser & Benson, 2005) that aimed at ensuring consistency of training across all participating countries and hereby optimizing the standardization of interviewer behavior and survey procedures and, ultimately, ensuring the cross-national comparability of the data. The international PIAAC Consortium trained the trainers (members of the national centers, and, if possible, field directors) as if they were the interviewers and provided countries with the full set of scripted material to be translated and adapted by national centers and subsequently used for their national trainings. For some of the material, training contents required relatively extensive national adaptation (e.g., administrative survey procedures, some aspects of the background questionnaire training), whereas, for other material, any national tailoring was strictly limited, if allowed at all (e.g., administration of the cognitive assessment). In Germany, a decision was made to depart from a decentralized training solution and to have the same trainer team instruct all training groups. Training was conducted immediately before interviewers started their fieldwork to allow them to directly apply and consolidate the procedures they had learned during training.[12]

Interviewer training generally addresses two basic aspects of interviewers' work: (a) contacting target persons and gaining cooperation, and (b) the interview administration according to survey protocols. Interviewers are required to carry out a wide variety of tasks requiring both adaptive behavior as well as the capability of adhering to standardized procedures. Adaptive behavior is essential for gaining the cooperation of the sample, whereas the measurement process itself requires the ability to follow prescribed procedures in a standardized way, although interviewers also sometimes need to adapt appropriately to certain situations during the interview (Lessler, Eyerman, & Wang, 2008). Interviewer training contributes to increased survey data quality by sensitizing interviewers to respondents' concerns and to the importance of tailoring their own responses (Groves & McGonagle, 2001), as well as by decreasing item nonresponse and increasing the amount and precision of the collected information (Billiet & Loosveldt 1988).

There is no single best way to address sample persons (e.g., Groves & Couper, 1998; Groves, Singer, & Corning, 2000). Instead of using a rote introduction, it is important for interviewers to *tailor* their behavior to specific respondent characteristics and concerns, and to apply strategies to *maintain interaction* and minimize the likelihood of evoking a *no* to the survey request (see Groves & McGonagle,

12 The majority of the German interviewers participated in the full training program for interviewers with experience in other surveys (31 hours of in-person training in five days); interviewers with PIAAC field test experience took part in a reduced training (22 hours of in-person training in three days).

2001).[13] Accordingly, during the session on gaining respondent cooperation in the German PIAAC interviewer training, interviewers practiced recognizing respondents' concerns and how to adapt their responses to specifically address these (i.e., how to tailor their responses). This included handling PIAAC-specific concerns, such as reluctance to complete the assessment. Although only experienced interviewers were assigned to PIAAC in Germany, this session was found to be invaluable, because it offered interviewers the opportunity to exchange notes and to expand their own repertoires and awareness.

Despite the fact the interviewers had extensive CAPI experience, one of the focuses of training was to specifically review the PIAAC background questionnaire and the required interviewing protocols. Groves et al. (2004) indicated that there is some evidence suggesting that experienced interviewers are not as compliant as new interviewers in reading the questions verbatim and adhering strictly to protocols. Thus, training firmly stressed the need for standardization as an important measure in reducing interviewer variance. Although the background questionnaire was developed with a view to minimizing interviewer discretion, probing, i.e., reiterating or rephrasing a question (see Cannell, Marquis, & Laurent, 1977), may sometimes be required if respondents answer inadequately. Appropriate probing techniques were also reviewed in this session.

In other sessions, interviewers were extensively trained on the administration of the cognitive assessment, which was a novel, non-standard task for them. The role change from that in the administration of the background questionnaire was an important focus. Whereas interviewers were active during the questionnaire administration, their role during the cognitive assessment was quite different. Here, interviewers had to create a quiet and supportive atmosphere, and, with the exception of technical problems, refrain from helping the respondent in any way.

Overall, training gave interviewers extensive and well-grounded knowledge about the background of the PIAAC survey, the PIAAC procedures, all components of the PIAAC interview, and the comprehensive set of materials required for the PIAAC interview. Key PIAAC standards were carefully reviewed and all measures of quality control were described in detail, to ensure full transparency. Training also included practice sessions on how to handle the novel international software. Hands-on practice exercises were found to be a crucial component of the interviewer training. Trainers circulated throughout practice interviews to observe and evaluate how interviewers were conducting the interview and whether there were any knowledge gaps to be filled or misunderstandings to clarify. Interviewers' evaluations of the training were very positive, both in the direct evaluation after training and in hindsight, as reported during the debriefings after fieldwork.

13 Interviewer training focusing on such refusal aversion strategies has been found to have certain positive effects on cooperation rates, also in face-to-face surveys (O'Brien, Mayer, Groves, & O'Neill, 2002).

Incentives

In view of the high response rate targets and the considerable respondent burden associated with the PIAAC interview, the standards and guidelines encouraged countries to consider using incentives. All planned incentives had to be signed off by the international Consortium prior to fieldwork. Whether or not a country finally opted to use an incentive for PIAAC was left to that country's discretion; in some countries, the use of incentives was not possible due to national regulations. Although prepaid incentives can be effective in increasing survey cooperation (Singer, 2002), the use of prepaid incentives for PIAAC in Germany was rejected from the outset, because prepaid incentives were not considered to be a justifiable use of taxpayers' money and are also liable to increase mistrust in and public criticism of the survey (see Börsch-Supan, Krieger, & Schröder, 2013). Three conditional incentive conditions were tested in the German PIAAC field test, to determine the best incentive for the main survey (a 10 € commemorative coin, 25 € in cash, or 50 € in cash).[14] Following evaluation of the results of the German field test incentive experiment, the largest of the tested incentives, the 50 € conditional cash incentive was chosen for the main survey (for details, see Martin, Helmschrott, & Rammstedt, in this volume). In addition, a non-monetary unconditional incentive (post-it notes featuring the PIAAC logo) was attached to the advance letter. In the re-issue phases, interviewers were given the option of deploying discretionary at-the-door non-monetary incentives. The 50 € incentive constitutes a substantive sum, in comparison to incentives offered by other national surveys (see Pforr et al., forthcoming). It reflects not only the national importance of the PIAAC survey, but also the substantial length of the interview, and acknowledges that participating in an assessment is an unusual and, for some, daunting aspect of the interview.

Contacting and Gaining Cooperation

Before any interview can be carried out, the interviewer has to locate the target person, establish contact, and gain their cooperation. The majority of the PIAAC standards and guidelines relating to contacting and callback rules, study materials and outreach tools, as well as techniques for dealing with nonresponse cases were in line with many national best practices in Germany. The contacting rules for PIAAC in Germany ascertained that at least four in-person contact attempts be made before a non-contact could be coded, with calls to take place at different times of the day and on different days of the week, to accommodate varying at-home patterns and facilitate reaching difficult-to-contact sample persons. In

14 All countries participating in Round 1 of PIAAC conducted a field test in 2010. Some information on the German PIAAC field test is provided in Zabal et al. (2014).

addition, interviewers were required to record each contact attempt and the disposition of each contact outcome. Prior to the first contact attempt, an advance letter, accompanied by a study flyer and the unconditional incentive, was sent to the sample person. Contrary to the survey organization's usual practice, a staggered mailing schedule was implemented that was individually attuned to each interviewer's personal contacting schedule, in an effort to reduce the time interval between receipt of the advance letter and the interviewer's first visit. Another new measure at the survey organization consisted in assigning interviewers exclusively to the PIAAC survey for four weeks during the first phase of fieldwork. Furthermore, the German National Center undertook considerable efforts to produce attractive study materials (e.g., not only a flyer but also a brochure), and, in targeted public relations activities, to increase the visibility of the PIAAC survey in Germany and emphasize the legitimacy of the interviewer's request (study website, toll-free hotline for respondents, press releases, and the targeted dispatch of the press releases to local newspapers in the PIAAC sample points).

During the re-issue phases, only a subset of the refusals could be re-approached, due to German legislation; "hard refusals" could not be re-contacted. For those cases that could be re-issued, additional refusal conversion measures were introduced: (a) tailored refusal conversion letters reinforcing specific aspects of the survey, (b) extended at-the-doorstep material that included an endorsement letter and translations of the advance letter and FAQs into five languages, (c) discretionary at-the-door non-monetary incentives, (d) re-assignment of interviewers, and (e) a selective deployment of travelling interviewers to difficult sample points.

One of the issues unique to countries with registry samples is that the selected addresses may be obsolete by the time a contact with the sample person is attempted. For example, persons who had re-located may not have correctly deregistered and re-registered. This problem is exacerbated in countries such as Germany that do not have a central population register but have nationally distributed registry offices. In registry-based high-quality surveys in Germany, it is common practice to classify cases with address-related dispositions as ineligibles (e.g., in the German General Social Survey, ALLBUS; see Wasmer, Scholz, Blohm, Walter, & Jutz, 2012). However, this was not an option in the context of the PIAAC standards. Instead, it was necessary to undertake special efforts to trace respondents who had moved or whose addresses proved to be invalid. To cope with this situation, a new procedure was introduced: cases with unresolved address-related dispositions[15] were collected at home office, and the registry offices were subsequently re-contacted with a request for updated information. This approach proved to be useful; for details, see Zabal et al. (2014). In addition, and contrary to common practice, respondents

15 Resolved address-related dispositions were cases in which the sample person had moved outside of the country or for which the interviewer was successful in obtaining a new address.

who had moved to non-PIAAC sample points were also pursued (within certain feasibility limits).

Interview Validation

Interviewer falsification denotes intentional interviewer deviations from the survey protocols, such as the fabrication of interviews (or parts thereof), the substitution of sample persons, misreporting disposition codes, or taking shortcuts through the interview (see Groves et al., 2004). One of the most important PIAAC standards with regard to identifying possible falsifications stated that 10% of each interviewer's finalized work had to be verified, including final nonresponse dispositions (OECD, 2010). In addition, one of the guidelines for this standard stipulated that cases for verification should be selected at random from all sampling units (including both respondents and nonrespondents). For PIAAC Germany, this standard and guideline were fundamentally problematic. The survey organization's common validation practice is to validate *all completed interviews*, and only these. This strategy is based on two considerations. First, because one of the potential drawbacks associated with the usual per piece payment is a higher risk of interviewer falsification (Rosen et al., 2011), the focus is clearly on identifying any potentially falsified interviews. Second, German legislation prohibits re-approaching hard refusals. After intensive deliberation, it was decided that departing from the survey organization's well-established validation procedures posed too great a risk. Thus, its standard validation procedure was adopted as a starting point. It essentially consists of sending all respondents a validation questionnaire by mail, and in exploiting one of the advantages of using a registry-based sample by checking the consistency of interview data with the basic information provided by the register when the sample is drawn (age, gender, and nationality). Furthermore, it includes a number of additional checks (e.g., interview time and length). Conspicuous cases are systematically followed up.

For PIAAC, the back-checks were extended to include other dispositions – as far as feasible, but with definite limitations, e.g., hard refusals could not be validated by law. Concretely, attempts were made to validate:

(a) certain ineligibles: via an internet search (ineligibles due to institutionalization),

(b) refusals due to disabilities: via a mail validation questionnaire,

(c) non-contacts: through a concerted telephone action, and

(d) soft refusals: in person.

Although the standard validation procedures worked smoothly and ensured that at least 10% of each interviewer's work was successfully validated, the attempted extensions of the validation scheme yielded only very modest returns. Checking

the ineligibles due to institutionalization proved to be practicable. The back-checks of refusals due to disabilities were not found to be advisable, due to certain ensuing ethical issues. The attempt to reach and validate non-contacts by phone was relatively unsuccessful, and the procedure for validating a non-contact was generally debatable. The main focus in re-approaching soft refusals remained refusal conversion (and not validation). Furthermore, the attempt to validate soft refusals in person during the refusal conversion phase did not work as smoothly as intended. With respect to the random selection specification, because all completed cases and all disabilities were selected for validation, the complete selection was superior to selecting a random subset. Given the legal restrictions in re-approaching hard refusals, a random selection from all dispositions was not possible in Germany.

Beyond carefully reviewing the survey organization's quality control results, the German National Center carried out a set of validation measures that complemented the basic validation described above. These included (a) monitoring the date and time of the interview and number of interviews per day, (b) monitoring the length of the interviews to identify suspicious outliers, especially scrutinizing very short interviews, (c) checking some interviews for routing shortcuts, (d) reviewing item nonresponse rates, (e) reviewing the quality of the entered responses to certain open format questions, and (f) checking the quality of the interviewers' scoring.[16]

Reviewing audio tapes of actual interviews provides direct information on the interview process (Fowler & Mangione, 1990). The PIAAC standards specified that each interviewer had to submit two tape recordings of interviews early on during fieldwork, with subsequent review of the recordings (OECD, 2010). This specification addressed the need to check if the PIAAC protocols and procedures taught in training were being applied appropriately. However, it should be noted that the use of tape recorders may lead interviewers to perform better (Billiet & Loosveldt, 1988). Although it was a non-standard requirement within the German framework for fieldwork, the vast majority of interviewers did, in fact, deliver audio tapes for monitoring. These audio tapes (specifically, the recording of the background questionnaire administration) were systematically reviewed at the German National Center. If deviations from the protocols were found, for example incorrect use of show cards or a tendency to not read each question fully and accurately, the survey organization was contacted with instructions to re-train specific interviewers on the identified issues. Ackermann-Piek and Massing (in this volume) describe these audio tape reviews in more detail and provide some evaluations of the interviewers' behavior.

16 Scoring denotes the evaluation of responses to cognitive tasks and coding them as correct or incorrect. One of the more difficult and training-intensive PIAAC interviewer tasks involved scoring responses to eight core assessment tasks that were part of the paper-based assessment.

Interview validation inspected the overall patterns of all these measures and closed followed-up on any conspicuous constellations. In Germany, no instance of falsification was detected. Further information on the interview validation and fieldwork quality control in PIAAC Germany can be found in Massing, Ackermann, Martin, Zabal, and Rammstedt (2013) and in Zabal et al. (2014).

5 Discussion

Section 4 describes key parameters of the German fieldwork strategy for PIAAC. These included the best possible and most comprehensive set of fieldwork measures that would work well within the national context, within the context of the national survey organization, and within the framework established by the PIAAC standards and guidelines. These procedures were followed rigorously during all phases of data collection, to obtain results of the highest possible quality. Overall, the German fieldwork strategy worked well and was effective in reaching national and international data collection goals. In the light of the general decline in response rates for face-to-face surveys in Germany and many other countries (e.g., Blohm & Koch, 2013; de Leeuw & de Heer, 2002), the achieved weighted response rate of 55% for PIAAC in Germany can be regarded as a particularly successful outcome.

Groves and Couper (1998) describe survey participation as a function of several factors that are grounded in features of the survey design, environmental features, individual characteristics of the sample persons, as well as in characteristics of the interviewer and the interviewer-sample person interaction. Thus, a wide variety of factors may affect a sample person's decision to participate in a survey, ranging from the survey climate, and the trustworthiness and respectability of the sponsor, to the subjective burden associated with the survey request, or the appeal of the offered incentive. Some of the international design specifications of the PIAAC survey that were beyond the control of the national implementation are potentially detrimental to gaining cooperation; for example, the interview length, and, at least to some extent, the request to participate in a cognitive assessment. On the other hand, a number of other factors are especially favorable, such as the long data collection period, and, in particular, the prominence of PIAAC, its international dimension, and its political relevance.

For Germany, PIAAC was a survey of particular national importance. This was a decisive factor that impacted on the national implementation, both directly and indirectly. Due to the priority of PIAAC, it was well funded and, as a consequence, the range of possible measures and interventions was larger than usual. This was an important element in realizing the sophisticated fieldwork strategy required to achieve internationally comparable and high-quality survey results. It also made it possible to offer an unusually attractive incentive of 50 €. Further-

more, the survey organization clearly acknowledged its internal prioritization of the PIAAC survey. It was therefore possible to initiate more novel components and modifications to standard procedures than usual. It should be noted that including the comprehensive PIAAC standards and guidelines as a part of the contract with the survey organization seems to have significantly contributed to triggering improved methods and departing from routine practice.

As previously indicated, while many of the PIAAC standards reflect national best practice, others do not. At the onset of the PIAAC survey, there were certain misgivings about the feasibility of a number of these standards in Germany. In some cases, these reservations proved to be wrong. For example, both the necessity and the feasibility of a five-day interviewer training were questioned. However, it turned out to be both necessary and feasible. The five days were indeed needed to review and transmit all the relevant information regarding the complex PIAAC interview, and to ensure that interviewers could smoothly bring together all the various components and procedures. The length of training was also justified by the need for standardization across the participating countries, by the introduction of diverse novel elements to the interviewers' work, and by the deviations from their usual practice. The latter should not be underestimated: Lynn (2003, p. 330) emphasizes that "The potential for errors and mistakes when people used to doing things one way are asked to do them in a slightly different way is considerable." Beyond these objective reasons, the interviewer training in Germany was found to have unforeseen and very positive motivational side-effects for interviewers and home office staff. Spending five intensive days together contributed to a sincere team building between all players – interviewers, supervisors, field directors, members of the German National Center – and created a strong identification with the PIAAC survey and its aims. To sum, in hindsight the interviewer training was vital for fieldwork success. However, this is not an appeal to widely implement five-day interviewer trainings for all German surveys. Many surveys will have neither a pressing need (in terms of the complexity of the interview and protocols), nor the resources for such (extended) in-person trainings. If extensive in-person trainings became a standard, they would also no longer have the unique motivational effect that they had for PIAAC. However, it is important to emphasize that even experienced interviewers can profit greatly from training on gaining cooperation and on standardized interviewing techniques.

Another example of a standard that was first thought to be problematic in the German context was the requirement to obtain audio tapes of actual interviews. Contrary to the initial forecasts, being asked to audio tape an interview did not cause significant friction, neither with the interviewer, nor the respondents. Admittedly, it remains unclear how well this would have been received without the PIAAC training, the weight of the international PIAAC standards, and the attractive conditional incentive. From the point of view of quality control, this direct monitoring

is especially suited to identifying interviewer mistakes in administering the interview. This aspect of quality control is not pursued in many national surveys, and experience with PIAAC shows that even very experienced interviewers can deviate from standardized survey protocols, and that monitoring the CAPI administration and providing timely feedback is important. It would therefore be recommendable to consider adopting this quality control element in other national surveys. However, it should be noted that reviewing the audio tapes required significant personnel resources at the German National Center, and that not all surveys will have the capacity needed for this work.

Some of the PIAAC data collection standards remained unfeasible in the German context, despite endeavors to achieve compliance. For example, the central component of the national interviewer remuneration remained a per piece and not a per hour payment. However, the national extensions to the standard remuneration practice captured the spirit of the international standards, which, in essence, consisted in providing an attractive and equitable payment for all aspects of the interviewers' tasks. In this case, it represented a viable compromise between the international requirements and national possibilities.

The case is different for the interview validation scheme. Here, national legislation and well-established validation procedures were diametrically opposed to the international standards. Without intending to imply that the national validation strategy cannot be improved upon, it is a strategy that harmonizes with other national fieldwork elements and makes sense in the German national context. Quality control back-checks are such a crucial element of fieldwork that completely changing the validation approach for one survey is neither feasible nor recommendable. The risk involved in departing completely from well-established procedures of this importance is significant. Based on the traditional national validation approach, and with every effort made to put in practice the entire array of additional possible checks, as well as introducing completely new ones to extend the range of validated dispositions, validation in PIAAC Germany was very thorough and comprehensive. However, the attempt to match the international validation scheme more closely by implementing new quality control back-checks for non-complete dispositions did not work very well.

From an international perspective, the detailed information provided by Montalvan and Lemay (2013a) about several aspects of fieldwork operations in the PIAAC Round 1 countries presents a useful overview of variations across countries. Montalvan and Lemay (2013b) also describe the quality assurance and quality control activities for the PIAAC survey operations and conclude that countries' compliance with the quality control program was high. As mentioned above, the comprehensive quality control mechanisms put in place for PIAAC culminated in a final data adjudication process. The development of the adjudication framework and the selection of indicators were undertaken relatively late in the survey life-

cycle. OECD (2013c, Appendix 7) describes the process and results of the final data adjudication. This data adjudication went beyond the mere evaluation of compliance with PIAAC standards. It aimed at evaluating the overall quality of the PIAAC data in terms of their "fitness for use", i.e., to assess whether the quality of the data was sufficient for the intended use (e.g., to inform policy-makers, for scientific purposes), or whether restrictions needed to be imposed on the dissemination and use of the data. Data collection was one of the four core domains scrutinized in the final data adjudication process; the other domains were sampling, coverage and nonresponse bias, and instrumentation. Each domain was evaluated according to a set of indicators, with three possible outcomes (pass, caution, or fail) that reflected whether the relevant requirements were fully met, met to an acceptable extent, or generally not fulfilled. The German data collection was given a pass, with a comment that the validation strategy met a reduced requirement (OECD, 2013c, Appendix 7). The requirements regarding response rates and coverage rates were considered as a part of the data adjudication domain "coverage and nonresponse bias". For Germany, the data adjudication report noted a caution for this domain but indicated that the extended nonresponse bias analysis "provides evidence that bias was reduced through the weighting adjustments" (OECD, 2013c, Appendix 7, p. 70). It is interesting to note that while the results of the nonresponse bias analyses were clearly essential for the evaluation of this domain, only the five countries with a final weighted response rate of 70% or above were given a pass. All countries with a final weighted response rate below 70% were assigned a caution (as was the case in Germany).

6 Conclusions and Outlook

Surveys such as PIAAC that strive to achieve cross-national comparability and produce data of the highest possible quality by implementing an effective system of quality assurance and control, and that receive high priority at international and national levels, have the potential to bring about welcome innovation to national survey practices. Countries participating in PIAAC had a strong incentive to reach the exacting international standards and, as such, these standards were often the gate-openers to adapting standard methods and procedures and adopting new survey operations. There were many instances of this in the fieldwork for PIAAC in Germany. Beyond the examples discussed in the previous section, many other details of fieldwork were adjusted or improved upon for the German implementation of PIAAC. Some of these may enrich future national surveys (e.g., address search).

Standardization of survey operations aims at achieving comparability. Even though the need for standardization in the data collection of cross-national com-

parative surveys is uncontested, there are also limits to standardization. Occasionally, comparability of results is better achieved by deliberately doing some things differently (Harkness, 2008; Koch et al., 2009; Lynn, 2003). Some of the PIAAC data collection standards and guidelines already explicitly allowed for different approaches, depending on countries' typical survey procedures. Furthermore, in the process of international quality control, certain country deviations were regarded as acceptable alternatives (Montalvan & Lemay, 2013a). Some of the other data collection standards, however, made no such allowances for national variability. The experience with the German PIAAC fieldwork, most aptly illustrated by the example of interview validation, points to the need for further reflection on how best to reach cross-national comparability in survey operations. It is thus with reservation that we note the recommendation proposed by Montalvan and Lemay (2013a) for future cycles of PIAAC calling for unconditional adherence to all validation standards and guidelines, specifically, the random selection of all finalized cases at a 10% level.

The "best" survey operations will differ, depending on the specific countries (and even on the specific survey organizations) involved. The challenge in defining an appropriate set of international standards is to strike the right balance between standardization and national adaptations (see Koch et al., 2009). Because cross-national differences exist, it may not always be possible to define single standards that are realistically achievable in all countries. Furthermore, the implementation of international survey operation standards will have different repercussions in different countries, including costs and timelines (Lynn, 2003). Thus, it is advisable to involve countries in the process of setting standards to make it, at least partly, a collaborative effort, with national conditions shaping the international survey standards and determining their relative importance. Lastly, in order to achieve full transparency in the program of quality assurance, it is crucial that not only the survey standards be known at the onset of the survey, but also the relevant framework and indicators for the data adjudication.

References

Ackermann-Piek, D., & Massing, N. (2014). Interviewer behavior and interviewer characteristics in PIAAC Germany. *methods, data, analyses, 8*(2), 199-222. doi: 10.12758/mda.2014.008

Alcser, K. H., & Benson, G. (2005). The SHARE train-the-trainer program. In A. Börsch-Supan & H. Jürges (Eds.), *The Survey of Health, Ageing, and Retirement in Europe – Methodology* (pp. 70-74). Mannheim: MEA. Retrieved from http://www.share-project.org/uploads/tx_sharepublications/Methodology_Ch6.pdf

Alcser, K. H., & Clemens, J. (2011). X. Interviewer recruitment, selection, and training (Revised 28. Nov. 2011). In *Cross cultural survey guidelines*. Retrieved from http://ccsg.isr.umich.edu/pdf/10InterviewerRecruitmentFeb2012.pdf

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly, 52*(2), 190-211.

Blohm, M., & Koch, A. (2013). Respondent incentives in a national face-to-face survey: Effects on outcome rates, sample composition and fieldwork efforts. *Methoden, Daten, Analysen (MDA), 7*(1), 89-122. doi: 10.12758/mda.2013.004

Börsch-Supan, A., Krieger, U., & Schröder, M. (2013). *Respondent incentives, interviewer training and survey participation*. SHARE Working Paper (12-2013). Munich Centre for the Economics of Ageing (MEA). Retrieved from http://www.share-project.org/publications/workingpapers0.html

Cannell, C. F., Marquis, K. H., & Laurent, A. (1977). *A summary of studies of interviewing methodology*. Washington, D.C.: US Government Printing Office.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresonse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York, NY: John Wiley & Sons.

European Social Survey. (2011). *Round 6 specification for participating countries*. London: Centre for Comparative Social Surveys, City University London.

European Social Survey. (2013). *Round 7 specification for participating countries*. London: Centre for Comparative Social Surveys, City University London.

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, California: Sage Publications.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M., & McGonagle, K. A. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics, 17*(2), 249-265.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly, 64*, 299-308.

Harkness, J. A. (2008). Comparative survey research: Goal and challenges. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56-77). New York, NY: Psychology Press.

Juran, J. M., & Gryna, F. M. (1970). *Quality planning and analysis*. New York, NY: McGraw-Hill Inc.

Koch, A., Blom, A. G., Stoop, I., & Kappelhof, J. (2009). Data collection quality assurance in cross-national surveys: The example of the ESS. *Methoden, Daten, Analysen (MDA), 3*(2), 219-247.

Lessler, J. T., Eyerman, J., & Wang, K. (2008). Interviewer training. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 442-460). New York, NY: Psychology Press.

Lyberg, L. E., & Biemer, P. P. (2008). Quality assurance and quality control in surveys. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 421-441). New York, NY: Psychology Press.

Lynn, P. (2003). Developing quality standards for cross-national survey research: Five approaches. *International Journal of Social Research Methodology, 6*(4), 323-336.

Martin, S., Helmschrott, S., & Rammstedt, B. (2014). The use of respondent incentives in PIAAC: The field test experiment in Germany. *methods, data, analyses,* 8(2), 223-242. doi: 10.12758/mda.2014.009

Massing, N., Ackermann, D., Martin, S., Zabal, A., & Rammstedt, B. (2013). Controlling interviewers' work in PIAAC – the Programme for the International Assessment of Adult Competencies. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys. Impact, reasons, detection and prevention* (pp. 117-130). Frankfurt am Main: Peter Lang.

Mohadjer, L., Krenzke, T., & Van de Kerckhove, W. (2013). Indicators of the quality of the sample data. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)* (Chapter 16, pp. 1-30). Paris: OECD. Retrieved from http://www.oecd.org/site/piaac/publications.htm

Montalvan, P., & Lemay, M. (2013a). Field operations. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)* (Chapter 10, pp. 1-38). Paris: OECD. Retrieved from http://www.oecd.org/site/piaac/publications.htm

Montalvan, P., & Lemay, M. (2013b). Quality control monitoring activities. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)* (Chapter 11, pp. 1-12). Paris: OECD. Retrieved from http://www.oecd.org/site/piaac/publications.htm

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Surve*y (NCES 98-053). Washington D.C.: U.S. Department of Education, National Center for Education Statistics.

O'Brien, E. M., Mayer, T. S., Groves, R. M., & O'Neill, G. E. (2002). Interviewer training to increase survey participation. *Proceedings of the Survey Research Methods Section* (pp. 2502-2507). Alexandria: American Statistical Association. Retrieved from http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-000530.pdf

OECD. (2010). *PIAAC technical standards and guidelines* (December 2010). Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm

OECD. (2013a). *OECD Skills outlook 2013: First results from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. (2013b). *The Survey of Adult Skills – Reader's companion*. Paris: OECD Publishing.

OECD. (2013c). *Technical report of the Survey of Adult Skills (PIAAC).* Paris: OECD. Retrieved from http://www.oecd.org/site/piaac/publications.htm

OECD & Statistics Canada. (2000). *Literacy in the information age. Final report of the International Adult Literacy Survey.* Paris: OECD Publishing.

OECD & Statistics Canada. (2011). *Literacy for life: Further results from the Adult Literacy and Life Skills Survey.* Paris: OECD Publishing.

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., . . . Rammstedt, B. (forthcoming). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly.*

Rammstedt, B. (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich – Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Rosen, J., Murphy, J., Peytchev, A., Riley, S., & Lindblad, M. (2011). The effects of differential interviewer incentives on a field data collection effort. *Field Methods, 23*(1), 24-36.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 163-177). New York, NY: John Wiley & Sons.

Statistics Canada & OECD. (2005). *Learning a living. First results of the Adult Literacy and Life Skills Survey*. Paris: OECD Publishing.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester: John Wiley & Sons.

Wasmer, M., Scholz, E., Blohm, M., Walter, J., & Jutz, R. (2012). *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2010* (GESIS-Technical Reports 2012/12). Köln: GESIS – Leibniz-Institut für Sozialwissenschaften.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann.

methods, data, analyses | Vol. 8(2), 2014, pp. 199-222 DOI: 10.12758/mda.2014.008

Interviewer Behavior and Interviewer Characteristics in PIAAC Germany

Daniela Ackermann-Piek[1,2] *& Natascha Massing*[2]

[1] *University of Mannheim*
[2] *GESIS – Leibniz Institute for the Social Sciences*

Abstract

Interviewers are the first in line when it comes to data collection. Therefore, it is important that they perform their tasks diligently, so that the data they collect are comparable and that errors are minimized. This paper analyzes how interviewers conducted interviews for the *Programme for the International Assessment of Adult Competencies* (PIAAC) and which kinds of mistakes they made. We approached these questions with audio interview recordings collected during the fieldwork of PIAAC in Germany (carried out in 2011/2012), as well as with an interviewer survey conducted with the German PIAAC interviewers. First, we introduce the data and the coding scheme used to evaluate interviewers' behavior with audio recordings. Subsequently, we describe the interviewers' actual behavior with regard to standardized interviewing techniques and investigate whether interviewer characteristics are associated with data quality. Our results demonstrate that interviewers do deviate from the expected behavior in all the aspects we examined. However, we identified only few associations with interviewers' background characteristics.

Keywords: PIAAC Germany, interviewer behavior, data quality, audio recordings, interviewer survey, interviewer characteristics

1 Introduction

Face-to-face surveys rely on interviewers for data collection. However, behavior regarding standardized interviewing techniques may differ across interviewers. As a result, interviewers can influence – intentionally or unintentionally – various aspects of the data collection process. Concerns about interviewer effects in interviewer-mediated surveys have accompanied generations of survey researchers. According to Groves et al. (1992), Loosveldt (2008), Schaeffer, Dykema, and Maynard (2010), and Blom and Korbmacher (2013), interviewers have many different roles in administering a survey: They contact sample persons and persuade them to participate, they clarify the goal of the survey and explain to respondents what is expected of them, as well as ask questions and record answers. Thus, the behavior of interviewers affects nearly all aspects of survey errors, including sampling (Eckman, 2013; Eckman & Kreuter, 2011; Tourangeau, Kreuter, & Eckman, 2012), nonresponse (e.g., Blom, de Leeuw, & Hox, 2011; Durrant, D'Addio & Steele, 2013; Jäckle, Lynn, Sinibaldi, & Tipping, 2013), measurement (Durrant, Groves, Staetsky, & Steele, 2010; Rice, 1929), and coding or editing of survey responses (e.g., Campanelli, Thompson, Moon, & Staples, 1997). The focus of the present paper is on the measurement perspective of interviewer behavior: interviewers' behavior with regard to deviations from standardized interviewing techniques during interviews.[1]

In terms of the total survey error framework, as many error sources as possible should be taken into account when designing a survey (for a survey see Groves & Lyberg, 2010). When it comes to errors during face-to-face interviews, standardized interviewing techniques are commonly used as a strategy to reduce errors introduced by interviewers (e.g., Fowler & Mangione, 1990; Mangione, Fowler, & Louis, 1992). In a standardized interview, interviewers are expected a) to read aloud questions, as well as instructions, as they are scripted, b) provide adequate

1 For more information regarding nonresponse in PIAAC Germany, see Helmschrott and Martin (in this volume).

Direct correspondence to
Daniela Ackermann-Piek, University of Mannheim, 68131 Mannheim, Germany
E-mail: ackermann-piek@uni-mannheim.de

Acknowledgment: We thank GESIS – Leibniz-Institute for the Social Sciences for its support. Furthermore, we thank the funders of PIAAC in Germany: The Federal Ministry of Education and Research, with the participation of the Federal Ministry of Labor and Social Affairs. We also thank Prof. Annelies Blom and the reviewers for many valuable comments and suggestions as well as Jana Kaiser for proof reading. Finally, we thank the German survey agency TNS Infratest for its cooperation in administrating the interviewer survey of PIAAC and all the interviewers who participated in the PIAAC interviewer survey.

nondirective probing, if necessary, and c) be unbiased towards respondents and record answers accurately (Fowler & Mangione, 1990, p. 14). All steps should be conducted in exactly the same way by each interviewer and therefore no differences between them should occur. Accordingly, all respondents are provided with identical stimuli and the "general assumption is that when all interviewers do their job in a standardized way and adhere to the interview rules, and when they interview a comparable group of respondents, they will get comparable answers." (Loosveldt, 2008, p. 216).

However, several studies have shown that interviewers deviate from standardized techniques. Hyman and Cobb (1954) were among the first to present results of errors introduced by interviewers who did not follow standardized interviewing techniques. Several other studies followed and revealed, for example, effects introduced by autonomously reworded text (e.g. Billiet & Loosveldt, 1988; Brenner, 1982; Haan, Ongena, & Huiskes, 2013; Ongena, 2005). Maynard and Schaeffer (2002) summarized the debate on standardization and concluded that understanding why interviewers deviate from the expected behavior helps to improve data quality.

Two approaches are commonly used to explain why interviewers deviate from standardized interviewing techniques. The first approach focuses on the survey instrument and the second on the interaction in the question-answer process. With respect to the survey instrument, many guidelines have been written on how survey questions should be scripted (e.g. Porst, 2008). Firstly, formulating survey questions of good quality reduces the bias introduced by interviewers, because they do not feel the need to deviate from the question text (Schaeffer, 1991; Schaeffer et al., 2010; Schaeffer & Maynard, 1996). Secondly, Schober and Conrad (2002) concluded that, due to the nature of communication, interviewers collaborate with respondents when trying to improve question understanding, which might affect responses. Additionally, interviewers might not want to appear ignorant or impolite and therefore tailor the question text (Ongena & Dijkstra, 2006). Further studies suggest that conversationally structured interviews reduce interviewers› burden and therefore minimize the chance of mistakes, because there are no rules for standardization (e.g. Cannell, Miller, & Oksenberg, 1981; Houtkoop-Steenstra, 2000; Schober & Conrad, 1997). Although these authors state that a flexible interviewing technique has many advantages – especially for interviewers – they admit that it is very time consuming and more challenging when controlling interviewers' work.

However, these two approaches do not fully explore the reasons for interviewers' deviations from standardized techniques. The literature suggests a third approach: using interviewer characteristics, such as attitudes or behavior, as predictors for nonresponse and measurement error (Blom & Korbmacher, 2013; Durrant et al., 2010). However, research into the effects of interviewers' background characteristics, such as gender, age or education, has yielded inconsistent findings

(for an overview see Schaeffer et al., 2010). Groves (2004) concluded that interviewers› characteristics are mostly associated with measured constructs when both are related (e.g., questions on respondents' weight might be affected by interviewers' gender). For example, interviewers' experience is often used to explain differences in the success of reaching contact or gaining cooperation.[2] Gfroerer, Eyerman, and Chromy (2002) related interviewers' experience to standardized interviewing techniques and found that less experienced interviewers tend to be more accurate in reading questions. Furthermore, Groves et al. (2009) and Groves and Lyberg (2010) reported that interviewers with more experience introduce greater measurement error to the data. However, other studies did not find an effect of experience and conclude that any effects might be overcome with training (e.g. Collins, 1980).

Nevertheless, detailed data on interviewers' actual behavior during the interview and interviewers' characteristics are often not available. Because these data are available for the Programme for the International Assessment of Adult Competencies (PIAAC) Germany, we used the third approach. The combination of detailed background information about interviewer characteristics with actual interview behavior is special and enables us to fill a gap in the literature and explain deviations of interviewers› behavior from standardized interviewing techniques. We first describe the behavior of the interviewers in the standardized structured background questionnaire of PIAAC Germany. Subsequently, we present findings from analyses of the association between interviewer behavior during the PIAAC interview and interviewer characteristics.

2 Data Description

In comparison to many other studies that use auxiliary data to evaluate interviewers' behavior, we could rely on factual data from the German PIAAC survey. We used data about interviewers that were either on the interviewer level or on the respondent level. Data on interviewers' background characteristics came from an interviewer questionnaire that was implemented in order to collect more data on interviewers, their attitudes, and reported behavior. Data on interviewers' actual behavior regarding standardized interviewing techniques were derived from audio recordings of interviews collected during the fieldwork in PIAAC Germany. In the following section, we first briefly explain the interviewers› role in PIAAC Germany[3] and then describe both data sources in more detail.

---

2 This relationship is usually linear (e.g. Jäckle et al., 2013) or, rarely, U-shaped (Singer, Frankel, & Glassman, 1983)

3 The description of PIAAC is based on our own experience during the implementation of PIAAC in Germany, as well as on the international technical report (OECD, 2013) and the German PIAAC technical report (Zabal et al., 2014).

2.1 PIAAC Germany and the Role of Interviewers

PIAAC is an international survey, initiated by the OECD (OECD, 2014) and implemented by an international Consortium. Its aim is to investigate how adults› competencies are distributed across and within countries. All participating countries collected data via face-to-face interviews with randomly sampled persons. In Germany – like in almost all other participating countries – the data collection took about eight months, between August 2011 and March 2012.[4] In total, 129 interviewers from the German survey organization TNS Infratest worked for PIAAC in Germany. The cases were organized in sample points based on a random sample of the adult population in Germany (16-65 years of age). Most interviewers worked in two or three sample points with 32 addresses per point. However, due to organizational arrangements, a few interviewers worked in only one or in up to five sample points. In total, the target size of approximately 5,000 respondents was exceeded, with a final number of 5,465 completed interviews.[5]

In PIAAC, the role of the interviewers differed somewhat from their normal tasks. The design of PIAAC included not only a computer-based background questionnaire, which interviewers are accustomed to administer, but also a computer-based assessment of every-day skills in the domains *literacy*, *numeracy* and *problem solving in technology-rich environments*. The background questionnaire was administered as a computer-assisted personal interview and contained questions about the respondent, such as education or the use of skills at work and in every-day life. The assessment was in a self-completion format administered under the supervision of the interviewer. Although we did not use the data collected in the skills assessment for the analysis in this paper, it is important to note that the interviewers had to adapt their behavior for the assessment, because they had to learn to be more passive in their role as test administrators.

To ensure that the PIAAC data were of high quality, specific and comprehensive technical standards and guidelines were defined by the international Consortium (OECD 2010) and each participating country had to comply with these standards when carrying out PIAAC. The implementation of the standards was monitored very closely by the Consortium and every single deviation from the standards had to be approved. One important aspect of the international requirements referred to quality control of the fieldwork: interviewers' work, as well as the data quality, had to be closely monitored.[6] The analyses in this paper that deal with deviations from standardized interviewing techniques were based on the information retrieved from audio recordings of interviews from the PIAAC background

4 This included two main fieldwork phases as well as several re-issue phases.

5 For a definition of a completed case in PIAAC, see OECD (2010).

6 All standards and guidelines related to interviewers are described in detail in Massing, Ackermann, Martin, Zabal, and Rammstedt (2013).

questionnaire that was collected and reviewed in this context. The international requirements for quality control stipulated that each interviewer had to produce two audio recordings (for more details, see below).

Another important aspect in the PIAAC standards and guidelines was that interviewers received intensive in-person trainings, to provide them with adequate information and practice for carrying out their various tasks. The training included a special focus on standardization for the data collection in the background questionnaire. Conducting such extensive interviewer trainings is relatively uncommon in Germany. In other countries, however, this is best practice and several studies have demonstrated a positive effect of interviewer trainings on response rates and on the overall data quality (e.g. Billiet & Loosveldt, 1988; Couper & Groves, 1992; Fowler & Mangione, 1990; Japec, 2008). Furthermore, German PIAAC interviewers were carefully selected.[7]

In addition to their training, interviewers were provided with substantial information material. For instance, they received an extensive manual that included detailed descriptions of each relevant aspect of PIAAC in Germany, as well as a small interviewer booklet. Providing interviewers with such extensive material is also uncommon in German surveys.

2.2 Interviewer Questionnaire

To date, interviewer behavior, or even interviewer effects, has often only been described but not explained, because data to explain those effects are lacking (Blom & Korbmacher, 2013; Brunton-Smith, Sturgis, & Williams, 2012). In Germany, detailed data on interviewer characteristics are normally not provided by survey agencies. To overcome this gap, additional data on the PIAAC interviewers were collected by the authors, using a questionnaire that was adapted from the questionnaire implemented in the Survey of Health, Ageing and Retirement in Europe (SHARE) 2011 (Blom & Korbmacher, 2013). Interviewers' participation was voluntary and the interviewers did not receive any kind of incentive. Data from the interviewer survey were not intended to be used for quality control measures during PIAAC but rather to gain more information about the interviewers, in order to analyze differences in interviewers' behavior and success, related to their characteristics. It contained questions about the interviewers' background, their attitudes, and their expectations, related to their fieldwork in PIAAC.[8] The questionnaire was sent to 128 interviewers and 115 interviewers completed and returned the questionnaire, resulting in a response rate of almost 90%. However, 15 questionnaires were received without an interviewer ID (see Table 1). These cases could not be matched

7 The selection criteria are described in detail in Zabal et al. (2014).

8 The source questionnaire is presented in Blom and Korbmacher (2013).

Table 1 Overview of the Interviewer Questionnaire

	n	Percent
Interviewer received questionnaire	128	100.0
Interviewer returned questionnaire	115	89.8
Questionnaire contained interviewer ID	100	78.1

Note. One interviewer was excluded after a short time. Therefore, the questionnaire was sent to 128 interviewers.

with interviewer behavior retrieved from the audio data. Therefore, they were excluded for joint analysis of interviewer characteristics and interviewer behavior. Their exclusion did not alter the results.

A summary of the interviewers' background characteristics, collected through the interviewer survey, is provided in Table 2. The results for gender and age were equivalent to the information provided by the survey agency TNS Infratest in their technical report (Zabal et al., 2014, p. 54). TNS Infratest provided additional information on how long interviewers had been working for their survey institute: 71% of the interviewers had worked for TNS Infratest for ten years or less. However, our results show that over 45% stated that they had worked as interviewers for more than ten years. Another interesting issue is related to the experience of PIAAC interviewers: compared to interviewers from other German surveys, PIAAC interviewers were very experienced (Blom, Ackermann, Korbmacher, Krieger, & Massing, 2013). This is not surprising, because one criterion for selection as a PIAAC interviewer required candidates to be a senior interviewer.

Table 2 Characteristics of the German PIAAC interviewers

		n	Percent
Gender	Male	62	53.91
	Female	53	46.09
	Total	115	100.00
Age	<= 45 years	10	8.70
	46 – 55 years	21	18.26
	56 – 65 years	51	44.35
	>= 66 years	33	28.70
	Total	115	100.00
Work experience	< 2 years	10	8.77
	2 – 5 years	31	27.19
	6 – 10 years	21	18.42
	11 – 15 years	10	8.77
	> 15 years	42	36.84
	Total	114	100.00
Education	Lower-level or medium-level school and no vocational or university qualification	1	0.93
	Medium-level school qualification and vocational education	36	32.73
	Advanced technical college entrance qualification or university entrance qualification	42	38.18
	Tertiary education	31	28.18
	Total	110	100.00
Working hours per week	<= 10 hours	6	5.66
	11 – 20 hours	31	29.25
	21 – 30 hours	36	33.96
	31 – 40 hours	18	16.98
	> 40 hours	15	14.15
	Total	106	100.00

Notes. Data from the PIAAC interviewer survey. 115 interviewers included in analysis. Number of cases varies because of item nonresponse.

2.3 Audio Recordings and Coding Scheme

As mentioned above, the PIAAC standards stated that each country had to evaluate at least two audio recordings, per interviewer, of interviews made during administration of the background questionnaire (OECD 2010). Analyzing recordings is considered to be a good way of monitoring interviewers' behavior and interviewing techniques, without affecting respondents' behavior (Fowler & Mangione, 1990; Sykes & Collins, 1992). In addition, such recordings provide insights into the complex interaction process between interviewers and respondents (Ongena, 2005). The audio recordings were taken early in the field period. The interview was recorded via an external digital voice recorder and the interviewer had to manually start and stop the recording. Table 3 shows an overview of the expected as well as the recordings actually delivered by the interviewers. In total, 258 recordings were expected. Recordings were not available for some interviewers, whilst others delivered more than two recordings. In total, 245 recordings were received, coded, and reviewed during quality control of the fieldwork in PIAAC Germany.

To use the information from the audio recordings for quality control, information first had to be coded. In the literature, several coding schemes are available, indicating that the choice of coding scheme depends on the purpose of the analysis (for an overview see Ongena & Dijkstra, 2006).

The main reason for evaluating interviewer behavior using audio recordings in PIAAC was quality control. The aim was to obtain information about the interviewers' interviewing techniques and their actual behavior during the interview as early as possible during the data collection in order to intervene, if necessary. Because coding and reviewing audio recordings is very time consuming[9] and information was needed as early as possible, we developed a simple coding scheme that focused on crucial deviant interviewer behavior in the background questionnaire.[10] A major problem was defined as a deviation from the standardized script that potentially affects the level of accuracy of the response (Ongena, 2005).

To avoid coder effects, coding was conducted by six different coders. It was ensured that two persons coded the recordings of one interviewer. Any inconsistencies or difficulties in the codes were resolved by two lead coders. After a review of the coding, a summary of the behavior of each individual interviewer was written by the lead coders and feedback was provided to the survey agency. All codes were derived directly from the audio recordings.

---

9 Coding the background questionnaire took about one hour per recording and was conducted directly from the recordings, using the software Audacity (Mazzoni & Dannenberg, 2012).

10 The PIAAC technical standards and guidelines only required this part of the interview to be reviewed via recordings.

Table 3 Overview of the audio recorded interviews

	n	Percent
Interviewer	129	100.0
Interviewer with no recordings	8	6.6
Audio tapes to be recorded	258	100.0
Received audio taped interviews	245	95.0
Interviewer with 1 recording	1	0.8
Interviewer with 2 recordings	116	95.9
Interviewer with 3 recordings	4	3.3

Note. Reference: Zabal et al. (2014).

For the present analysis, we reorganized the original coding scheme used for quality control in PIAAC, based on the coding scheme of Beullens, Loosveldt, and Dries (2013). Each single code represents one aspect of standardization. The resulting seven codes were grouped into three categories: administration, completeness, and probing (see Figure 1).

The first category contained administrative information that interviewers were asked to record at the beginning of the interview. The first code *admin I* consisted of a combination of the following information: the date of the interview, the interviewer ID and the respondent ID. Only if the interviewer ID or the respondent ID was recorded incorrectly (missing or incomplete) was this coded as incorrect interviewer behavior. *Admin II* covered whether interviewers announced the recoding to the respondent and whether they explicitly asked for permission to record the interview. This was especially crucial because data protection regulations are strict in Germany. Only if the announcement of the recording was completely absent on the recording was this coded as incorrect interviewer behavior. However, because a digital voice recorder, and not the laptop, was used to record the interview, it was obvious for all recordings that all respondents were aware that the interview was being recorded. This was further confirmed by the audio recordings, which contained no indication of any secret recording of interviews. Nevertheless, because this was a formal requirement, this code provided information on how accurately interviewers worked.

For the second category, *completeness*, the two codes referred to question text.[11] We will explain these codes by using the example of a question wording,

11 During quality control, two additional codes were used, referring to answer categories and showcards. However, coding could not be derived from the audio recordings for all cases and we thus excluded these codes from our analysis.

	Administration	Completeness	Probing
Admin I: collected date of interview, interviewer ID, respondent ID	x		
Admin II: collected permission to record interview from respondent	x		
Question is read out (not incorrect skipped)		x	
Question is completely read out		x	
Probing (if applicable)			x
Probing overall correct			x
3-point scale for probing quality			x

Notes. ID = Identification Number. Admin = Administration.

Figure 1 Coding scheme for audio recordings of the background questionnaire of PIAAC in Germany

provided in Figure 2, to illustrate deviations from standardized interviewing techniques.

We coded each single incidence of an incorrectly skipped question as incorrect interviewer behavior. With respect to the question wording provided in Figure 2, we found that interviewers often deviated from the script, using information from the previous part of the interview. For example, in one interview, the interviewer assumed that the respondent was a student instead of part-time employed, because both talked about forthcoming holidays. Because the question was not asked, the interviewer collected incorrect information. As a consequence, various filters of the following questionnaire did not fit the respondent's situation and data were incorrect. Although incorrectly skipped questions do not necessarily result in incorrect data, this example shows that each piece of information obtained from the previous conversation has to be verified by asking each single question (Ongena 2005). Luckily, in our example, the respondent realized the error introduced by the interviewer and asked to go back, to change the information that applied to her situation.

With respect to the second aspect of completeness, we assume that rewording or shortening a question has either no, a minor, or a major impact on the respondents' answers, and use the example provided in Figure 2 to explain the differences. For the wording presented in Figure 2, the interviewer might simply leave out the

Question

Please look at this card and tell me which ONE of the statements best describes your current situation. If more than one statement applies to you, please indicate the statement that best describes how you see yourself.

Instruction

1. Hand show card 9.
2. Mark only one answer.

Answer Categories

01 Full-time employed (self-employed, employee)

02 Part-time employed (self-employed, employee)

03 Unemployed

04 Pupil, student

05 Apprentice, internship

06 In retirement or early retirement

07 Permanently disabled

08 In compulsory military or community service

09 Fulfilling domestic tasks or looking after children/family

10 Other

DK

RF

Notes. DK = don't know. RF = refused. DK and RF were not printed on showcards in general.

Figure 2 Example of a question from the PIAAC background questionnaire

first word "Please". We assume that this has no effect on question understanding. However this rewording could also have a minor effect, if respondents think that the question is not worded very politely or that the interviewer is impolite. We assume that minor rewordings have no major effect on the accuracy of responses. On the other hand, we assume that complete rewordings of the original question text (e.g., changing the question wording presented in Figure 2 to: "Are you employed?") will have major effects on the accuracy of responses, if further information is not provided by the interviewer about how respondents should answer the question and, thus, respondents do not have the opportunity to assign themselves to the correct answer category. In comparison, a minor effect of this completely reworded question could be that respondents ask for clarification and interviewers probe to provide respondents with the missing information. As mentioned above, we decided to focus on major problems and did not code minor rewordings as incor-

rect interviewer behavior during quality control. We only coded major deviations from the original question text which, we assumed, would have major effects on the responses, as incorrect interviewer behavior.

Finally, three codes referred to probing (Figure 1), an interviewing technique in which additional information is provided on request. This is usually triggered by respondents, when, for example, they ask for clarification of the question or give an inaccurate answer (e.g., one that does not fit the answer scheme). Each time interviewers had to probe, the quality of the probing was coded. The first code included information on whether probing was necessary or not. We subsequently constructed a dichotomy code that included information about whether probing was correct or not. Because there is a wide range of probing quality, we decided to additionally build a three-point scale to differentiate between a) excellent probing, b) probing that was not good, but for which it was assumed that it would not have a major negative effect on the respondent's answer and, c) poor probing. The scale was constructed by combining the number of good and poor probes, based on the overall distribution: More than three correct probes were considered to be excellent probing on the scale; if an interviewer conducted only bad probing, without any good probing, we considered this to be poor probing, and everything in between was assigned to the middle category. A good probe is nondirective and neutral, which means that it does not influence the content of the answer. In contrast, a poor probe influences the answer of the respondent (Fowler & Mangione, 1990). Due to limited details in the original coding schema, this scale could be applied to approximately only half of the recordings.

3 Results

In this section, we present results of the descriptive analysis of the interviewer behavior retrieved from the audio recordings. We start by describing how many interviews we identified in which interviewers collected administrative information incorrectly and then proceed to provide information on interviewers' behavior using standardized interviewing techniques such as reading questions without incorrect skipping or rewording. Finally, we provide information on interviewers' probing behavior. In the second part of this section, we show whether interviewers' behavior in the interviews is associated with interviewers' background characteristics. For this purpose, we crossed the information from the audio recordings with interviewers' characteristics from the PIAAC interviewer survey and calculated several regressions. All results in the following section are based on those cases for which the interviewer ID was available from the interviewer questionnaire. Nevertheless, results including all cases do not differ substantively.

Table 4 Interviewer behavior for collecting administrative information

Admin I **collected date of interview, interviewer ID, respondent ID**		
	n	Percent
Incorrect	94	43.32
Correct	123	56.68
Total	217	100.00
Admin II **collected permission to record interview from respondent**		
	n	Percent
Incorrect	52	23.96
Correct	165	76.04
Total	217	100.00

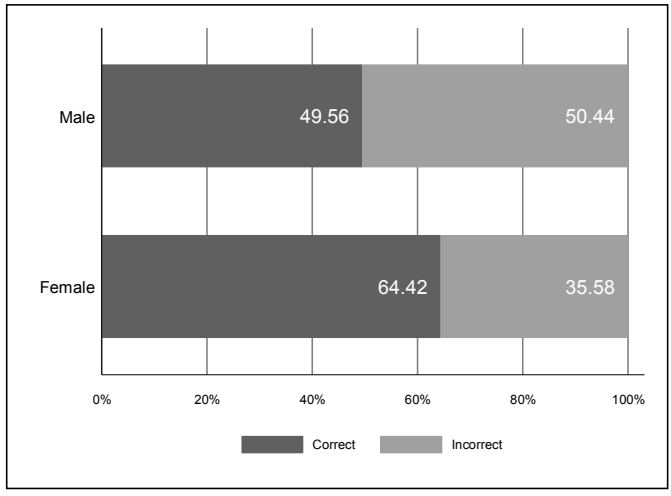
Note. Data based on 107 interviewers and 217 recordings.

3.1 Administration

The interviewers were asked to record some administrative information, such as the date of the interview or the interviewer ID. The results presented in Table 4 show that, in 43% of the recordings, either the date of the interview, the respondent ID or the interviewer ID were missing on the recording (admin I). Furthermore, it was a formal requirement for interviewers to record the permission of the respondent for recording the interview (admin II). In almost 25% of the cases, the recording was not announced in the standardized way; i.e., according to the instructions the interviewers had received. As already mentioned, we did not find any case in which recordings were not announced at all to respondents.

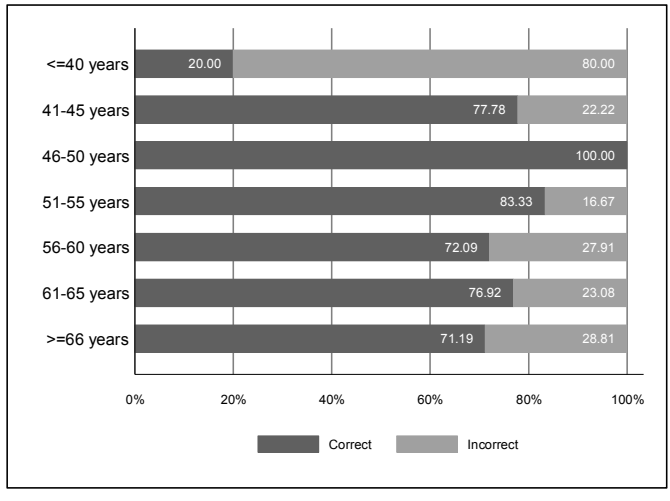
Crossing admin I with interviewers' characteristics revealed that there were significantly fewer mistakes in recording the date of the interview, the interviewer ID, as well as the respondent ID in interviews conducted by female interviewers, compared to interviews conducted by their male colleagues (Figure 3). In terms of age, working experience, education, and working hours, a clear pattern was not evident. Results of a logistic regression that included all five interviewer characteristics in one model did only reveal a positive significant association with gender (Odds Ratio = 0.1853, $p = 0.048$).

For collecting permission to record the interview (admin II), our analyses yielded a significant association with age and working hours per week: For age, no clear pattern was found (Figure 4). However, we found significantly more mistakes



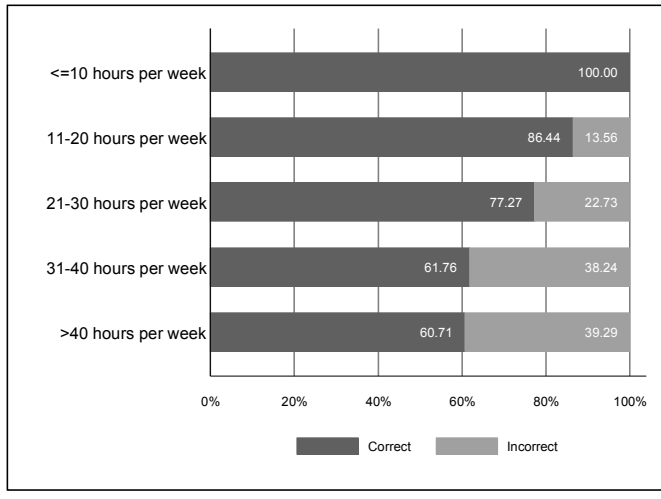
Notes. $\chi^2 = 4.8742$. $p = 0.027$. Data based on 107 interviewers and 217 recordings.

Figure 3 Interviewer behavior for collecting administrative information I and interviewer's gender



Notes. $\chi^2 = 17.2574$. $p = 0.008$. Data based on 107 interviewers and 217 recordings.

Figure 4 Interviewer behavior for collecting administrative information II and interviewer's age



Notes. $\chi^2 = 14.8856$. $p = 0.005$. Data based on 98 interviewers and 199 recordings.

Figure 5 Interviewer behavior for collecting administrative information II and interviewer's working hours per week

in interviews conducted by interviewers with longer working hours per week (Figure 5). For interviewers' gender, experience, and education, a significant association was not evident. Results of a logistic regression including all five interviewer characteristics in one model supported these results: a significant negative association was present only for working hours per week (Odds Ratio = 0.5882, $p = 0.001$).

3.2 Completeness

We investigated several aspects of completeness, including the correct use of filters (questions not incorrectly skipped) and the accuracy of reading a question as scripted. Starting with the number of incorrectly skipped questions, our results showed that, in 55% of the recordings, every question was read out (Table 5). In 27% of the cases, up to two questions were incorrectly skipped and, in 10%, five or more questions were incorrectly skipped. No significant differences were identified for any of the tested interviewer characteristics, neither through cross tabulation nor with a linear regression.[12]

With regard to reading questions as they are scripted (e.g. shortening or rewording), our results showed that, in 58% of all recordings, up to ten questions were read incorrectly. Additionally, more than ten questions were not read correctly in 26% of the recordings (see Table 6). Examples of how interviewers reworded

12 Results available from corresponding author upon request.

Table 5 Interviewer behavior regarding incorrect skipping of questions

Number of incorrect skipped questions	*n*	Percent
0	120	55.30
1	43	19.82
2	16	7.37
3	8	3.69
4	8	3.69
>= 5	22	10.14
Total	217	100.00

Notes. Data based on 107 interviewers and 217 recordings. On average, around 160 questions were asked per case.

Table 6 Interviewer behavior regarding incorrect reading of questions

Number of incorrect read questions	*n*	Percent	Cummul. percent
0	35	16.13	16.13
1	32	14.75	30.88
2	17	7.83	38.71
3	16	7.37	46.08
4	12	5.53	51.61
5	13	5.99	57.60
6	10	4.61	62.21
7	8	3.69	65.90
8	5	2.30	68.20
9	9	4.15	72.35
10	3	1.38	73.73
11 - 20	36	16.59	90.32
21 - 30	12	5.53	95.85
> 30	9	4.15	100.00
Total	217	100.00	100.00

Notes. Cummul. = cumulative. Data based on 107 interviewers and 217 recordings. On average, around 160 questions were asked per case.

questions are provided in section 2.2. No significant differences were identified for any of the tested five interviewer characteristics, using cross tabulation or a linear regression model.[13]

13 Results available from corresponding author upon request.

Table 7 Interviewer behavior regarding probing quality

	n	Percent
Excellent probing	35	29.41
Satisfying probing	62	52.10
Inaccurate probing	22	18.49
Total	119	100.00

Note. Data based on 84 interviewers and 119 recordings.

3.3 Probing

In almost all recorded interviews, respondents triggered interviewers to probe for at least one question (96%). In these cases, 29% of the interviewers performed excellently, probing was satisfactory in 52%, and probing was inadequate in almost 19% (Table 7). No significant association was found for any of the five tested interviewer characteristics.[14]

4 Discussion

Using data from PIAAC Germany, we provide detailed information on interviewers' behavior regarding several aspects of standardized interviewing techniques, such as using correct filters without skipping questions incorrectly, reading questions as scripted, and neutral communication. Furthermore, we investigated how interviewers' background characteristics were associated with deviations from the expected behavior with regard to these standardized interviewing techniques. During field work, some problems – such as incorrect reading of questions or incorrect probing[15] – were detected; analyses of interviewer behavior therefore seemed worthwhile. The overall results showed that the majority of the interviewers fulfilled the requirements and predominantly used standardized interviewing techniques. Some further analyses focused on the following aspects: Do the interviewers capture administrative information correctly? Do interviewers read each single question correctly (including answer categories)? Do interviewers probe accurately?

Capturing administrative information is one part of interviewers› daily work. Nonetheless, over 40% of interviewers did not correctly capture information, such

14 Results available from corresponding author upon request.

15 In total, 14 out of 129 interviewers were identified who had major problems with their interviewing technique and, consequently, received re-training.

as their own interviewer ID, on the recordings, and, in almost 25% of the cases, the interviewers did not announce the recording in the mandatory way. We consider the source of this error to be the way interviewer trainings are typically conducted. Usually, interviewer trainings in Germany have focused on providing study-specific information, such as how specific questions need to be administered. We assume that aspects of interviewers' daily work, especially accuracy of simple tasks, are covered in more general trainings that are often only conducted at the beginning of an interviewer's career. According to our analyses, there is a need to improve interviewers' understanding on how important it is to accurately capture administrative data, for example, for monitoring and controlling the fieldwork.

Another aspect of a standardized survey interview is that each single question is read completely as it is scripted. On average, around 160 questions were asked per case in the PIAAC background questionnaire. Results showed that, in almost half of the recorded interviews, interviewers incorrectly skipped at least one question and, in one fourth of the interviews, they even skipped more than two questions incorrectly. Additionally, in approximately one third of the recorded interviews, more than ten questions were not read out as scripted. Instead of reading out the question, interviewers, for example, used information from the previous part of the interview to answer the question by themselves. Yet, by not reading a question at all, interviewers "may overlook specific terms of questions or specific situations that the respondent did not report" (Ongena, 2005, p. 25). There is a real chance that the resulting data are incorrect and results drawn from this data contain errors. The same applies for reworded questions: While slightly rewording a question might have no, or even a positive effect, e.g., Haan et al. (2013), major deviations are more likely to affect the accuracy of responses (see also Ongena & Dijkstra, 2006; Smit, Dijkstra, & Van der Zouven, 1997). Differences across respondents may thus be artifacts of the effect interviewers had during the response process (Fowler & Mangione, 1990).

Furthermore, we examined the probing quality: for about one third of the interviews, the probes were excellent. However, we identified inaccurate probing in one fifth of our recordings (e.g., directive probing or providing incorrect information). According to Smit et al. (1997), suggestive probing has an impact on respondents› answers and can be considered to be a serious problem. Again, interviewers have to be made aware of the importance of correct probing and should be continuously trained and re-trained to make proper use of interviewer instructions and supportive material.

In most cases, we did not find significant differences in deviant behavior with regard to standardized interviewing techniques that were related to interviewers' characteristics (gender, age, education, experience, and working hours). With respect to education it is not surprising that significant differences are not found, because the level of education among the interviewers is relatively homoge-

neous. On the other hand, some associations were identified. For example, our data showed that, for interviews conducted by female interviewers, fewer mistakes were made in capturing administrative data such as interviewer or respondent ID. This might be mediated through other factors, because, for example, women tend to be more conscientious (Weisberg, DeYoung, & Hirsh, 2011). Training and monitoring activities could be adapted accordingly to intensify the attention on the way men perform their work.

Our results showed that, for interviews conducted by interviewers who reported having longer working hours per week, permission to record the interview was significantly less frequently collected. The interviewers' workload is likely to have an effect on the accuracy of interviewers' daily work. The amount of time interviewers can spend per case is lower when they have many cases to work on. Survey administration should ensure that interviewers' workload is manageable, as, for example, already stated early in the fifties by Collins (1980) and recently confirmed by Japec (2008), since this is one way of reducing interviewers' burden. However, it is not always possible to reduce interviewers' workload, for example, due to the availability of interviewers. Additionally, we are aware that some of the interviewers work for more than one survey agency, which we, unfortunately, cannot account for in this analysis.

Although interviewers were aware of the recordings, because they started the recording themselves manually, our results showed that interviewers did not always follow standardized interviewing techniques. In this study, some interviewers received feedback on their interviewing techniques after we had reviewed their audio recordings. Accordingly, they might have adapted their behavior. However, we have not checked their behavior again and we only provided feedback to those interviewers for whom we detected serious deviant interviewer behavior. According to Biemer (2010), interviewers tend to divert from standardized procedures in the same way over repeated interviews (e.g., they always read out a particular question incorrectly). In summary, we consider that recordings are a good way to gain information on interviewers' overall behavior, and we assume that our results can be generalized across interviews.

5 Conclusion and Outlook

In PIAAC Germany, extensive interviewers trainings were conducted, which is relatively uncommon in Germany (Zabal et al., 2014, p. 54f). An emphasis was placed on the importance of standardized interviewing techniques. However, even with this more intense training, it was not possible to completely avoid deviant interviewer behavior with regard to standardized interviewing. This suggests that, in many surveys, the problem of deviant behavior is underestimated. Of course, as interviewers are human beings, some degree of deviation from the standardized

script has to be expected. Nonetheless, deviations may affect data quality and thus results in quantitative studies conducted by interviewers.

Our analyses did not show many associations between interviewers' behavior, with regard to standardized interviewing techniques, and interviewers' background characteristics. Thus, the trainings might have been effective in reducing the variability between interviewers (see also Collins, 1980). This is consistent with our preliminary analyses with regard to interviewer effects on cooperation, using the same database. Here, we find that only 1.7% of the variability in cooperation rate can be attributed to interviewers (Blom et al., 2013; Massing & Ackermann, 2013). In comparison to similar surveys, which report interviewer effects of approximately 7% (Blom et al., 2013), this is particularly low. Another explanation for the lack of associations between interviewers' background characteristics and deviant interviewing might be that interviewer characteristics other than socio-economic ones are more important in this respect (for an overview see Schaeffer et al., 2010).

Deviations from standardized interviewing techniques result in inhomogeneous answers and hence may reduce the quality of the data or introduce measurement error, and should therefore be minimized. Several studies have already concluded that formulating good survey questions, intensive, tailored interviewer training and supervision as well as several monitoring strategies are a good way to minimize such effects. Based on a joint analysis of interviewers' success in gaining contact or cooperation and measurement, Brunton-Smith et al. (2012) suggest monitoring measures of interviewers' success, such as the contact or cooperation rate, which are indicators of key aspects of interviewer performance. This can lead to significant improvements in overall survey quality. We suggest, additionally, checking measures related to data quality by using recordings and giving feedback to interviewers on a regular basis during fieldwork. Simply training interviewers before they start to work might not be enough to keep them motivated and to ensure that they work consistently in the best possible way throughout the entire field period.

In this paper, our intention was not to explain interviewer effects but rather to demonstrate how interviewers deviated from expected behavior with regards to standardized interviewing techniques and to examine first associations between deviations and interviewers› background characteristics. Further analyses that make use of the rich data PIAAC Germany offers are necessary to explain the results. For example, other interviewer characteristics, such as interviewers' attitudes and expectations, respondents' characteristics or question characteristics can be used to explain deviation from standardized interviewing techniques. Based on analyses by Brunton-Smith et al. (2012), a combination of the relationship of interviewers' contact behavior and their workload is also worth analyzing. It would also be worthwhile to address the important issue of question quality, in order to reduce interviewers' burden.

References

Beullens, K., Loosveldt, G., & Dries, T. (2013). *Explaining interviewer effects on interviewer fieldwork performance indicators*. Paper presented at the 5th Annual Conference of the European Survey Research Association, Ljubljana.

Biemer, P. P. (2010). Overview of design issues: Total survey error. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2 ed., pp. 27-57). Bingley, UK: Emerald Group Publishing Limited.

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey question by interviewer training. *Public Opinion Quarterly, 52*, 190-211. doi: 10.1086/269094

Blom, A. G., Ackermann, D., Korbmacher, J. M., Krieger, U., & Massing, N. (2013). *Comparing interviewer characteristics and explaining interviewer effects on nonresponse across three German face-to-face surveys*. Paper presented at the Workshop: Explaining Interviewer Effects in Interviewer-Mediated Surveys, Mannheim.

Blom, A. G., de Leeuw, E. D., & Hox, J. J. (2011). Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics, 27*(2), 359-377.

Blom, A. G., & Korbmacher, J. M. (2013). Measuring interviewer characteristics pertinent to social surveys: A conceptual framework. *Survey Methods: Insights from the Field, 1*(1). doi: 10.13094/SMIF-2013-00001

Brenner, M. (1982). Response-effects of "role-restricted" characteristics of the interviewer. In W. Dijkstra & J. van der Zouwen (Eds.), *Response behavior in the survey-interview* (pp. 131-165). London, UK: Acad. Press.

Brunton-Smith, I., Sturgis, P., & Williams, J. (2012). Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance? *Public Opinion Quarterly, 76*(2), 265-286. doi: 10.1093/poq/nfr067

Campanelli, P., Thompson, K., Moon, N., & Staples, T. (1997). The quality of occupational coding in the United Kingdom. In L. Lyberg (Ed.), *Survey measurement and process quality* (pp. 437-456). New York, NY: John Wiley and Sons.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (Vol. 17, pp. 389-437). San Francisco, CA: Jossey-Bass Publishers.

Collins, M. (1980). Interviewer variability: A review of the problem. *Journal of the Market Research Society, 22*(2), 77-95.

Couper, M. P., & Groves, R. M. (1992). The role of the interviewer in survey participation. *Survey Methodology, 18*(2), 263-277.

Durrant, G. B., D'Addio, J., & Steele, F. (2013). Analysing interviewer call record data by using a multilevel discrete-time event history modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 251-269. doi: 10.1111/j.1467-985X.2012.01073.x

Durrant, G. B., Groves, R. M., Staetsky, L., & Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly, 74*(1), 1-36. doi: 10.1093/poq/nfp098

Eckman, S. (2013). Do different listers make the same housing unit frame? Variability in housing unit listing. *Journal of Official Statistics, 29*(2), 249-259. doi: 10.2478/jos-2013-0021

Eckman, S., & Kreuter, F. (2011). Confirmation bias in housing unit listings. *Public Opinion Quarterly, 75*(1), 139-150. doi: 10.1093/poq/nfq066

Fowler, F. J. J., & Mangione, T. W. (1990). *Standardized survey interviewing. Minimizing interviewer-related error* (Vol. 18). Newbury Park, CA: Sage Publications.

Gfroerer, J., Eyerman, J., & Chromy, J. (2002). *Redesigning an ongoing National Household Survey: Methodological issues*. DHHS Publication No. SMA 03-3768. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Groves, R. M. (2004). *Survey error and survey costs*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey interviewing*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error past, present and future. *Public Opinion Quarterly, 74*(5), 849-879. doi: 10.1093/poq/nfq065

Haan, M., Ongena, Y. P., & Huiskes, M. (2013). Interviewers' question: Rewording not always a bad thing. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention* (pp. 173-194). Frankfurt am Main: Peter Lang.

Helmschrott, S., & Martin, S. (2014). Nonresponse in PIAAC Germany. *methods, data, analyses, 8*(2), 243-266. doi: 10.12758/mda.2014.009

Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionnaire*. Cambridge: Cambridge University Press.

Hyman, H. H., & Cobb, W., J. (1954). *Interviewing in social research*. Chicago, IL: University of Chicago Press.

Jäckle, A., Lynn, P., Sinibaldi, J., & Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods, 7*(1), 1-15.

Japec, L. (2008). Interviewer error and interviewer burden. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 187-211). Hoboken, NJ: John Wiley & Sons.

Loosveldt, G. (2008). Face-to-face interviews. In E. D. De Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 201-220). New York, NY: Taylor & Francis Group.

Mangione, T. W., Fowler, F. J. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics, 8*(3), 293-307.

Massing, N., & Ackermann, D. (2013). *Interviewer characteristics & interviewer effects in PIAAC Germany*. Paper presented at the 5th conference of the European Survey Research Association, Ljubljana.

Massing, N., Ackermann, D., Martin, S., Zabal, A., & Rammstedt, B. (2013). Controlling interviewers' work in PIAAC - the Programme for the International Assessment of Adult Competencies. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention* (pp. 117-130). Frankfurt am Main: Peter Lang.

Maynard, D. W., & Schaeffer, N. C. (2002). Standardization and its discontents. In R. M. Groves, G. Kalton, J. N. K. Rao, N. Schwarz & C. Skinner (Eds.), *Standardization and tracit knowledge: Interaction and practice in the survey interview* (pp. 3-45). New York, NY: John Wiley & Sons.

Mazzoni, D., & Dannenberg, R. (2012). Audacity. Retrieved March 2014 from http://audacity.sourceforge.net/?lang=de.

OECD. (2010). PIAAC technical standards and guidelines (December 2010). Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm.

OECD. (2013). Technical report of the Survey of Adult Skills (PIAAC). Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm.

OECD. (2014). Retrieved November 2014 from http://www.oecd.org/site/piaac/.

Ongena, Y. P. (2005). *Interviewer and respondent interaction in survey interviews*. Amsterdam: Vrije Universiteit.

Ongena, Y. P., & Dijkstra, W. (2006). Question-answer sequences in survey-interviews. *Quality & Quantity, 40*, 983-1011. doi: 10.1007/s11135-005-5076-4

Porst, R. (2008). *Fragebogen: Ein Arbeitsbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rice, S. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology, 35*(3), 420-423.

Schaeffer, N. C. (1991). Conversation with a purpose - or conversation? Interaction in the standardized interview. In P. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 369-391). Hoboken, NJ: John Wiley & Sons.

Schaeffer, N. C., Dykema, J., & Maynard, D. W. (2010). Interviewers and interviewing. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 437-470). Bingley, UK: Emerald Group Publishing Limited.

Schaeffer, N. C., & Maynard, D. W. (1996). From paradigm to prototype and back again. In N. Schwarz & S. Sudman (Eds.), *Answering questions* (pp. 65-88). San Francisco, CA: Jossey-Bass Publishers.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*(4), 576-602. doi: 0033-362X/97/6104-0002$02.50

Schober, M. F., & Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer & J. van der Zouven (Eds.), *Standardization and tracit knowledge: interaction and practice in the survey interview* (pp. 67-94). New York, NY: John Wiley & Sons.

Singer, E., Frankel, M., & Glassman, M. B. (1983). The effect of interviewer characteristics and expactations on response. *Public Opinion Quarterly, 47*(1), 68-83.

Smit, J. H., Dijkstra, W., & Van der Zouven, J. (1997). Suggestive interviewer behaviour in surveys: an experimental study. *Journal of Official Statistics, 13*(1), 19-28.

Sykes, W., & Collins, M. (1992). Anatomy of the survey interview. *Journal of Official Statistics, 8*(3), 277-291.

Tourangeau, R., Kreuter, F., & Eckman, S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly, 76*(3), 453-469. doi: 10.1093/poq/nfs033

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J., B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology, 2*(178). doi: 10.3389/fpsyg.2011.00178

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann.

methods, data, analyses | Vol. 8(2), 2014, pp. 223-242 DOI: 10.12758/mda.2014.009

The Use of Respondent Incentives in PIAAC: The Field Test Experiment in Germany

Silke Martin[1], *Susanne Helmschrott*[2] & *Beatrice Rammstedt*[1]

[1] *GESIS – Leibniz Institute for the Social Sciences*
[2] *University of Mannheim*

Abstract

In PIAAC, each participating country was required to attain a response rate of at least 50%, as long as evidence was provided that there was either no or only low nonresponse bias in the data. Achieving 50% is a challenge for face-to-face surveys in most Western countries and also in Germany. Previous research showed that the use of incentives is an effective tool to increase response rates in different kinds of surveys. However, incentives may have differential effects on certain socio-demographic groups, because the perceived benefits of an incentive are subjective. To assess the effects of incentives on response rate and nonresponse bias, an experiment with three incentive treatments (€10-coin, €25 and €50 in cash) was implemented in the German PIAAC field test. Results show that response rates increased as the incentive increased. With regard to nonresponse bias, the results are less explicit. According to logistic regressions, the main factors for participation in the €50 condition are age, citizenship, and municipality size and in the €25 condition, only municipality size. Bivariate analyses put these results into perspective. For all treatment groups, a low potential for bias is visible, and there is no statistical evidence that response distributions of the realized sample across treatments are different.

Keywords: Incentive, response rate, experiment, sample composition, PIAAC

1 Introduction

The Programme for the International Assessment of Adult Competencies (PIAAC) aimed at producing a high-quality database with reliable and comparable data across the participating countries. Achieving a high response rate was one central indicator for quality in PIAAC. As defined in the international PIAAC technical standards and guidelines, and in accordance with similar cross-national studies, such as the European Social Survey (ESS; Koch, Fitzgerald, Stoop, Widdop, & Halbherr, 2012), the target response rate for each country was set to 70% (OECD, 2010a). Response rates of between 50% and 70% were typically accepted, as long as evidence was provided that there was either no or only low nonresponse and undercoverage bias in the data. Countries not meeting the minimal response rate requirement of 50% were usually not included in the international data set and reports (OECD, 2010a).[1]

Passing the benchmark of a minimum response rate of 50% was a challenge for several countries in PIAAC, because non-participation in large-scale face-to-face surveys is a growing concern all over the world (Atrostic, Bates, Burt, & Silberstein, 2001; Couper & de Leeuw, 2003; de Leeuw & de Heer, 2002; Dixon & Tucker, 2010). As Blohm and Koch (2013) report, for example, four of the 27 countries in the 2010 round of the ESS (European Social Survey, 2012) and eight of the 27 countries participating in 2011 in the European Quality of Life Survey (Eurofound, 2012) failed to reach response rates of 50%.

A serious and constant decrease in response rates for registry-based surveys is also clearly detectable in Germany. National probability surveys in Germany, such as the German General Social Survey (Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS), have experienced a decline in response rates throughout recent years: from approximately 54% in 1994 (Koch, Gabler, & Braun, 1994) to 34% in 2010 (Wasmer, Scholz, Blohm, Walter, & Jutz, 2012). Analogously, in the first round of the ESS, Germany achieved a response rate of approximately 56% (European Social Survey, 2002), whereas in the last two rounds 5 and 6, response

1 Countries were only to be included if analyses indicated that the potential bias is not greater than a potential bias introduced by a response rate between 50% and 70%.

Direct correspondence to
Silke Martin, GESIS – Leibniz Institute for the Social Sciences,
P.O. Box 12 21 55, 68072 Mannheim, Germany
E-mail: silke.martin@gesis.org

Acknowledgment: The authors sincerely thank Michael Blohm and Achim Koch for their valuable input in the decision process identifying an appropriate incentive strategy in PIAAC.

rates of only 31% and 34%, respectively, were realized (European Social Survey, 2012, 2013).

In addition to achieving a substantial response rate, keeping the nonresponse and undercoverage bias negligible, or at least low, was a second major quality criterion in PIAAC (OECD, 2010a). If non-participation in a survey follows a systematic pattern, such that certain groups of sampled persons are less likely to participate than others, nonresponse may cause bias in the data and thus have an impact on the quality of the data (Groves, 2006). Offering an incentive could have a differential effect on the propensity to participate for certain groups and can thus either introduce or reduce nonresponse bias.

Given these standards for PIAAC, the recommendation to use incentives was explicitly embedded in the PIAAC technical standards and guidelines and countries were encouraged to adopt an incentive strategy for improving response rates (OECD, 2010a). The vast majority of the participating countries used some form of incentive or a selection of several incentives during the PIAAC field test (OECD, 2010b). However, only in five countries, Denmark, Germany, Norway, the United Kingdom, and the United States, an incentive experiment was implemented. The incentive experiment in Germany employed three different monetary incentive conditions and aimed to identify the most suitable incentive strategy for the main study.[2]

2 Past Research on Incentives

Previous studies have demonstrated that response rates increase when incentives are provided (e.g., Börsch-Supan, Krieger, & Schröder, 2013; Church, 1993; Singer, 2002; Singer & Kulka, 2002; Singer, Van Hoewyk, Gebler, Raghunathan, & McGonagle, 1999; Singer & Ye, 2013). In his meta-analysis, Church (1993) analyzed 38 mail surveys that have commonly used monetary and non-monetary incentives over the last decades, and concluded that, overall, incentives have a positive effect on the response rate. In particular, the results showed that (a) prepaid incentives work better than conditional incentives, (b) monetary incentives are more effective than non-monetary incentives, and (c) response rates increase with the monetary value of the incentive. In face-to-face or telephone surveys, the effectiveness of incentives has been investigated less (Blohm & Koch, 2013). The most prominent study in this context is the meta-analysis by Singer et al. (1999) of 39 incentive experiments conducted in interviewer-administered surveys in the United States and Canada. They verified that the previously identified effects of incentives

2 PIAAC in Germany was funded by the Federal Ministry of Education and Research (BMBF) with the participation of the Federal Ministry of Labor and Social Affairs (BMAS).

on response rates, although generally smaller than in mail surveys, are also present in face-to-face and telephone surveys. Further international research on incentive experiments conducted in face-to-face or telephone surveys, some of them panel surveys, more or less replicated these findings (e.g., Castiglioni, Pforr, & Krieger, 2008; Eyerman, Bowman, Butler, & Wright, 2005; Jäckle & Lynn, 2008; Rodgers, 2011; Schröder, Saßenroth, Körtner, Kroh, & Schupp, 2013; Singer & Kulka, 2002).

For cross-sectional large-scale assessment surveys, like PIAAC, there is only limited published evidence on the use of incentives or on incentive experiments, to date. This type of survey places some large burdens on respondents because, in addition to a long interviewer-administered interview, respondents have to complete a cognitive assessment on their own. Incentives can be a helpful tool to compensate for this additional burden. In the two central international adult assessment surveys that precede PIAAC, the International Adult Literacy Survey (IALS) and the Adult Literacy and Life Skills Survey (ALL), the use of monetary incentives was prohibited (Murray, Kirsch, & Jenkins, 1997; Statistics Canada, 2002). In IALS, however, Sweden and Germany deviated slightly from this guideline and offered small symbolic incentives. In ALL 2003, the United States provided a conditional incentive of $35 (Krenzke, Mohadjer, & Hao, 2012).

Berlin et al. (1992) and Mohadjer et al. (1997) reported results from an incentive experiment implemented in the 1992 National Adult Literacy Survey (NALS; U.S. Department of Education, 2001) that are in line with the literature and show that incentives significantly increase the response rates. Results from the other PIAAC field test experiments go into the same direction. In the United Kingdom, vouchers (worth £20 or £30) were offered and results showed a significant difference in response rate in favor of the higher incentive (Department for Business Innovation & Skills, 2013).[3] In the United States two incentive conditions ($35 and $50) were concurrently tested. Krenzke et al. (2012) showed that the refusal rate was significantly lower in the $50 condition.

Incentives may have differential effects on certain socio-demographic groups because the perceived benefits of an incentive are subjective and therefore could affect the sample composition (Singer & Kulka, 2002). The effect may be positive and reduce nonresponse bias, when incentives draw individuals into the sample who would otherwise be more prone to refuse (Singer & Ye, 2013). Only a few studies have investigated the effects of incentives on sample composition and response distributions, to date (e.g., Blohm & Koch, 2013; Eyerman et al., 2005; McGrath, 2006; Nicolaas & Stratford, 2005; Singer, 2002; Singer et al., 1999). In summary, these studies provide mixed results. Whereas some studies found no (major) effects of providing incentives on the sample composition (e.g., Blohm & Koch, 2013; Eyerman et al., 2005), other studies report evidence that offering an incentive sup-

3 Results of the Danish and the Norwegian experiment are not available, to date.

ported the recruitment of respondent groups into the sample that otherwise would be underrepresented in the survey, such as e.g., low-income or minority respondents (e.g., Singer, 2002; Singer et al., 1999). In the 1992 NALS, Berlin et al. (1992) found evidence for self-selection of better-educated and wealthy people into the zero-incentive condition, resulting in an overestimation of the population's literacy level in this treatment group.

In the German context, Pforr et al. (forthcoming) have compiled information on ten incentive experiments conducted in eight large-scale face-to-face surveys (two cross-sectional surveys, ALLBUS and PIAAC, and six panel surveys[4]). Given the variation in study and experimental design of these eight surveys, the findings always only refer to some of the analyzed surveys. Pforr et al. found evidence that incentives increase response and retention rates and demonstrated that an increase of the monetary incentive value results in a higher response rate. Cash incentives were more effective than lottery tickets. Mixed results were found with regard to the effects of incentives on nonresponse bias. For several socio-demographic variables, the variable distributions across experimental conditions were analyzed. The results for cross-sectional face-to-face studies indicated that incentives did not affect the sample composition for the selected variables, whereas, for some panel studies, evidence emerged that some groups of respondents were more attracted by incentives than others.

The present paper aims to investigate two central questions in the context of the German PIAAC field test experiment: Do incentives have a positive effect on the response rate? Is there a differential effect of incentives on the sample composition and response distribution?

3 Method

The PIAAC field test had the function of a dress rehearsal for the main study and aimed to define and evaluate, amongst other key aspects, sampling, interviewer training, and survey operation procedures. In Germany, all procedures were implemented as closely as possible to the PIAAC main study parameters. However, given a shorter data collection period than in the main study,[5] some of the main study fieldwork measures (e.g., refusal conversion in re-issue phases or tracing addresses of sampled persons that had moved) could not be realized in the field test.

A registry-based, three-stage stratified and clustered sample design was implemented for the PIAAC field test in Germany, and a gross sample of 3,455 persons

4 German Internet Panel, National Educational Panel Study, German Family Panel, Panel Study "Labor Market and Social Security", Survey of Health, Aging and Retirement, and the Socio-Economic Panel.

5 In Germany, field test data collection took place from April to June 2010.

was selected. In order to depict a representative distribution of small, medium, and large municipalities in Germany, but on a smaller scale, the field test was conducted in only five federal states: Bavaria, Hamburg, Saxony, Schleswig-Holstein, and Thuringia (for more information see Zabal et al., 2014). To obtain a sufficient number of selected persons per federal state, the sampling occurred disproportionately, with oversampling Hamburg and Schleswig-Holstein and selecting fewer cases in Bavaria.

The PIAAC field test in Germany employed three monetary incentives: a €10 commemorative silver coin, engraved with a motif of the 2006 Soccer World Cup, €25 in cash, and €50 in cash. They were randomly allocated to individuals in the gross sample within each sample point with the ratio of 20:40:40.[6] In general, a sample point was allocated to one single interviewer, which ensured that each interviewer worked in all three incentive conditions.

Given the fact that the incentive experiment was not an independent scientific endeavor, the experimental design had a clear limitation. The PIAAC interview, consisting of an interviewer-administered background questionnaire and a self-administered cognitive assessment part, had an average duration of 1 hour and 40 minutes. The decision to use a €10-coin as the baseline was made to account for this substantial interview burden. Thereby, however, we were not able to analyze effects of an incentive compared to a zero-incentive condition. The analyses of the incentive experiment were based on a gross sample of 3,383 eligible cases and a net sample of 1,183 cases (unweighted counts).[7]

Sampled individuals were informed about the survey and the incentive through an advance letter that was sent to them prior to the first interviewer contact. Similarly, interviewers knew which incentive amount was assigned to each sampled individual and could use this information deliberately as a door-opener. Interviewers were instructed to provide the incentive to respondents at the end of an interview.

In contrast to other response rate calculations standards, such as defined by AAPOR (The American Association for Public Opinion Research, 2011), the response rate in PIAAC is a product of the background questionnaire response rate and the cognitive assessment response rate (cf. Mohadjer, Krenzke, & Van de Kerckhove, 2013, p. 12). In accordance with this definition, response rate analyses were calculated by counting full interviews and refusals of the assessment in the numerator as completed cases[8] and subtracting ineligibles and impairments from

6 Overall, the allocation of incentives with this ratio had been implemented successfully across the treatments.

7 Six cases were excluded, because respondents received €50 instead of the pre-assigned incentive amount.

8 We deviated from the completed case definition (see Mohadjer et al., 2013) by excluding literacy-related nonrespondents.

the group of sample persons in the denominator. The following dispositions were summarized as ineligibles and impairments: Death, sample person moved (a) into institution or (b) outside country, hearing and blindness/visual impairment, physical and other disability.

In the PIAAC field test, a proxy variable of proficiency was calculated for each respondent, instead of producing a set of plausible values[9] for each skill domain, as in the main study. This proxy variable is a standardized logit score based on a transformation of the proportion of correct responses to the assessment items (PIAAC Consortium, 2010).

Analysis Plan

In order to answer our research questions, we first compared differences in response rates and nonresponse rates across the treatment conditions. Subsequently, we analyzed whether the incentive conditions potentially introduced some bias. We used variables from the sampling frame, such as age (in five categories), gender, citizenship (in two categories: German and other), and municipality size in three categories (large, medium, and small) that were available for both respondents and nonrespondents, and ran logistic regressions with response as the dichotomous dependent variable for each incentive condition separately. We decided to not include any data from the interviewer case folders or a consumer marketing database (Microm) in these analyses. Although they are, in general, available for all eligible units, they have quality limitations. Case folder information is subject to measurement error, because in the field test, information was not collected in the standardized way like in the main study. Microm variables do not reflect individual case-wise information, but are rather aggregated (information from up to 500 households is combined) and some have a substantial amount of missing data, most probably because sampled addresses could not be categorized.

In a next step, we looked at response distributions of several socio-demographic variables for each treatment group and compared them to the corresponding data from the German 2008 Microcensus, provided by the Federal Statistical Office. We used 2008 Microcensus data because, in 2010, when we first analyzed the experimental data to make a decision for the main study incentive, these were the most current official and available data at that time.

Additional analyses focused on the effects of incentives on the sample composition by comparing response distributions across the incentive treatments, using Chi-Square-Tests of Independence and propensity weighting.

---

9 For definition and computation of plausible values see Yamamoto, Khorramdel, & Von Davier, 2013.

4 Results

To analyze the extent by which response rates increase when a monetary incentive is provided, response rates of the three incentive treatments were compared by means of Chi-Square-Tests of Independence. Table 1 shows the response and nonresponse rates for the overall sample as well as for each treatment group separately. The nonresponse rate is split into nonresponse due to refusal, non-contact, address-related issues, and other reasons.

In the €50 condition, the achieved response rate was 41.7%, compared to 35.4% in the €25 treatment and 26.5% in the €10-coin group. All differences are significant. Even though the PIAAC target response rate of 50% is not achieved for any of the treatment groups, the results clearly demonstrate an increase of the response rate with increasing incentive size.

In general, nonresponse was particularly due to refusals (41.1%, for the overall sample, compared to 22.7% for the remaining reasons). While the response rate increased from lowest to highest incentive amount, the refusal rate developed in the opposite direction: the higher the incentive, the lower the refusal rate. The refusal rates for both the €25 and the €50 condition differed significantly in comparison to the €10-coin group ($p<.01$ and $p<.001$, respectively). Further, the rates for non-contacts, address-related issues, and other reasons for non-participation were also slightly lower in the €50 condition, but these differences did not reach statistical significance.

The second research question addresses the aspect of selectivity in response across treatment groups and differential effects on the sample composition. At first, effects of socio-demographic frame variables on response behavior (1 = response; 0 = nonresponse) were tested for each treatment group separately by means of logistic regressions with the following explanatory variables from the frame:

(a) Age: 16-25 (reference category)/26-35/36-45/46-55/56-65;

(b) Gender: Male (reference category)/female;

(c) Citizenship: German (reference category)/other;

(d) Municipality size: Large with 100,000 or more inhabitants (reference category)/medium with 20,000 to under 100,000 inhabitants/small with under 20,000 inhabitants.

Distributions of the explanatory variables, separately for respondents and nonrespondents, are given in Table A1.1 in the Appendix. Results of the logistic regressions are summarized in Table 2 and indicate no significant effects for the €10-coin incentive group. For both the €25 and the €50 condition, the results demonstrated that individuals living in small municipalities have a significantly higher propensity to participate, compared to individuals residing in large municipalities ($p<.001$). In the €50 condition, this effect was also found for sampled persons living in medium

Table 1 Response and nonresponse rates by incentive treatment

	Overall (n = 3,381)	€10-coin (n = 660)	€25[a] (n = 1,374)	€50[b/c] (n = 1,347)
	%	%	%	%
Response rate	36.2	26.5	35.4***	41.7***/**
Refusal rate	41.1	48.6	40.6**	37.9***/n.s.
Non-contact rate	8.4	9.9	8.5n.s.	7.4n.s./n.s.
Nonresponse rate - address issues	6.8	8.2	6.8n.s.	6.2n.s./n.s.
Nonresponse rate - other reasons	7.5	6.8	8.7n.s.	6.8n.s./n.s.

Notes: Number of cases = eligible sample. To account for disproportionality in sampling, data are adjusted by a correction factor.
a χ^2-Test for comparison of €10-coin and €25
b χ^2-Test for comparison of €10-coin and €50
c χ^2-Test for comparison of €25 and €50
* = $p < .05$, ** = $p < .01$, *** = $p < .001$, n.s. = not significant

municipalities (p<.01). In addition, the €50 incentive seemed to be more attractive for younger individuals and persons with German citizenship. The 36-45 (p<.01), 46-55 (p<.05), and 56-65 (p<.01) age-groups responded significantly less often than the 16 to 25 year-olds. In the €50 treatment, citizenship also had an effect on participation; individuals with non-German citizenship had a lower propensity of providing an interview, but this result was only significant at the 5%-level. While the pseudo R^2 in the €50-model is the highest across all models, overall, the values of the pseudo R^2 are low for all models, indicating only a weak explanation of response behavior through the independent model variables. In addition, significant correlations (p<.01) showed only low strengths between response status and municipality size in the €25 condition (r=-.116) and between response status and age (r=-.085), citizenship (r=.086) and municipality size (r=-.103) in the €50 condition.

In a second step, we compared the response distributions of central sociodemographic variables, for each incentive condition separately, with the corresponding distributions from the German 2008 Microcensus.[10] The response distributions for several frame and survey-relevant outcome variables, such as highest school qualification and employment status, are given in Table 3. With regard to

10 When comparing response distributions with reference data, differences are not only induced by nonresponse bias, but can be due to other error sources (e.g., noncoverage or sampling). The noncoverage rate was low (cf. Zabal et. al, 2014, for main study), and sampling bias is expected to be low, due to probability sampling.

Table 2 Logistic regression of response behavior on frame variables for each incentive treatment

	€10-coin		€25		€50	
	β	*SE*	*β*	*SE*	*β*	*SE*
Gender						
Male (*ref. cat.*)						
Female	-.292	(.183)	-.071	(.116)	-.036	(.115)
Age						
16 to 25 (*ref. cat.*)						
26 to 35	-.073	(.291)	.007	(.194)	-.082	(.193)
36 to 45	-.281	(.285)	-.223	(.182)	-.479**	(.178)
46 to 55	-.376	(.289)	-.172	(.187)	-.391*	(.178)
56 to 65	-.202	(.310)	-.316	(.198)	-.521**	(.190)
Citizenship						
German (*ref. cat.*)						
Other	-.340	(.378)	-.159	(.223)	-.596*	(.232)
Municipality size (No. of inhabitants)						
100,000+ (*ref. cat.*)						
20,000 to <100,000	-.068	(.294)	.105	(.188)	.535**	(.178)
<20,000	.254	(.208)	.565***	(.135)	.474***	(.131)
Constant	-.702**	(.270)	-.636***	(.174)	-.209	(.173)
n	618		1,289		1,282	
Pseudo R^2	0.020		0.027		0.040	

Notes: To account for disproportionality in sampling, data are adjusted by a correction factor.
* = $p < .05$, ** = $p < .01$, *** = $p < .001$

gender, the samples of all treatments included more men than women, when compared to the Microcensus. Whereas the distribution is fairly close to the reference data for both of the cash alternatives, the difference in the €10-coin distribution is obvious. This could be due to the motif of the €10-coin, which was related to the 2006 Soccer World Cup and might have been more attractive to male individuals.

Similar to the effects observed in the multivariate analyses, it can be seen that the €50 incentive was more attractive for the youngest age group. However, at the bivariate level, the proportion of 16 to 25 year-olds is only slightly higher, compared to the Microcensus data, and the share of 36 to 45 year-olds is slightly smaller. The distribution of age in the €25 condition shows the best fit with the Microcensus data, while there are some minor deviations from the expected distri-

Table 3 Comparison of survey estimates with German Microcensus data 2008

	€10-coin %	€25 %	€50 %	MC 08 %
Gender				
Male	57.7	51.9	52.4	50.7
Female	42.3	48.1	47.6	49.3
Age				
16 to 25	18.9	18.1	21.4	18.2
26 to 35	21.7	18.9	18.3	18.3
36 to 45	22.9	24.5	21.7	24.2
46 to 55	20.6	22.0	22.4	21.9
56 to 65	16.0	16.5	16.2	17.5
Citizenship				
German	94.3	93.2	94.7	91.4
Other	5.7	6.8	5.3	8.6
Municipality size (No. of inhabitants)				
100,000+	30.3	26.7	26.4	29.1
20,000 to <100,000	12.6	13.0	17.3	17.3
<20,000	57.1	60.3	56.3	53.6
Highest school qualification				
Low	32.4	25.7	28.4	33.7
Medium	39.9	37.1	39.5	37.7
High	27.7	37.1	32.1	28.6
Employment status				
Employed	80.3	74.5	75.2	70.2
Unemployed	4.0	3.9	5.2	8.3
Inactive	15.6	21.6	19.6	21.5

Notes: To account for disproportionality in sampling, data are adjusted by a correction factor. Microcensus estimates are based on data for the target group of 16 to 65 year olds in Bavaria, Hamburg, Saxony, Schleswig-Holstein, and Thuringia.

bution in the €10-coin condition. Altogether, there is no indication that any of the three distributions of age clearly deviates from the Microcensus distribution.

At the bivariate level, it can be seen that each of the three incentives attracted more target persons with German than with non-German citizenship into the sample, although the effect of citizenship on response behavior in the logistic regression model only reached statistical significance in the €50 condition. Overall, the €25 condition had a slightly better distribution than the €50 condition or the €10-coin group, in comparison to the reference data.

Results observed in the multivariate analysis for municipality size were also visible in the bivariate analysis. Distributions across categories of the variable municipality size showed some deviations from Microcensus distribution in all incentive treatments. Whilst in the €10-coin group, the proportion of persons living in large municipalities was closest to official data, in the €50 condition, the proportion of persons living in medium municipalities matched the Microcensus data perfectly. Altogether, the observed distribution in the €25 treatment deviated clearly from the Microcensus distribution, mainly because the share of residents in medium municipalities is considerably lower and the share of residents in small municipalities considerably higher than in the Microcensus.

Regarding educational attainment, measured as the highest German general school leaving qualification obtained, the €10-coin group revealed a distribution that closely followed the Microcensus distribution, whilst both the €25 and the €50 conditions differed, in comparison to the Microcensus. However, a comparison of these two conditions reveals that the response distribution in the €50 condition was closer to the reference data than the response distribution in the €25 condition, mainly due to a considerable underrepresentation of persons with a low educational level and an overrepresentation of persons with a high educational level in the €25 group.

Next to educational attainment, employment status is considered as a central outcome variable in PIAAC, because skills and employment status are closely linked (Klaukien et al., 2013; OECD, 2013). The distribution of employment status differed considerably from the Microcensus distribution in each treatment group. Particularly in the €10-coin treatment, employed individuals are overrepresented, whereas unemployed and inactive persons are underrepresented.

In order to investigate differential effects of incentives on the sample composition, we analyzed differences in the response distributions for a range of variables across treatment groups by using Chi-Square-Tests of Independence. Results summarized in Table 4 indicate that neither the €25, nor the €50 condition revealed any significant differences in the response distributions for any of the variables, when compared to the €10-coin treatment or to one another.

In addition, we investigated if the incentive treatments differed in the mean outcome variable, the proxy of proficiency. This logit score in the German PIAAC net sample ranges from -4.5360, a value that indicates a lower proficiency level, to 2.7591, a value that represents a higher skill level. Given an average of -.1475 (with a standard deviation of 1.1110) for the overall sample, all of the corresponding logit score means in the three treatment groups were fairly close to this average. While the logit score means of the €10-coin and the €25 treatment (-.1216 and -.1279, respectively; see Table 4) were slightly higher, the logit score mean in the €50 condition was slightly lower (-.1726). Results of the *t*-test for independent samples, however, revealed no significant differences between the treatment groups.

Table 4 Comparison of survey estimates across incentive treatments

	€10-coin %	€25 %	€50 %
Gender (χ^2-*Test*)		(n.s.)	(n.s./n.s)
Male	57.7	51.9	52.4
Female	42.3	48.1	47.6
Age (χ^2-*Test*)		(n.s.)	(n.s./n.s)
16 to 25	18.9	18.1	21.4
26 to 35	21.7	18.9	18.3
36 to 45	22.9	24.5	21.7
46 to 55	20.6	22.0	22.4
56 to 65	16.0	16.5	16.2
Citizenship (χ^2-*Test*)		(n.s.)	(n.s./n.s)
German	94.3	93.2	94.7
Other	5.7	6.8	5.3
Municipality size (χ^2-*Test*)		(n.s.)	(n.s./n.s)
100,000+	30.3	26.7	26.4
20,000 to <100,000	12.6	13.0	17.3
<20,000	57.1	60.3	56.3
Highest school qualification (χ^2-*Test*)		(n.s.)	(n.s./n.s)
Low	32.4	25.7	28.4
Medium	39.9	37.1	39.5
High	27.7	37.1	32.1
Employment status (χ^2-*Test*)		(n.s.)	(n.s./n.s)
Employed	80.3	74.5	75.2
Unemployed	4.0	3.9	5.2
Inactive	15.6	21.6	19.6
	Mean	Mean (n.s.)	Mean (n.s./n.s)
Proficiency proxy (*t-Test*)	-.1216	-.1279	-.1726

Notes: To account for disproportionality in sampling, data are adjusted by a correction factor. n.s. = not significant

Table 5 Comparison of survey estimates for non-propensity and propensity weighted data across incentive treatments

	€10-coin		€25		€50	
	non-propensity weighted	propensity weighted	non-propensity weighted	propensity weighted	non-propensity weighted	propensity weighted
	%	%	%	%	%	%
Highest school qualification						
Low	32.4	31.8	25.7	25.0	28.4	27.3
Medium	39.9	40.2	37.1	36.2	39.5	39.3
High	27.7	28.0	37.1	38.8	32.1	33.4
Employment status						
Employed	80.3	80.8	74.5	73.7	75.2	75.4
Unemployed	4.0	3.8	3.9	4.0	5.2	5.3
Inactive	15.6	15.3	21.6	22.3	19.6	19.3
	Mean	Mean	Mean	Mean	Mean	Mean
Proficiency proxy	-.1216	-.1451	-.1279	-.1201	-.1726	-.1922

In order to simulate a sample distribution that would have resulted if all sample persons had participated, we finally weighted the data with propensity weights that accounted for differential response behavior and which were computed in the logistic regression models for each incentive treatment separately. Distributions of propensity weighted variables were further compared to the distributions of variables without propensity weights (see Table 5). Considerable differences in the distributions would be an indicator that differential response behavior has an effect on the sample composition.

As Table 5 depicts, propensity weights had hardly any effect on the distribution of educational attainment in any of the incentive treatment groups. With regard to employment status, there was no effect on the distribution from propensity weighting for the €10-coin and the €50 condition. In the €25 treatment, however, the share of employed persons was slightly reduced and the share of inactive individuals increased slightly through the weighting.

Propensity weighting in both the €10-coin and the €50 condition resulted in somewhat lower means of the proxy variables for proficiency, whereas the average proficiency score in the €25 condition became slightly higher. Given a range of 7.2951 for this variable, these changes, however, can be considered negligible.

5 Discussion

Previous studies on the use of incentives showed that incentives have a positive effect on response rates. Only few studies, however, investigated effects of incentives on the sample composition and response distributions. In particular, there is only limited published evidence on the use of incentives in adult assessment surveys. In the present study, we analyzed results from the experiment of testing three incentive conditions in the German PIAAC field test. Two central questions were addressed: Do incentives have a positive effect on the response rate? Is there a differential effect of incentives on the sample composition and response distribution?

Results from the response rate analysis of this experiment are straightforward. As expected from the literature, we found that incentives are an effective tool for increasing the response rate. For the PIAAC field test incentive experiment, we observed that response rates increased significantly with increasing amounts of the incentive.

With regard to the second research question the results are less explicit. While results of the multivariate analyses indicate a potential for bias in the €25 and €50 condition for municipality size, these results are put into perspective, at least partly for age and citizenship, based on the bivariate analyses, e.g., by comparing response distributions to Microcensus data or across incentive treatments. Response distribution of citizenship and municipality size differ across all treatment groups when compared to the reference data. Results thus indicate that non-German individuals and persons who live in large municipalities have, in general, a lower response propensity.

The €50 incentive was, however, more attractive for 16 to 25 year-olds. This effect is significant in the logistic regression and results in a slightly higher proportion of 16 to 25 year-olds, compared to the Microcensus. However, the difference in the response distribution of age in the €50 condition does not reach statistical significance when compared to the response distribution of age in the €25 or €10-coin condition.

For educational attainment, the results reveal that the variable's distribution in the €10-coin group showed the best match with Microcensus data. Both cash alternatives introduced some bias in the data, but compared to the €25 treatment, educational levels are better represented in the €50 condition. With regard to employment status, none of the response distributions of any incentive treatment matched the distribution of the Microcensus data well.

In general, comparisons of response distributions of central socio-demographic variables and of the mean logit scores across all incentive conditions did not provide evidence that the incentive size changed the sample composition in any treatment group in a substantial way. In contrast to results reported by Berlin et al. (1992) for the NALS survey in 1992, findings from the German PIAAC field test do

not confirm the observation that there is a self-selection of more educated or more skilled individuals in the condition with the smallest monetary amount (€10-coin) and, thus, an overestimation of the proficiency level in this treatment group. However, the results are not perfectly comparable, because in the NALS experiment, the control group received no incentive at all. Moreover, in the PIAAC field test data, only an approximation of proficiency was used as indicator.

Finally, by using propensity weights, obtained from the logistic regression, we see that the differential effects for response hardly changed the response distributions of educational attainment, employment status, and mean proficiency score.

In conclusion, the €50 incentive had the strongest positive effect on the achieved response rate. In this condition, some groups of people had a higher propensity to participate. This had, however, only a minor impact on the sample composition. Moreover, there is a low potential for bias in the data for each treatment group, because response distributions of some variables show minor deviations in each of the treatments, compared to Microcensus data. When response distributions of each treatment were compared with one another, statistical evidence that they are different could not be found.

For future cycles of PIAAC, it would be interesting to assess whether the current findings can be replicated in other participating countries and to which extent different survey operation designs, protocols and procedures (e.g., sampling designs, different types of data collection agencies or fieldwork instructions for interviewers) moderate the results. In the context of large-scale adult assessment surveys, further research on the impact of incentives on final proficiency scores (as computed in terms of plausible values) would be beneficial in order to evaluate potential motivational effects of the incentive amount on the respondent's effort to accomplish the cognitive assessment part.

References

Atrostic, B. K., Bates, N., Burt, G., & Silberstein, A. (2001). Nonresponse in U.S. Government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics, 17*(2), 209-226.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I. S., Rock, D., & Yamamoto, K. (1992). An experiment in monetary incentives. In: American Statistical Association, *Proceedings of the Survey Research Methods Section 1992* (pp. 393-398). Retrieved September 2014 from http://www.amstat.org/sections/srms/Proceedings/allyearsf.html

Blohm, M., & Koch, A. (2013). Respondent incentives in a national face-to-face survey: Effects on outcome rates, sample composition and fieldwork efforts. *methods, data, analyses,* 7(1), 89-122.

Börsch-Supan, A., Krieger, U., & Schröder, M. (2013). *Respondent incentives, interviewer training and survey participation*. SHARE Working Paper Series 12-2013. Munich: Munich Center for the Economics of Aging (MEA).

Castiglioni, L., Pforr, K., & Krieger, U. (2008). The effect of incentives on response rates and panel attrition: Results of a controlled experiment. *Survey Research Methods,* 2(3), 151-158.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly,* 57(1), 62-79.

Couper, M. P., & de Leeuw, E. D. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 157-177). New York: John Wiley & Sons.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse. A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York: John Wiley & Sons.

Department for Business Innovation & Skills. (2013). *The International Survey of Adult Skills 2012: Adult literacy, numeracy and problem solving skills in England - Appendices.* BIS Research Paper No. 139A. London: Department for Business, Innovation and Skills. Retrieved September 2014 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/248689/bis-13-1221an1-international-survey-of-adult-skills-2012-appendices.pdf

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 593-630). Bingley: Emerald.

Eurofound. (2012). *Third European Quality of Life Survey – Quality of life in Europe: Impacts of the crisis.* Luxembourg: Publications Office of the European Union.

European Social Survey. (2002). *ESS1-2002 Documentation Report. The ESS Data Archive.* (6.3 ed.). Bergen: Norwegian Social Science Data Services. Retrieved September 2014 from http://www.europeansocialsurvey.org/docs/round1/survey/ESS1_data_documentation_report_e06_3.pdf

European Social Survey. (2012). *ESS5-2010 Documentation Report. The ESS Data Archive.* (3.0 ed.). Bergen: Norwegian Social Science Data Services. Retrieved March 2014 from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e03_0.pdf

European Social Survey. (2013). *ESS6-2012 Documentation Report. The ESS Data Archive* (1.3 ed.). Bergen: Norwegian Social Science Data Services. Retrieved March 2014 from http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_data_documentation_report_e01_3.pdf

Eyerman, J., Bowman, K., Butler, D., & Wright, D. (2005). The differential impact of incentives on refusals: Results from the 2001 national household survey on drug abuse incentive experiment. *Journal of Economic and Social Measurement,* 30, 157-169.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly,* 70(5), 646-675.

Jäckle, A., & Lynn, P. (2008). Respondent incentives in a multi-mode panel survey. Cumulative effects on nonresponse and bias. *Survey Methodology,* 34(1), 105-117.

Klaukien, A., Ackermann, D., Helmschrott, S., Rammstedt, B., Solga, H., & Wößmann, L. (2013). Grundlegende Kompetenzen auf dem Arbeitsmarkt. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich. Ergebnisse von PIAAC 2012* (pp. 127-165). Münster: Waxmann.

Koch, A., Fitzgerald, R., Stoop, I., Widdop, S., & Halbherr, V. (2012). *Field Procedures in the European Social Survey Round 6: Enhancing Response Rates.* Mannheim: European Social Survey, GESIS. Retrieved 16 November 2014 from http://www.europeansocialsurvey.org

Koch, A., Gabler, S., & Braun, M. (1994). *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 1994*. ZUMA-Arbeitsbericht 94/11. Mannheim: ZUMA.

Krenzke, T., Mohadjer, L., & Hao, H. (2012). *Programme for the International Assessment of Adult Competencies: U.S. Incentive Experiment*. Paper presented at the 67th Annual Meeting of the American Association of Public Opinion Research, May 2012, Orlando, FL.

McGrath, D. E. (2006). *An incentives experiment in the U.S. consumer expenditure quarterly survey.* Paper presented at the Joint Statistical Meetings, ASA Section on Survey Research Methods, August 2006, Seattle, WA.

Mohadjer, L., Berlin, M., Rieger, S., Waksberg, J., Rock, D., Yamamoto, K., . . . Kolstad, A. (1997). The role of incentives in literacy survey research. In A. C. Tuijnman, I. S. Kirsch & D. A. Wagner (Eds.), *Adult basic skills: Innovations in measurement and policy analysis* (pp. 209-244). Cresskill, NY: Hampton Press.

Mohadjer, L., Krenzke, T., & Van de Kerckhove, W. (2013). Indicators of the quality of the sample data. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)* (Chapter 16, pp. 1-30). Paris: OECD. Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1997). *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey.* Washington D.C.: National Center for Education Statistics.

Nicolaas, G., & Stratford, N. (2005). A plea for the tailored use of respondent incentives. In: C. van Dijkum, J. Blasius, & C. Durand (Eds.), Proceedings of Sixth International Conference on Social Science Methodology: Recent developments and applications in social research methodology (pp. 1–15). Amsterdam: Budrich Verlag.

OECD. (2010a). *PIAAC Technical Standards and Guidelines* (December 2010). Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm

OECD. (2010b). *Survey Operations Quality Control Field Test Monitoring Report. PIAAC(2010_12)FT_SurveyOperation_QCFinalReport_11Dec10.docx*. Unpublished document.

OECD. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*: Retrieved March 2014 from http://www.oecd.org/site/piaac/Skills%20volume%201%20%28eng%29--full%20v12--eBook%20%2804%2011%202013%29.pdf

PIAAC Consortium. (2010). *Standardized_Logit_File_Description.pdf.* Unpublished document.

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., . . . Rammstedt, B. (forthcoming). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly.*

Rodgers, W. L. (2011). Effects of increasing the incentive size in a longitudinal study. *Journal of Official Statistics*, 27(2), 279-299.

Schröder, M., Saßenroth, D., Körtner, J., Kroh, M., & Schupp, J. (2013). Experimental evidence of the effect of monetary incentives on cross-sectional and longitudinal response: Experiences from the Socio-Economic Panel (SOEP). *SOEPpapers on Multidisciplinary Panel Data Research,* No. 603. Berlin: DIW.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, D. A. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 163-177). New York: John Wiley Co.

Singer, E., & Kulka, R. A. (2002). Paying respondent for survey participation. *Survey methodology program, No. 092*.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics,* 15(2), 217-230.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science,* 645(1), 112-141.

Statistics Canada. (2002). *Public use microdata file. User's manual*. Ottawa.

The American Association for Public Opinion Research. (2011). *Standard definitions: Final dispositions of case codes and outcome rates for surveys. 7th edition*.

U.S. Department of Education, National Center for Education Statistics (2001). *Technical report and data file user's manual for the 1992 National Adult Literacy Survey.* Washington, DC: National Center for Education Statistics.

Wasmer, M., Scholz, E., Blohm, M., Walter, J., & Jutz, R. (2012). *Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2010.* GESIS-Technical Reports. Köln: GESIS – Leibniz-Institut für Sozialwissenschaften.

Yamamoto, K., Khorramdel, L., & Von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the Survey of Adult Skills (PIAAC)* (Chapter 17, pp. 1-33). Paris: OECD. Retrieved March 2014 from http://www.oecd.org/site/piaac/surveyofadultskills.htm

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann.

Appendix

Table A1.1 Distributions of socio-demographic variables for respondents and nonrespondents

	€10-coin		€25		€50	
	R	NR	R	NR	R	NR
	%	%	%	%	%	%
Gender	(*n*=175)	(*n*=485)	(*n*=486)	(*n*=888)	(*n*=561)	(*n*=786)
Male	57.7	51.3	51.9	49.4	52.4	50.6
Female	42.3	48.7	48.1	50.6	47.6	49.4
Age	(*n*=175)	(*n*=479)	(*n*=486)	(*n*=876)	(*n*=561)	(*n*=781)
16 to 25	18.9	15.4	18.1	15.7	21.4	15.4
26 to 35	21.7	18.8	18.9	17.8	18.3	15.5
36 to 45	22.9	23.8	24.5	25.3	21.7	25.7
46 to 55	20.6	25.7	22.0	21.7	22.4	23.6
56 to 65	16.0	16.3	16.5	19.5	16.2	19.8
Citizenship	(*n*=175)	(*n*=452)	(*n*=486)	(*n*=823)	(*n*=561)	(*n*=725)
German	94.3	91.6	93.2	90.8	94.7	89.9
Other	5.7	8.4	6.8	9.2	5.3	10.1
Municipality size (No. of inhabitants)	(*n*=175)	(*n*=485)	(*n*=486)	(*n*=888)	(*n*=561)	(*n*=786)
100,000+	30.3	34.2	26.7	36.3	26.4	37.6
20,000 to <100,000	12.6	16.1	13.0	15.9	17.3	13.7
<20,000	57.1	49.7	60.3	47.8	56.3	48.7

Notes: R = respondents; NR = nonrespondents. To account for disproportionality in sampling, data are adjusted by a correction factor.

methods, data, analyses | Vol. 8(2), 2014, pp. 243-266 DOI: 10.12758/mda.2014.010

Nonresponse in PIAAC Germany

Susanne Helmschrott[1] *& Silke Martin*[2]

[1] *University of Mannheim*
[2] *GESIS – Leibniz Institute for the Social Sciences*

Abstract

Nonresponse is of concern for the quality of survey data, because it may introduce bias into the collected sample. To date, only few studies deal with nonresponse in skills or educational surveys. This paper aims at contributing to this field by identifying the main factors that influenced participation in the first wave of PIAAC Germany, a survey assessing skills of the adult population, conducted in 2011/2012. Using bi- and multivariate analyses, we found that age, citizenship, the level of education, the type of house the sampled persons live in, and municipality size were the main factors influencing response to PIAAC Germany. Our findings suggest that, for the effective reduction of nonresponse in skills or education studies, researchers should target persons with a low level of education, foreigners, those living in larger housing units, and big-city dwellers by using appropriate measures at the different stages of the survey process.

Keywords: Nonresponse bias, PIAAC, Germany, skills, education

1 Introduction

The full drop out or underrepresentation of certain groups of sample persons, because of nonresponse, may cause bias in the achieved sample data (Groves, 2006). Survey researchers have several possibilities to tackle nonresponse at different stages of the survey process. A common approach is to apply weighting techniques after data collection, thus alleviating nonresponse bias in the resulting sample (Lynn, 1996). However, it is just as important to avoid nonresponse right from the outset with the help of specific survey design features and fieldwork strategies. Exploring factors that influence nonresponse is thus useful for choosing suitable strategies in future waves of PIAAC, the "Programme for the Assessment of Adult Competencies"[1], or similar survey projects.

Although, to date, a multitude of studies on nonresponse across countries, survey topics, and designs has been published (e.g., Blom, 2012; Blom, de Leeuw, & Hox, 2011; de Leeuw & de Heer, 2002; Groves, 2006; Groves & Couper, 1998), publications on nonresponse in skills and education surveys are scarce (e.g., Kleinert, Ruland, & Trahms, 2013; Darcovich, Binkley & Cohen et al., 1998; Van de Kerckhove, Krenzke, & Mohadjer, 2009). However, if a survey includes the assessment of competencies or knowledge, nonresponse might be different because the greater time and cognitive burden associated with the assessment could yield a specific profile of nonrespondents. This article aims at contributing to this field of research by analyzing survey participation in the first wave of PIAAC in Germany, conducted in 2011/2012.

Our research objective is to identify the main factors influencing survey participation in PIAAC Germany. First, we provide a review of theories on survey participation, in order to identify potential correlates of nonresponse in PIAAC, and derive hypotheses on survey participation across different groups. Among the multitude of potential influencing factors, we will focus on socio-demographic, economic, and geographic characteristics of the sample persons. They are not directly causal for survey participation, but influence the latent social and psychological constructs driving the response process (Groves & Couper, 1998). However, nonresponse is only of concern if nonrespondents differ from respondents in terms of the central study outcome(s). The characteristics identified thus need to be significantly related to both response status and the central study outcome(s) (See Section 2). Hence, in order to identify those characteristics with the potential to introduce bias

1 A description of the aim and methodology of the study is given by Rammstedt & Maehler in this volume.

Direct correspondence to
Susanne Helmschrott, University of Mannheim, L13, 17, 68131 Mannheim, Germany
E-mail: helmschrott@uni-mannheim.de

in the data set, we subsequently explore these relationships at the bivariate level. Next, we fit multivariate models of survey participation and isolate the main factors influencing participation in PIAAC Germany. The results could be useful for researchers in other skills and education surveys attempting to reduce nonresponse bias from the outset of their studies.

2 Nonresponse and its Effects on Sample Quality

Nonresponse constitutes one of several sources of error that can arise during the design and implementation of a survey (Groves & Lyberg, 2011). "…[It] occurs when a sampled unit does not respond to the request to be surveyed or to particular survey questions" (Dillman, Eltinge, & Groves et al., 2002, p. 3). As the definition implies, two types of nonresponse exist: "Unit nonresponse occurs when a selected element does not provide information at all, that is, the questionnaire form remains empty. Item nonresponse occurs when some questions have been answered but no answer is obtained for some other, possibly sensitive questions" (Betlehem 2009, p. 209). Because this article deals with unit nonresponse only, "nonresponse" will always refer to unit nonresponse here.

Under certain conditions, nonresponse can yield biased survey estimates. According to Betlehem (2002), every member of the survey population has a certain propensity to respond to a survey. As the following formula shows, nonresponse bias in the respondent mean $\left(\overline{y}_r\right)$ can be approximated by the ratio of the covariance between the response propensity (p) and the survey variable (y), and the mean response propensity $\left(\overline{p}\right)$:

$$Bias\left(\overline{y}_r\right) \approx \frac{\acute{o}_{yp}}{\overline{p}}$$

The formula implies that nonresponse bias in the respondent mean depends not only on the response rate (the mean response propensity), but also on the strength of the relationship between the response propensity and the variables measured in the survey. Indeed, "… [t]he stronger the relationship between the target variable and response behavior, the larger the bias" (Betlehem, 2002, p. 276). As described by Groves (2006) in the "survey variable cause model", the most severe case of nonresponse bias is given in the case of a perfect correlation between the survey variable of interest and response propensity. Here, the survey variable is the cause of nonresponse and groups that differ from respondents, in terms of the survey variable, are completely missing from the sample. This type of nonresponse is also called "nonignorable nonresponse" (Little & Rubin, 2002), because nonresponse adjustment techniques may not be successful (Betlehem et al., 2011). Described by Groves (2006) as the "common cause model", nonresponse bias can also occur if

response propensity and survey variable arise from the same variable or set of variables. Here, the covariance between the survey variable (y) and response propensity (p) is due to a common cause of both variables. However, if the nonresponse adjustment techniques are based on the variables that caused both y and p, correction of the bias is possible (Betlehem et al., 2011).

3 Factors Influencing Survey Participation

In the following section, we provide an overview of theoretical approaches to survey participation. The aim of this synopsis is to identify common correlates of nonresponse and develop hypotheses for factors influencing participation in PIAAC in Germany.

3.1 Rational Choice Approaches

Most theories that try to explain survey participation have their roots in rational choice theory; for instance, the "opportunity cost hypothesis" (Groves & Couper, 1998) or the social exchange theory (e.g., Dillman, 1978; Goyder, 1987). Rational choice approaches assume that, when confronted with a survey request by an interviewer, sample persons weigh up all potential benefits of a survey against their costs and base their decision on the outcome of the calculation. Although rational choice approaches, in their strictest sense, assume that sample persons take their time for a careful consideration of the pros and cons of participation, Groves and Couper (1998) specify that this is rarely the case in survey practice. For example, most refusals in telephone surveys take no longer than 30 seconds (Groves et al., 2009). Instead, they suggest that the decision to participate is a heuristic act based on a superficial and quick cost-benefit analysis that is influenced by a variety of external and situational factors (Groves & Couper, 1998). In their "leverage-saliency-theory", Groves, Singer, and Corning (2000) specified that costs and benefits are not static across sample persons and that individuals differ in the "leverages" they attach to the various design features. This means that sample persons can have different perceptions about whether specific survey characteristics are benefits or costs and can differ in their evaluation of how high the respective costs and benefits of participation are.

The costs and benefits of a survey depend on its specific design features. For example, many surveys aim to encourage participation with monetary or nonmonetary incentives, which can be paid either conditionally upon participation or unconditionally to each household or sample person (Groves & Couper, 1998). Even if the type and value of the incentives differ across survey projects, types, and countries, some indisputable effects of incentives have been repeatedly demonstrated:

Incentives do increase response rates, monetary incentives are more effective than non-cash incentives, and prepaid incentives are more effective than conditional incentives (e.g., Singer, Van Hoewyk, Gebler, Raghunathan, & McGonagle, 1999; Singer & Ye, 2013). Obviously, the topic of a survey can also be seen as a cost or a benefit of participation. From a rational choice perspective, sample persons might expect greater benefits for themselves when participating in a study that is of interest to them (Groves & Couper, 1998). In an experiment on differences in response rates among various interest groups, Groves, Presser, & Dipko (2004) did indeed find higher response rates among groups interested in a topic, such as school teachers on educational topics, than in the general population. Furthermore, interview burdens are supposed to play a part in the sample persons' cost-benefit analyses. Surveys can show enormous variation regarding the time requested for completion of the questionnaire, the cognitive burden imposed by answering the questions, and the emotional burden that opening up to a stranger on sensitive topics might imply (Groves & Couper, 1998). In addition, the survey sponsor influences the decision to participate. As discussed below, surveys conducted by public authorities generally achieve higher response rates than those organized by private companies (e.g., Lyberg & Dean, 1992). Furthermore, a design feature that might positively impact the cost-benefit analysis is sending advance letters. In addition to informing the household or specific sample persons that they have been selected for participation, these letters also generally aim at encouraging cooperation by highlighting the most attractive benefits of participation (Groves & Couper, 1998).

The notion that sample persons perceive costs and benefits differently implies that some socio-demographic groups might be more attracted by specific survey design features than others. For example, rational choice approaches would assume that those with little discretionary time available, such as those in employment, should consider the time needed for completion as a greater burden than those with more available time. However, empirical tests either failed to verify the hypothesis that the amount of discretionary time available is related to survey cooperation (Groves & Couper, 1998) or found that the unemployed are more likely to refuse (Durrant & Steele, 2009). Furthermore, costs and benefits should be perceived differently across different levels of education. For example, the less educated could perceive the cognitive burden stemming from the need to complete a test such as administered in PIAAC as a higher cost than the highly educated. Furthermore, the latter might see a greater benefit in contributing to a study such as PIAAC, which informs policy makers on educational policy topics, because they are probably more interested and are better informed about the matter. Indeed, research on nonresponse repeatedly found less educated groups to be prone to nonresponse (Koch, 1998; Watson & Wooden, 2009). However, research on nonresponse in skills and educational studies report mixed results. Whereas Kleinert et al. (2013) could confirm, for an educational study conducted in Germany - "Arbeiten und

Lernen im Wandel" (ALWA), that groups with a low level of education were underrepresented, Van de Kerckhove et al. (2009) did not find a significant relationship between response status and the level of education in the US-American section of the Adult Literacy and Lifeskills Survey (ALL) in 2003. Furthermore, Darcovich et al. (1998) found, for the International Adult Literacy Survey (IALS) conducted in the US in 1994, the highest response rates among both the lowest and highest educational groups.

The question of whether specific survey design features might even introduce bias into the sample data has predominantly been discussed with reference to incentives (e.g., Singer & Ye, 2013). For example, rational choice approaches would assume that low-income groups and groups with correlated characteristics, such as a low level of education, consider incentives as a greater benefit, compared to high-income groups. However, research has found few consistent effects of incentives on the sample composition. Although a relationship between income and the attractiveness of incentives was verified by several studies (Juster & Suzman, 1995; Singer & Kulka, 2002), results regarding the level of education are mixed. Indeed, both Petrolia and Bhattacharjee (2009) and Berlin et al. (1992) reported that incentives were particularly successful in attracting less educated respondents, but Jäckle and Lynn (2008), found no such effect when studying different achievement groups among 16-17-year-old students. Analyses of an incentive experiment conducted during the PIAAC field test indicate that the payment of an incentive of 50€ in the German PIAAC main study may have had some effect on the sample composition. In the field test, the 50€-incentive was more successful in attracting the 16-25-year-olds, German citizens, and persons living in small and medium-sized towns than the other incentives (see Martin, Helmschrott & Rammstedt in this volume). The distributions of the level of education, gender and household size were not significantly different across the incentives (Pforr et al., forthcoming).

Another approach to explain nonresponse based on notions of rational choice is the application of social exchange theories to survey participation (Dillman, 1978; Dillman, Smyth, & Christian, 2009; Goyder, 1987). They assume that individuals are in constant social interaction with other individuals or institutions and expect long-term rewards from those relationships if they are equal with respect to favors (Blau, 1964). In the survey context, this concept is particularly useful when applied to governmental surveys. These are characterized by a special relationship between the sample person and the survey sponsor, including mutual rights and duties. When confronted with such a survey request, the reaction of the sample persons is supposed to depend on their past and expected future relationship with the governmental institutions, for example, with respect to government services (Groves & Couper, 1998).

Because government services vary between socio-economic groups, indicators reflecting the sample persons' socio-economic status (SES) have been used

to test the theory. However, two opposing hypotheses have developed: The first hypothesis suggests a negative linear relationship between survey participation and SES. According to the hypothesis, low SES groups feel more bound to participate as a sort of repayment, because they might receive governmental benefits, while high SES groups do not feel this obligation, because they pay more than they receive. The second hypothesis suggests a curvilinear relationship, with both the low and the high SES groups refraining from participation. The suggested explanation is that the low SES groups constantly feel unjustly disadvantaged in society, and survey interviewers – as agents of the more fortunate – might evoke memories of their disadvantages (Groves & Couper, 1998).

However, research has failed to find consistent support for either of these hypotheses. For example, Groves and Couper (1998) found support for greater cooperation among lower SES groups, as proposed by the first hypothesis, whereas Durrant and Steele (2009) and Demarest et al. (2012) found lower participation rates among households with a low SES. By contrast, Smith (1983) reported that middle SES groups were more likely to refuse than low or high SES categories. These inconsistencies might be explained by variations in the operationalization of socio-economic status. While some studies rely only on a single indicator for SES, such as income (Smith, 1983) or education (Demarest et al., 2012), others use combinations of various proxy indicators (Groves & Couper, 1998). Furthermore, it should be difficult to find consistent effects of SES on survey participation in studies conducted in different countries, because the type and magnitude of duties towards the government, such as taxes and government services, vary considerably across countries.

The application of this approach to PIAAC might be limited by the fact that it is not a "government survey", in its strictest sense. Even though the study was funded by two federal ministries, fieldwork has been conducted by a commercial survey organization. Thus, the interviewers might not have been perceived as agents of the government. However, as suggested by the social exchange theory, mentioning the survey sponsors might evoke memories of past exchanges with the government, in its broadest sense, and encourage participation as a reciprocal act for any kind of received benefits.

3.2 Social Isolation Theories

Social isolation theories are closely related to social exchange theories. They suggests that individuals or groups with a long history of negative exchange experiences with society feel socially isolated. The repeated frustration of such groups, e.g. due to unequal treatment, leads to the deliberate denial of mainstream societal norms. In the context of survey participation, this could restrain potential participants from seeing their participation as their "civic duty" (Groves & Couper, 1998).

In the literature, the theory has been tested extensively with socio-demographic proxy indicators for social isolation. For example, it is conceivable that persons living in a single-person household show lower response rates, due to less social integration, whereas households with children have higher response rates because they are highly integrated into the community, e.g., through school networks. The theory also implies that sample persons living in large, multiunit structures are less inclined to participate, due to weaker ties with neighbors and the local community (Groves & Couper, 1998). Indeed, a multitude of studies found lower response propensities for single-person households (e.g., Ekholm & Laaksonen; Groves & Couper, 1998; Smith, 1983) and households in multiunit structures (e.g., Goyder, Lock, & McNair, 1992; Groves & Couper, 1998), while there is consistent proof of higher response rates from households with children (e.g., Groves & Couper, 1998).

Furthermore, the theory suggests that immigrants and ethnic minorities have lower response rates than native citizens or the ethnic majority group. Prior research has found that these groups are less likely to be respondents (Blohm & Diehl, 2001; Feskens, Hox, Lensvelt-Mulders, & Schmeets, 2007). However, some studies failed to find differences (e.g., DeMaio, 1980; Smith, 1983) or reported above-average response rates among minority groups (Groves & Couper, 1998). This might be due to the fact that both of these groups are very different across and within countries and may thus show large variations in response behavior. In addition, lower participation rates by the elderly have been explained by their stronger disengagement from society, compared to younger age groups (Krause, 1993). However, results are inconsistent, with other studies finding either higher response rates for the elderly (Groves & Couper, 1998) or no age effect (Nicoletti & Peracchi, 2003). Regarding gender, some researchers claim that men are less likely to participate in surveys than women, because women more often take over the role of maintaining social interaction with friends, relatives, or neighbors (Groves & Couper, 1998). Also here, results are mixed, with most studies reporting higher response rates for women or failing to find a gender effect (e.g., Brehm 1993; Smith, 1983).

3.3 Further Factors Influencing Survey Participation

The theories presented above introduced useful notions about the mechanisms underlying the decision to participate in a survey and specific factors influencing the propensity to respond to a survey. However, with their respective focus, they fail to fully grasp the complexity of the survey participation process and its various influence levels.

First of all, they focus on that stage of the survey process at which the interviewer is already in contact with the sample person. However, as outlined by Groves and Couper (1998), nonresponse can already arise at an earlier stage: when the interviewer tries to locate or contact the sample person. The success of establishing

contact depends on the variability of the interviewers' contact attempts throughout the day and the week and on the at-home patterns of the sample persons. This implies that the reason for the low participation rates of some groups is that they spend little time at home. Research has repeatedly found that persons in employment and younger respondents (Lynn, 2003), single-person households, big-city dwellers, high-income, and well-educated groups are more difficult to contact than the elderly or households with children (Durrant & Steele, 2009; Goyder, 1987). It has also been assumed that women might, overall, be met more often at home than men. They more often take care of young children without holding a paid job, in comparison to men, or only have a teleworking or part-time job (Groves & Couper, 1998). In addition, research has found that lower participation rates of immigrants can be largely explained by their low contact rates (Koch, 1997). Reasons for this could be that immigrants spend prolonged time periods in their home countries (Blohm & Diehl, 2001) or that they are more likely to live in urban areas where contact difficulties are more pronounced (Feskens et al., 2007).

Moreover, an important reason for nonresponse at the cooperation stage that has not been reflected by the theories is the inability to participate in a survey. For example, immigrants and ethnic minority groups may simply not be able to participate because they do not speak the survey language and no interpreter is provided by the survey organization. Moreover, persons with a disability or health problems might not participate because their physical or mental problems impede them from understanding, reading, or correctly answering the survey questions (Groves, 2009; Stoop, 2005).

In addition, the theories focus on reasons related to the sample persons and their reactions to specific survey design features. However, Groves and Couper (1998) stress that the survey process is more complex and additional factors play a role in the decision to participate or not. One such factor is the social environment, which may be negatively influenced when privacy concerns are widely shared in society, or when citizens are often confronted with survey requests ("over-surveying effect") (Groves & Couper, 1998). Another environmental factor that has consistently proven to be related to survey participation is urbanicity. Residents of small towns and rural areas are generally more likely to be respondents, whereas big-city dwellers are usually both less cooperative and harder to contact (Blom, 2012; Stoop, Billiet, Koch, & Fitzgerald, 2010). Furthermore, Groves and Couper (1998) stress that the interviewers play an important role in gaining both contact and cooperation in interviewer-administered surveys, which has been widely acknowledged in the literature (Blom et al., 2011; Jäckle, Lynn, Sinibaldi, & Tipping, 2013; Pickery & Loosveldt, 2004).

3.4 Hypotheses on Factors Influencing Survey Participation in PIAAC Germany

Based on the literature review above, we derive the following hypotheses on survey participation in PIAAC Germany.

Regarding *age*, we assume that the youngest age group was most likely to respond (Hypothesis 1). Even though, in empirical studies, young respondents have been found to be difficult to contact, the incentive experiment conducted during the German field test showed that the 50€-incentive was particularly successful in attracting the 16-25-year-olds. Hypotheses proposed by the theories for other age groups seem to be hardly applicable to PIAAC. For example, theories on social isolation suggest lower response rates among the elderly because of their disengagement from society. Rational choice approaches propose that they feared a higher cognitive burden from the skills assessment than younger sample persons. In addition, they are supposed to be more likely to suffer from a reading and/or writing difficulty or an impairment.

However, the oldest age group in PIAAC comprises the 55-65-year-olds, who are generally still active members of society and in good health. Hence, we expect their willingness to respond to be similar to that of other age groups. Furthermore, we expect *women* to show higher response rates than men (Hypothesis 2), largely because empirical studies found them to be more often met at home than men.

Several reasons make us expect that *non-Germans* were less likely to participate than Germans (Hypothesis 3). First, they have been shown to be difficult to contact. Second, according to social isolation theories, as non-citizens, they could have felt less obliged to contribute to a study that is useful for the German society. Third, rational choice approaches would assume that non-Germans who are not proficient in German might have refrained from participation because they feared higher cognitive and time burdens than Germans. This is due to the fact that the skills assessment was conducted solely in German; an interpreter could be used only for the completion of the background questionnaire. Fourth, those non-Germans without German language skills might not have seen the benefit of completing a questionnaire without being able to participate in the skill assessment, and thus refrained from participation. Finally, the incentive experiment of the German field test has shown that the 50€-incentive was less attractive for non-Germans than for Germans.

Furthermore, we expect that persons with *lower levels of education* were less willing to participate in PIAAC than those with a high level of education (Hypothesis 4). Rational choice approaches suggest that those with lower educational attainment feared higher costs, in the form of cognitive survey burden, due to the need to complete a skills assessment. Those with a high level of education might have been more interested in the topic and more curious about completing a skills assessment.

Moreover, they probably expected lower costs from the cognitive survey burden and a higher personal benefit from being part of a study whose results serve policy makers.

Regarding *urbanicity*, we expect big-city dwellers to have lower response rates than those living in smaller cities (Hypothesis 5). This might be related to the hypothesis of social isolation theories that big-city dwellers live more anonymous lives and avoid contact with strangers, and also because they are less likely to be reached at home, due to busy life-styles. Furthermore, the results of the incentive experiment of the German field test indicate that the 50€-incentive was more successful in convincing residents of smaller and medium-sized cities to participate. Closely related to urbanicity, we also expect that sample persons living in large *multiunit houses* were less likely to participate (Hypothesis 6), in keeping with the social isolation theories.

Additionally, we assume that persons with a low *socio-economic status*[2] have lower response rates (Hypothesis 7). The curvilinear hypothesis of the social exchange theories and the social isolation theories would suggest that this is due to a reduced feeling of civic obligation to contribute to a research project benefitting society. Even though the curvilinear hypothesis of the social exchange theory would also predict low response rates for groups with high socio-economic status, we expect that these groups were more likely to respond. Those with a high socio-economic status tend to have a high level of education; as outlined above, we expect the highly educated to be more inclined to participate.

Regarding the sample persons' *work status* (Hypothesis 8), we can derive two hypotheses from the theory. Notions of social isolation or social exchange theories suggest that the unemployed and those out of the labor force are less interested in participating in a study useful for a society they do not feel to be a part of. Furthermore, rational choice theory proposes that they might fear higher survey burdens by having their skills tested, because they could be afraid of having lower skills, compared to respondents holding a job. However, rational choice theory also assumes that those in employment and the self-employed are less likely to respond because they fear higher costs from the time burden imposed by a survey. Furthermore, they are probably more difficult to be contacted, because they are met at home less often.

With regard to *household size*, the theories predict that sample persons living in single-person households were less inclined to participate than persons living in multi-person households with children (Hypothesis 9). Social isolation theories assume that the former are more isolated from society than the latter and thus are less willing to contribute to a survey beneficial for society. Furthermore, single-person households are more difficult to contact.

---

2 To test this hypothesis, we use the variables "socio-economic status", "condition of the house" and "purchasing power". The variable "socio-economic status" is a combination of the level of income and the level of education in the area the sample person lives in.

4 Nonresponse in PIAAC Germany

Following a description of the data and the analyses we used, we explore in this section, which of the described characteristics are the main factors influencing the decision to participate in PIAAC or not. Since the non-contact rate in PIAAC Germany was only 3.4% (Zabal et al., 2014), we focused on overall nonresponse, rather than explicitly distinguishing between non-contact and non-cooperation.

4.1 Data Description

To analyze nonresponse, auxiliary variables are needed that are available for both respondents and nonrespondents. As described in Zabal et al. (2014), the basic socio-demographic and geographic information we had at our disposal (age, gender, citizenship and municipality size) is part of the sample frames provided by the Federal Statistical Office and local population registries. Furthermore, interviewers were required to assess the sample persons' level of education and social class, type and condition of the house they lived in, and whether an intercom existed, and provide this information in their contact protocols. This evaluation had to be done prior to the first contact with the sample person. Finally, we used a commercial consumer-marketing database provided by Microm, which includes further socio-demographic and economic information on sample units at an area level. The data we used from this source are unemployment rate, socio-economic status (a combined variable of the level of education and income), purchasing power per household, and the prevailing family structure (i.e., the share of single households and households with children) in an area (Microm, 2011).

4.1.1 Quality of the data

Among these sources, the information provided by the sample frames is assumed to be of the highest quality. These data are regularly updated by the administrative authorities, are available at the individual level and rarely contain missing values. The contact protocol information also contained only few missing values. However, these data are prone to error, because interviewers were advised to collect them prior to their first contact with the sample persons, in order to make the data from respondents and nonrespondents comparable. This instruction might have had little effect on questions such as the type and condition of the house or whether an intercom existed (Sinibaldi, Durrant, & Kreuter, 2013). However, interviewers' evaluations of social class and level of education are potentially subject to measurement error, because they are based solely on environmental factors such as the neighborhood or features of the housing (Olson, 2013; West, 2013). We assessed the accuracy of the interviewers' judgments of the sample persons' level of educa-

tion; for this variable, we had comparable data available from the PIAAC interview. By calculating the percentage of correct estimations[3], we found that, overall, only approximately half of assessments were correct (48.4%). However, only very few interviewers gave a completely wrong assessment by assigning a low level of education when, in fact, the respondent had a high level of education, and vice versa (5.5%). We thus conclude that the interviewers' assessments of the respondents' level of education were reasonably accurate. However, the results have to be treated with caution, because comparable data were not available for nonrespondents and only the interviewer evaluations of the respondents' level of education could be verified.

Microm data also have quality limitations, because they are aggregated over an area comprising between five and approximately 500 households, with an average of about eight households (Microm, 2011). In addition, for about 5% of the sampled units, Microm data were not available (Zabal et al., 2014).

4.1.2 Definition of response status and sample size

Participants in PIAAC first had to complete a questionnaire collecting background information that was administered by the interviewers on a laptop computer. The questionnaire was followed by an assessment that respondents performed in the domains literacy, numeracy, or problem solving in technology-rich environments[4] (Zabal et al., 2014). We defined *respondents* as participants who completed the PIAAC background questionnaire or had answered a sufficient proportion of the questionnaire, as defined in OECD (2013). *Nonrespondents* were defined as sample persons who did not start the interview because they were, for example, not able to be contacted, refused, did not respond due to literacy-related reasons or due to a disability, or broke off the interview before reaching the designated threshold[5]. Literacy-related reasons are language problems, difficulties with reading or writ-

3 The interviewers had to assess whether the sample person's level of education was "low", "medium" or "high". For the comparison, the information on the respondent's ISCED level (International Standard Qualification of Education) collected during the interview was recoded as "low": below ISCED 1, ISCED 1 & 2, "medium": ISCED 3 & 4, and "high": ISCED 5 & 6.

4 The assessment comprised, generally, a combination of two of the domains mentioned. However, one sixth of respondents received only items in problem solving in technology-rich environments.

5 There were only three breakoffs in the PIAAC background questionnaire. Two cases were counted as respondents, one as nonrespondent.

ing and a learning or mental disability (Zabal et al., 2014). Ineligible cases were excluded.[6]

The gross sample in PIAAC Germany comprised n = 10,240 individuals, out of which n = 10,086 were eligible. According to the definition outlined above, n = 5,379 sample persons were counted as respondents, and n = 4,707 as nonrespondents.

4.1.3 Weights used for analyses and variance estimation

For all analyses presented, the PIAAC unknown eligibility weight was used. This is a base weight correcting for differential selection probabilities that occurred because of an erroneous selection algorithm used during sample selection in PIAAC Germany (Zabal et al. 2014). Moreover, this base weight adjusts for unknown eligibility: Those whose eligibility could not be verified, e.g., because they had moved and their new address could not be traced, were weighted down according to the proportion of ineligibles among those with known eligibility. In order to account for an increased variance due to the complex sample design, for each of the weights used in PIAAC, 80 replicate weights had been calculated by the international consortium (OECD, 2013). For the correct estimation of variance, the unknown eligibility weight was thus used, together with its 80 replicate weights.[7]

4.2 Main Factors Potentially Introducing Nonresponse Bias in PIAAC Germany

In this section, we examine which of the socio-demographic, economic, and geographic characteristics suggested by the literature are the main factors influencing survey participation in PIAAC Germany and test whether our hypotheses could be verified.

As outlined above, only those factors that are both related to the central study outcome(s) and response status have the potential to introduce bias into the data set (Groves, 2006). Thus, we first examined, at the bivariate level, whether the charac-

6 This definition differs slightly from the one used in the official PIAAC nonresponse bias analyses. In this paper, the literacy-related nonrespondents are coded as nonrespondents, because the inability to participate due to literacy-related reasons is considered as an important reason for nonresponse. Due to technical reasons related to the weighting process, the literacy-related nonrespondents were excluded from the analyses for the official PIAAC nonresponse bias analyses. The results presented here are thus not directly comparable to the results of similar analyses published in Zabal et al. (2014).

7 Weights exceeding $3.5^* \sqrt{1+CV^2}$ the median unknown eligibility weight were trimmed by the authors, in line with the trimming procedure for the PIAAC final weights (see OECD, 2013).

teristics frequently identified as drivers of nonresponse were significantly associated with both response status and the central study outcome, which is *proficiency*[8] in PIAAC. The variables not significantly related to proficiency and response status are irrelevant for nonresponse bias in PIAAC and were thus omitted from further analyses.

We used proficiency in literacy (in the following called "proficiency") as the key study outcome, because literacy can be regarded as a basic skill that is highly relevant for the acquisition of the other skills measured in PIAAC (Zabal et al., 2013). For the correct estimation of variance, due to both the complex sample design and the imputed plausible values, the "PIAACTOOLS"[9] that have been developed for Stata were used. Because these tools do not include Pearson's r correlation analyses, we ran linear regression models with proficiency as dependent variable and each variable investigated as individual predictor. The Pearson's r values were obtained by calculating the radical of the coefficient of determination of the regression models.

4.2.1 Factors with the potential to introduce nonresponse bias into the data set

As can be seen in Table 1, most explanatory variables were highly significantly correlated with proficiency at the 0.1% level. The strongest correlations were observed for the level of education and social class (both $r = 0.3$, $p = 0.000$), followed by age, citizenship, the condition of the house, socio-economic status, and purchasing power ($r = 0.2$, $p = 0.000$). The unemployment rate and the type of house (both $r = 0.1$, $p = 0.000$), gender, municipality size, and the family structure in the area (all $r < 0.1$, $p < 0.05$) showed the lowest correlations. Given that the correlation coefficients are only of a low to medium strength, the potential for bias in the proficiency score is only moderate. The presence of an intercom at the sample persons' houses showed no significant correlation with proficiency and was thus omitted from the logistic regression analysis below.

Results of the χ^2-tests of independence[10] between the explanatory variables and response status revealed that nearly all characteristics were significantly related

8 The proficiency scales in PIAAC have been modelled for each of the skill domains, based on Item Response Theory (IRT). This reflects both the difficulty of the task and the respondents' skill level on one scale. The scales range from 0-500; the higher the value on the scale, the higher the skill level needed to solve a task. For each respondent, 10 "plausible values" were estimated per scale, in order to improve the accuracy of the proficiency estimates for the subpopulations and the overall population (Zabal et al., 2013).

9 http://www.oecd.org/site/piaac/PIACTOOLS_16OCT_for_web.pdf (retrieved November 2014)

10 In order to account for the complex sample design, the Pearson's χ^2-statistic was corrected with the second-order correction of Rao and Scott (1981) and converted into an F-statistic.

Table 1 Associations of explanatory variables with proficiency and response status

	Proficiency		Response status	
Explanatory variable	Pearson's *r*	*p-value*	*F***	*p-value*
Registry				
Age	0.2	0.000	21.7	0.000
Gender	<0.1*	0.007	3.8	0.056
Citizenship	0.2	0.000	48.4	0.000
Municipality size	<0.1	0.02	5.7	0.000
Contact protocol				
Level of education	0.3	0.000	30.7	0.000
Social class	0.3	0.000	20.5	0.000
Intercom	<0.1	0.302	-	-
Type of house	0.1	0.000	27.9	0.000
Condition of house	0.2	0.000	16.2	0.000
Microm				
Socio-economic status in area	0.2	0.000	8.5	0.000
Unemployment in area	0.1	0.000	3.6	0.030
Purchasing power in area	0.2	0.000	6.2	0.003
Family structure in area	<0.1	0.024	22.1	0.000

* The output of the PIAACTOOL regression displays the R^2 only with two decimals after the point. In case the value of the displayed R^2 is 0.00, an exact result for *r* cannot be calculated.

** See footnote No. 10.

to response status; most at the 0.1% level. Gender was not significantly related to response ($F = 3.78$, $p = 0.056$). However, because the 5% level of significance was only marginally missed and, in the multivariate setting, this covariate could be more significant, it was included in the regression analysis.

4.2.2 Main influencing factors on participation in PIAAC Germany

In this section, we analyze which of the characteristics significantly associated with proficiency and response status at the bivariate level had an effect on participation in PIAAC when controlled for by other covariates in logistic regression analyses predicting response. These characteristics are identified as the main factors influencing response to PIAAC in Germany. Results of the analyses serve to test whether our hypotheses on nonresponse in PIAAC Germany can be verified.

First, a full model was estimated that included all factors that had been shown to have the potential to introduce bias into the PIAAC data, with the exception of the dummy variables for social class. In an analysis of multicollinearity, they showed a high variance inflation factor, indicating that its inclusion might bias the results (low social class: VIF = 6.93, tolerance = 0.14, middle social class: VIF = 5.16, tolerance = 0.19). Subsequently, those variables without a significant contribution in the full model were removed and a final model was fitted.

As displayed in Table 2, results from the first full model showed that, when controlling for other factors, only age, citizenship, the level of education, the type of house, residence in a metropolitan area (500,000 inhabitants and more), and a high unemployment rate in the area had a significant influence on survey participation in PIAAC Germany. By contrast, gender, the condition of the house, the predominant socio-economic status, purchasing power and household size in the area the sample person lives in, proved not to be significant predictors of response. However, a goodness-of-fit test indicated a lack of fit of the full model ($p = 0.044$)[11]. We thus removed the insignificant covariates to estimate a final model that has an improved model fit ($p = 0.104$). In this final model, we see that the unemployment rate no longer had a significant influence on response, whereas the remaining effects were similar to those in the full model. Age, citizenship, a low level of education and the type of house the sample persons live in were highly significant predictors of response at the 0.1% level, and having a medium level of education and living in a metropolitan area were significant at the 5% level. Living in a smaller or medium-sized city did not have a significant effect.

The results of the multivariate analyses indicate that only some of our hypotheses on nonresponse to PIAAC in Germany were substantiated. Even though gender (Hypothesis 2), the predominant socio-economic status, the condition of the house and purchasing power (all Hypothesis 7), the unemployment rate (Hypothesis 8) and the household size in the area the sample person lives in (Hypothesis 9) were significantly related to response status at the bivariate level, in the multivariate setting they proved not to be significant predictors of response to PIAAC. However, a closer look at the results of the final model reveals that our hypotheses on age, citizenship, the level of education and urbanicity could be verified. As expected, the 16-25-year-olds were distinctly more likely to participate than the other age groups (Hypothesis 1). Moreover, non-Germans (Hypothesis 3), persons with lower levels of education (Hypothesis 4), big-city-dwellers (Hypothesis 5) and those living in larger housing units (Hypothesis 6) were less likely to participate than their respective counterparts.

11 As goodness-of-fit test, the F-adjusted mean residual test was used, which takes the complex sample design into account. A small p-value indicates a lack of fit (For details of the method, see Archer & Lemeshow, 2006).

Table 2 Logistic regression models predicting response

	Full model		Final model	
Variable	Coefficient	S.E.	Coefficient	S.E.
Age (*Reference = 16-25*)				
26-35	-0.460***	(0.085)	-0.466***	(0.082)
36-45	-0.526***	(0.081)	-0.532***	(0.079)
46-55	-0.631***	(0.068)	-0.613***	(0.068)
56-65	-0.633***	(0.096)	-0.645***	(0.093)
Gender (*Reference = Female*)				
Male	-0.063	(0.053)		
Citizenship (*Reference = German*)				
Non German	-0.368***	(0.096)	-0.379***	(0.094)
Level of education (*Reference = High level of education*)				
Low level of education	-0.332***	(0.084)	-0.389***	(0.068)
Medium level of education	-0.166*	(0.070)	-0.173*	(0.065)
Type of House (*Reference = Farmhouses, single and terrace houses*)				
House with three to eight flats	-0.193**	(0.061)	-0.225***	(0.057)
Houses with 9 flats and more	-0.337***	(0.072)	-0.379***	(0.068)
Municipality size (*Reference = 1-4,999 inhabitants*)				
5,000-49,999 inhabitants	0.026	(0.080)	-0.027	(0.076)
50,0000-499,999 inhabitants	-0.065	(0.091)	-0.055	(0.084)
500,000-99,999,999 inhabitants	-0.204*	(0.097)	-0.190*	(0.087)
Condition of the house (*Reference = Very good condition of the house*)				
Bad condition	-0.097	(0.090)		
Good condition	-0.011	(0.070)		
Unemployment rate (*Reference = below average unemployment rate*)				
Average unemployment rate	0.059	(0.065)	0.016	(0.059)
Above average unemployment rate	0.168*	(0.075)	0.059	(0.057)
Purchasing power per household (*continuous variable*)				
Purchasing Power	0.001	(0.002)		
Socio-economic status (*Reference = Above average status*)				
Below average status	-0.072	(0.091)		
Average status	0.006	(0.067)		
Family structure (*Reference = above average share of families with children*)				
Above average share of single HH	-0.061	(0.088)		
Mixed family structure	-0.078	(0.059)		

Table 2 Logistic regression models predicting response (cont.)

	Full model		Final model	
Variable	Coefficient	S.E.	Coefficient	S.E.
Constant	0.912**	(0.278)	0.987***	(0.105)
N	9367		9832	
Nb. of Replicates	80		80	
Design df	79		79	
Prob > F	0.000		0.000	
P-value of the *F*-adjusted mean residual test	0.044		0.104	

Dependent variable: 1 = response 0 = nonresponse
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

5 Discussion

In multivariate analyses, we found that non-Germans, those with lower levels of education, those living in larger housing units, and big-city dwellers were significantly less likely to participate in PIAAC than their respective counterparts. Furthermore, we found that 16-25-year-olds were significantly more willing to take part in PIAAC than other age groups. Age, citizenship, the level of education, the type of house the sample person lives in, and municipality size can therefore be identified as main factors influencing participation in PIAAC Germany. However, given that the correlation coefficients of these variables with the central study outcome proficiency are only of weak to medium strength (r = 0.1-0.3), the potential for nonresponse bias in the data set is only moderate.

Our hypotheses about foreigners, big-city dwellers and those living in larger housing units being less likely to participate in PIAAC have been validated and thus confirm corresponding findings in the existing literature (DeMaio 1980; Feskens et al. 2007; Goyder et al., 1992; Groves & Couper, 1998). Theoretical approaches to survey participation, such as hypotheses on social isolation, suggest that these groups feel isolated from either their local communities or from society as a whole and thus lack the feeling of a "civic duty" to participate in surveys useful for society. In addition, a multitude of studies has shown that these groups are difficult to contact (Durrant & Steele, 2009; Koch, 1997). Furthermore, because the PIAAC skills assessment was conducted in German, we suppose that, among non-German citizens with little or no German language skills, the higher cognitive burden related to this assessment, or the inability to complete it, impeded participation.

Our expectation that those with lower levels of education were less willing to participate was met, too. This indicates that rational choice approaches, which sug-

gest these groups might fear higher survey burdens from the skills assessment and might be less interested in participating in an educational study, have good explanatory power to justify the reluctance of these groups to participate. Furthermore, we can confirm empirical studies that reported a similar effect of the level of education on response (Kleinert et al., 2013; Koch, 1998; Watson & Wooden, 2009).

Our hypotheses regarding gender, socio-economic status, work status, and the household size could not be verified because they proved not to be significantly related to response status, when controlled for by other covariates in the multivariate setting. However, because they are all significantly related to survey participation at the bivariate level, they could still be valid when tested separately.

Thanks to the rich information from three data sources, we were able to test which of the bivariately significant factors were the strongest predictors of participation at the multivariate level. However, it should be noted that the three data sources used are of different quality. Most of the variables that proved not to have a significant independent effect on survey participation contain information aggregated at an area level. These variables might not accurately describe the situation of all persons in the sample and they are thus weaker predictors of survey response than individual level data. Moreover, the information on the sample persons' level of education is prone to measurement error. The evaluation had to be performed prior to the first contact with the sample person and interviewers had to base their evaluation on neighborhood or housing characteristics. Even though we have demonstrated that the assessments of the level of education were reasonably accurate, a certain degree of error still remains.

6 Conclusion

The analyses presented in this paper aimed at identifying the main factors influencing survey participation in PIAAC Germany. Although a multitude of influence levels exists, we focused on socio-demographic, economic, and geographic characteristics of the sample persons. Because only few publications on nonresponse in skills and education studies exist, to date, this work yields valuable insights for researchers in this field when addressing nonresponse at different stages of the survey process.

In our analyses, we identified age, citizenship, the level of education, the type of the house the person lives, and municipality size as the main factors influencing participation in PIAAC Germany. We established that non-Germans, persons with lower levels of education, those living in larger housing units, and residents of metropolitan areas were less likely to participate.

These results indicate that skills and educational survey researchers can most effectively address nonresponse bias if they concentrate on these central factors.

In particular, the reluctance of those with the lowest level of education should be taken seriously, because this group can be expected to behave very differently with respect to educational topics, such as skills assessments or knowledge tests. This problem could be minimized by, for example, specifically addressing this group in tailored advance letters that might reduce potential fears about a test situation. The low participation of foreigners could be addressed by providing both the questionnaires and tests in the most common minority languages (Blohm & Diehl, 2001). Obviously, the usefulness and feasibility of the suggested measures depend on design features, such as the goals of the study or the study sample. For example, in PIAAC, a deliberate decision was made to conduct the skills assessment only in the official country language(s) or only in those languages of groups representing an important share of the population. This is due to the fact that the aim of the study was to measure skills that are needed for successful participation in the national society, which, in general, include speaking the country's language. Furthermore, the translation of tests and questionnaires or the use of interpreters for the questionnaires is costly. In countries without official information on the sample persons' level of education and citizenship, it will also be difficult to identify the relevant sample persons for targeted measures such as tailored advance letters.

Our analyses focused on overall nonresponse; the possibility of nonresponse due to contact difficulties was discussed only at the theoretical level. In addition, our analyses comprised only a selection of the various factors potentially influencing nonresponse. Future research could thus yield further valuable insights for the reduction of nonresponse in skills and educational studies by distinguishing between noncontacts and noncooperation and exploring the effects of other sources of influence, such as the interaction of interviewers with the sample persons or the countries' survey climates.

References

Archer, K. J., & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal 6*(1), 97-105.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I. S., Rock, D., & Yamamoto, K. (1992). *An experiment in monetary incentives.* In: American Statistical Association, Proceedings of the Survey Research Methods Section 1992 (pp. 393-398). Retrieved September 2014 from http://www.amstat.org/sections/srms/Proceedings/allyearsf.html

Bethlehem, J. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 275-287). New York: Wiley.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. Hoboken, NJ: Wiley.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, NJ: Wiley.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley

Blohm, M., & Diehl, C. (2001). Wenn Migranten Migranten befragen. Zum Teilnahmeverhalten von Einwanderern bei Bevölkerungsbefragungen. *Zeitschrift für Soziologie, 30*(3), 223-424.

Blom, A. G. (2012). Explaining cross-country differences in survey contact rates: Application of decomposition methods. *Journal of the Royal Statistical Society: Series A, 175*(1), 217-242.

Blom, A. G., de Leeuw, E. D., & Hox, J. J. (2011). Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics, 27*(2), 359-377.

Brehm, J. (1993). *The phantom respondents: Opinion surveys and political representations*. Ann Arbour: University of Michigan Press.

Darcovich, N., Binkley, M., Cohen, J., Myrberg, M., & Persson, S. (1998). Non-response bias. In T. S. Murray, I. S. Kirsch & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey* (NCES 98-053, pp. 55-71). Washington D.C.: U.S. Department of Education.

de Leeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York: Wiley.

DeMaio, T. J. (1980). Refusals: Who, where and why? *Public Opinion Quarterly, 44*, 223-233.

Demarest, S., Van der Heyden, J., Charafeddine, R., Tafforeau, J., Van Oyen, H., & Van Hal, G. (2012). Socio-economic differences in participation of households in a Belgian National Health Survey. *European Journal of Public Health, 23*(6), 981-985.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002): Survey nonresponse in design, data collection, and analysis. In: Groves, R. M, Dillman, D. A., Eltinge, J. L., & Little, R. J. A (Eds): *Survey Nonresponse* (pp. 3-26). New York: Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail and mixed-mode surveys: The tailored design method*. Hoboken, N. J.: Wiley.

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and noncontact in household surveys: Evidence from six UK government surveys. *Journal of the Royal Statistical Society: Series A, 172*(2), 361-381.

Ekholm, A., & Laaksonen, S. (1991). Weighting via response modeling in the finnish household budget survey. *Journal of Official Statistics, 7*(3), 325-337.

Feskens, R., Hox, J. J., Lensvelt-Mulders, G., & Schmeets, H. (2007). Nonresponse among ethnic minorities: A multivariate analysis. *Journal of Official Statistics, 23*(3), 387-408.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Cambridge: Polity Press.

Goyder, J., Lock, J., & McNair, T. (1992). Urbanization effects on survey nonresponse: A test within and across cities. *Quality and Quantity, 26*, 39-48.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646-675.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Hoboken, NJ: Wiley.

Groves, R. M., & Lyberg, L. (2011). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849-879.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quaterly, 68*(1), 2-31.

Groves, R. M., Singer, E., & Corning, A. D. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly, 64*(3), 299-308.

Jäckle, A., Lynn, P., Sinibaldi, J., & Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods, 7*(1), 1-15.

Juster, F. T., & Suzman, R. (1995). An overview of the health and retirement study. *Journal of Human Resources, 30*(5), 7-56.

Kleinert, C., Ruland, M., & Trahms, A. (2013). *Bias in einem komplexen Surveydesign: Ausfallprozesse und Selektivität in der IAB-Befragung ALWA*. FDZ Methodenreport (Vol. 02/2013). Nürnberg: IAB.

Koch, A. (1997). Teilnahmeverhalten beim Allbus 1994. Soziodemographische Determinanten von Erreichbarkeit, Befragungsfähigkeit und Kooperationsbereitschaft. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 49*(1), 99-122.

Koch, A. (1998). Wenn „mehr" nicht gleichbedeutend mit „besser" ist: Ausschöpfungsquoten und Stichprobenverzerrungen in allgemeinen Bevölkerungsumfragen. *ZUMA-Nachrichten, 42*, 66-90.

Krause, N. (1993). Neighborhood deterioration and social isolation in later life. *International Journal of Aging and Human Development, 36*(1), 9-38.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Lyberg, L., & Dean, P. (1992, May). *Methods for reducing nonresponse rates: A review*. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg.

Lynn, P. (1996). Weighting for non-response. In R. Banks, J. Fairgrieve, L. Gerrard, T. Orchard, C. Payne, & A. Westlake (Eds.), *Survey and Statistical Computing 1996* (pp. 205-214). Chesham: Association for Statistical Computing.

Lynn, P. (2003). PEDAKSI: Methodology for collecting data about survey non-respondents. *Quality and Quantity, 37*(3), 239-261.

Martin, S., Helmschrott, S., & Rammstedt, B. (2014). The use of respondent incentives in PIAAC: The field test experiment in Germany. *mehods, data, analyses*, 8(2), 223-242, doi: 10.12758/mda.2014.009

Microm. (2011). *Datenhandbuch. Arbeitsunterlagen für Microm MARKET & GEO*. Neuss.

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: Microlevel evidence from the european community household panel. *Journal of the Royal Statistical Society: Series A,168*(4), 763-781.

OECD. (2013). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD.

Olson, K. (2013). Paradata for nonresponse adjustment. *The annals of the American Academy of Political and Social Science, 645*(1), 142-170.

Petrolia, D. R., & Bhattacharjee, S. (2009). Revisiting incentive effects: Evidence from a aandom sample mail survey on consumer preferences for fuel ethanol. *Public Opinion Quarterly, 73*(3), 537-550.

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., . . . Rammstedt, B. (forthcoming). Are incentive effects on response rates and nonresponse bias in large-

scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly.*

Pickery, J., & Loosveldt, G. (2004). A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers. *Journal of Official Statistics, 20*(1), 77-89.

Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association (76*), 221-230.

Singer, E., & Kulka, R. A. (2002). Paying respondents for survey participation. In M. Ver Ploeg, R. A. Moffitt & C. F. Citro (Eds.), *Studies of welfare populations: Data collection and research issues* (pp. 105-128). Washington, D.C.: National Academy Press.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics, 15*(2), 217-230.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The annals of the American Academy of Political and Social Science, 645*(1), 112-141.

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly, 77*(S1), 173-193.

Smith, T. W. (1983). The Hidden 25 Percent: An analysis of nonresponse on the 1980 general social survey. *Public Opinion Quarterly, 47*(3), 386-404.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response. Lessons learned from the European Social Survey.* Chicester: Wiley.

Stoop, I. A. L. (2005). *The Hunt for the last respondent: Nonresponse in sample surveys*. The Hague: Social and Cultural Planning Office of the Netherlands.

Van de Kerckhove, W., Krenzke, T., & Mohadjer, L (2009). *Adult literacy and lifeskills survey (ALL) 2003: U.S. nonresponse bias analysis.* (NCES 2009-063). Washington D.C.: U.S. Department of Education.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 157-181). Chichester: Wiley.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the national survey of family growth. *Journal of the Royal Statistical Society: Series A, 176*(1), 211-225.

Zabal, A., Martin, S., Klaukien, A., Rammstedt, B., Baumert, J., & Klieme, E. (2013). Grundlegende Kompetenzen der erwachsenen Bevölkerung in Deutschland im internationalen Vergleich. In B. Rammstedt (Ed.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich – Ergebnisse von PIAAC 2012* (pp. 31-76). Münster: Waxmann.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical Report*. Münster: Waxmann.

methods, data, analyses | Vol. 8(2), 2014, pp. 267-280 DOI: 10.12758/mda.2014.011

A Simulation Approach to Estimate Inclusion Probabilities for PIAAC Germany

Siegfried Gabler, Sabine Häder & Jan-Philipp Kolb
GESIS – Leibniz Institute for the Social Sciences

Abstract

In PIAAC *(Programme for the International Assessment of Adult Competencies)* inclusion probabilities have to be known for every respondent at each sampling stage in all participating countries. However, in some cases it is not possible to calculate inclusion probabilities for a sample survey analytically – although the underlying design is probabilistic. In such cases, simulation studies can help to estimate inclusion probabilities and thus ensure that the necessary basis for the calculation of design weights is available. In this section, we present a Monte Carlo simulation using the German sample data. During the selection process for PIAAC Germany an error had occurred. Because of that, it was not possible to determine the inclusion probabilities analytically. Therefore a simulation study with 10,000 runs of the erroneous selection process was set up. As a result it was possible to compute the inclusion probabilities for the sample of PIAAC Germany.

Keywords: Monte Carlo simulation, inclusion probabilities, sampling for PIAAC Germany

1 Sampling for Comparative Surveys

Cross-national surveys have become very popular during the last decades. The reason for this is the multiplicity of questions that can be answered with the help of this kind of data. Lynn et al. (2006, p. 10) identify three main objectives for cross-national surveys, such as PIAAC (*Programme for the International Assessment of Adult Competencies)*:

a) Comparisons of estimates of parameters for different countries

b) Rankings of countries on different dimensions such as averages or totals

c) Estimates for a supra-national region such as the European Union aggregated from estimates of different countries.

Sampling strategies have to ensure the equivalence and/or combinability of these estimates. For this, both sample designs and estimation strategies have to be chosen carefully.

Kish (1994, p. 173) gives a theoretic basis for the application of sample designs in cross-cultural surveys:

"Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements."

Following this idea, an optimal sample design for cross-national surveys should consist of the best random sampling practice used in each participating country. The choice of a special sample design depends on the availability of frames, experience, but also mainly on the costs in different countries. Once the survey has been conducted, and adequate estimators have been chosen, the resulting values become comparable. To ensure this comparability, design weights have to be computed for each country. For this, the inclusion probabilities of every respondent at each stage of selection must be known and recorded. Furthermore, the inclusion probabilities for non-respondents must also be recorded at every stage where the necessary information is available to have possibilities for the compensation of the nonresponse (see Helmschrott/Martin in this volume) by suitable weighting procedures.

Direct correspondence to
Siegfried Gabler, GESIS – Leibniz Institute for the Social Sciences,
PO Box 12 21 55, 68072 Mannheim, Germany
E-mail: siegfried.gabler@gesis.org

Acknowledgment: We thank Silke Martin, GESIS, for her comments to a draft of this paper.

In the following section basic requirements for the PIAAC sampling are explained. For the German survey the sample design is described in detail. Furthermore, the erroneous procedure applied by the survey institute during the selection of the PIAAC gross sample is presented. Then, the simulation setup is demonstrated. The simulation results are evaluated in section 3. Finally, conclusions are drawn in the last section.

2 Basic Requirements and Sample Design Features of PIAAC Germany

Derived from the principles of sampling for cross-cultural surveys mentioned above the international PIAAC-Consortium expressed the following basic requirements for sample designs in the participating countries (OECD 2009, p. 6):

- Clustered and stratified designs were advised since these design features ensure both cost efficiency and variation of socio-demographic variables.
- A variety of designs could be applied because different countries have different access to frames and varying experience with the application of sample designs. Self -weighting designs of dwelling units or individuals should be preferred.
- All countries had to use probability based sampling methods at each stage of selection.
- The target population was defined as non-institutionalized adults between the ages of 16 and 65 (inclusive).

2.1 Sample Design and Sample Selection in Germany[1]

The sample design can be described as stratified two-stage probability design.

Stage 1

The PSUs (municipalities = Primary Sampling Units) were explicitly stratified by the variables federal states (Bundesländer), administrative regions (Regierungsbezirke), districts (Kreise) and ten grades of urbanization.

The sample points within the PSUs consisted of a pre-specified number of individuals to be selected at the second sampling stage from the person register held by the municipalities. In the vast majority of cases, sample points corresponded to one municipality only, while very large municipalities were drawn more than once and therefore covered more than one sample point. The number of sample points was

1 For a more detailed description of the PIAAC sampling procedure, see Zabal et al. (2014) as well as Lynn et al. (2014).

Table 1 Allocation of the sample sizes to municipalities

Number of inhabitants	Sample size
- 99,999	60
100,000 - 499,999	120
500,000 and more	180

set to 320. This resulted in the selection of 277 municipalities. In every municipality the sample spread over the whole area, i.e. there was no local clustering. Some larger municipalities had more than one sample point. If there were *k* sample points in a municipality the number of persons selected was multiplied by *k* (see table 1).

The PSUs were allocated proportionally to the size of the target population within each stratum. As only whole numbers can be selected as PSUs, the exact number of sample points to be selected from each stratum was determined using the procedure for unbiased controlled rounding by Cox (1987). This so-called Cox-Algorithm assures that the cell totals as well as the marginal totals of the allocation table remain nearly unchanged by the rounding procedure so that the structural properties of the population are not lost due to rounding (see Lynn et al., 2014).

Stage 2

To ensure an equal selection process in each selected municipality the following instructions were sent to the registration offices:

A simple systematic random sample of individuals, with a random start number and a sampling interval had to be drawn. The sample size in each municipality depended on its population size according to table 1. Personal information such as name, address, age, gender, nationality had to be provided for each selected individual by the registration offices. Data delivered by them were checked for different aspects. For more details see Zabal et al. (2014, pp 51).

All individuals (= person addresses) per point were allocated to a matrix defined by the variables age (six groups) and gender (see Sample Frame in figure 1). With an Iterative Proportional Fitting procedure (IPF) 32 individuals per sample point were selected from the frame under the constraint to meet the age and gender distribution in the federal state (for the result of the selection process, see Allocation Matrix in figure 1).

The selection of the individuals from the pool of addresses per community was done systematically with a selection interval. Unfortunately, in this process a programming as well as a sorting error did occur. The length of the interval was computed by “number of cases on the sampling frame” divided by “number of

Official Statistics		
Sex	Age Group	Freq[(1)]
m	16 – 19	256511
m	20 – 29	660824
m	**30 – 39**	**672291**
m	40 – 49	939744
m	50 – 59	727218
m	60 – 65	323953
f	16 – 19	243785
f	20 – 29	648787
f	30 – 39	668194
f	40 – 49	899595
f	50 – 59	726175
f	60 – 65	330558
total		7097635

Sample Frame		
Sex	Age Group	Point 163
m	16 – 19	2
m	20 – 29	5
m	**30 – 39**	**6**
m	40 – 49	5
m	50 – 59	3
m	60 – 65	2
f	16 – 19	0
f	20 – 29	11
f	30 – 39	3
f	40 – 49	5
f	50 – 59	6
f	60 – 65	2
total		50

Allocation Matrix		
Sex	Age Group	Point 163
M	16 – 19	1
M	20 – 29	3
m	**30 – 39**	**4**
M	40 – 49	3
M	50 – 59	2
M	60 – 65	1
F	16 – 19	0
F	20 – 29	7
f	30 – 39	2
f	40 – 49	3
f	50 – 59	5
f	60 – 65	1
total		32

(1) Source: Statistisches Bundesamt Genesis Table 12411-0012 at 31.12.2009

Sex	Age	selected
m	31	1
m	32	0
m	32	1
m	34	1
m	35	0
m	38	1

Figure 1 Example for the functionality of the optimization algorithm (variables sex and age)

cases to be selected" (see Allocation Matrix) and was not rounded. For the start number, a random number between 0 and the length of the interval was generated. If the start number was between 0 and 1.5, the program rounded always to 1. If the start number was at least 1.5, the program rounded to the closest integer number (based on commercial rounding). From a statistical point of view it would have been correct to always round up to the closest integer number.

The example from figure 1 illustrates the optimization algorithm. The adjustment algorithm always results in the same solution of number of elements to be selected in each cell (unrounded, exact allocation). In the example this value is equal to 3.97. The rounding of this number to the closest integer number is done randomly. 3.97 is rounded to 4 in 97% of the cases and to 3 in 3% of the cases. Due to the random procedure in the example, the exact number of persons to be selected from males, age 30 to 39 in sample point 163 was set to 4. Thus, the interval length is 6/4. In a next step the algorithm computed a random start number between 0 and 1.5, which was here 1.1. If the algorithm had worked correctly, 1.1 would have been rounded in some occasions to 1 and in other occasions to 2. However, due to

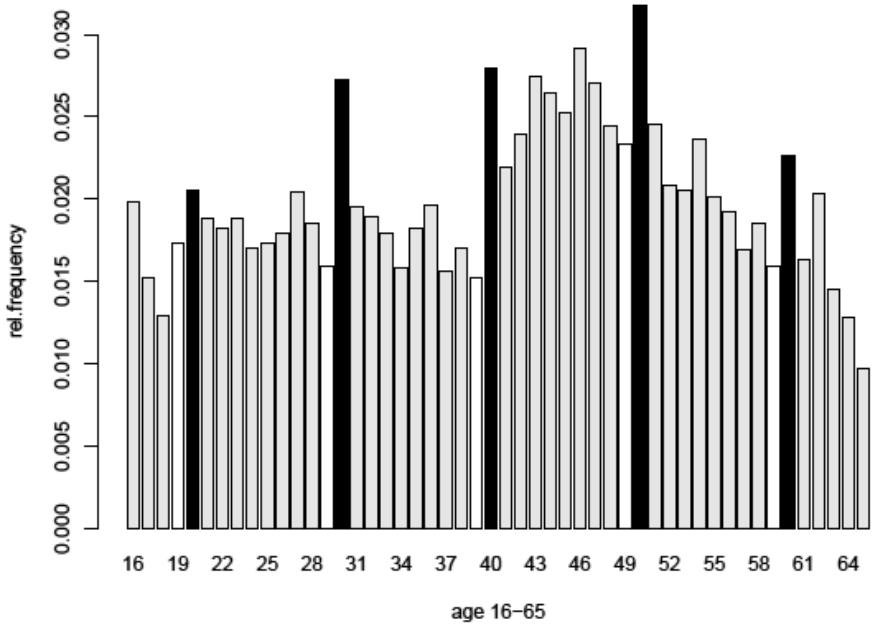


Figure 2 Age distribution (PIAAC sample unweighted) resulting from the erroneous algorithm

the error in the algorithm program, a random number of 1.1 would have always been rounded down to 1, and thus the chance for the first person on the frame to be selected was higher.

Summary of the selection process in the example
Number of cases due to IPF: 3.97 (rounded to 4)
Length of interval: 6/4 = 1.5
Start value: 1.1
Selected units unrounded: 1.1, 1.1+1.5, 1.1+1.5+1.5, 1.1+1.5+1.5+1.5
Selected units after commercial rounding: 1, 3, 4, 6

According to common practice of the survey institute, the pool of addresses on the sample frame is randomly ordered by the Fisher-Yates Shuffle before the sample is drawn. This procedure was done with the pool of addresses for the PIAAC sample as well. However, for some quality control checks the sample frame was sorted by age and this sorting order was unfortunately kept for the drawing. This mistake in accordance with the programming error (rounding error of the start number and thus higher chances of selection for the first person on the frame) both had a very negative impact (see Figure 2): Some age-groups (those ending with 0) are over-represented, others (in particular those ending with 9) are under-represented.

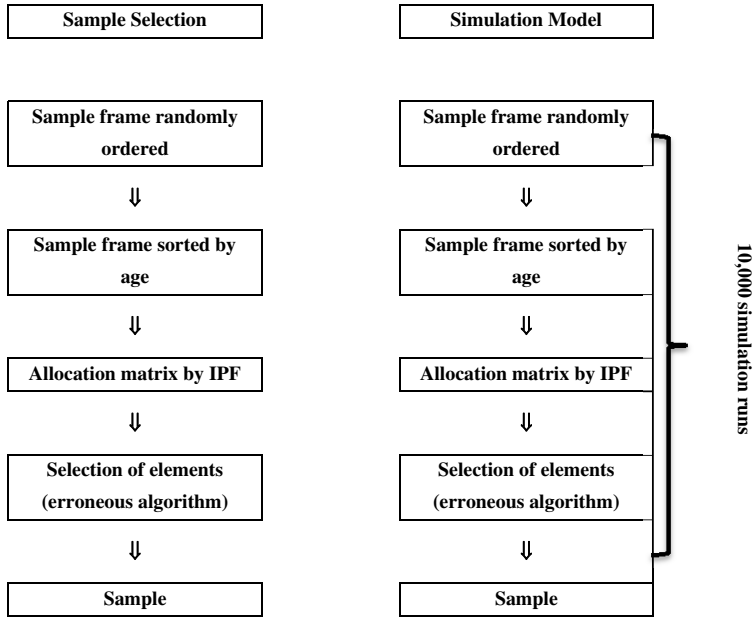


Figure 3 Comparison of the erroneous sample selection process and the simulation model

2.2 Simulation of the Selection Procedure

As a consequence, the gross sample has no longer the characteristic of equal selection probabilities for all elements. Instead, the selection probabilities for persons varied. Since it was too time consuming to model the incorrect selection probabilities, we decided to compute them through simulations, i.e. through a repetition of the erroneous optimization algorithm for 10,000 times. The idea was to rebuild the erroneous sampling procedure. Thus, the selection of the individuals from the pool of addresses was repeated 10,000 times. This was the basis of the simulation. The simulation model is described in figure 3.

In our model the random shuffle was repeated each time before a new iteration occurred – as it was done in the original optimization process. Thus, the following steps were repeated 10,000 times:

- The sample frame was randomly ordered according to the Fisher-Yates Shuffle.
- The sample frame was sorted by age.
- The sample was drawn.

Again, in order to estimate the selection probability of each element on the sample frame, a count was made of how many times an element was selected in each of the 10,000 samples.

The results of the simulation study are presented in the next section.

3 Evaluation of the PIAAC Sampling Procedure

To evaluate the results of the simulation – it was a Monte Carlo simulation – some theoretical considerations have to be explained first. The sample selection of the PIAAC sample is the result of a random experiment. For the PIAAC sample the random experiment to generate the PIAAC sample consists of several random experiments. If the random experiment would have been conducted as planned by the survey institute, the result would be that each person of the population would have the same selection probability. Due to the described error in the course of the random experiment, the equal probability is interfered, but not the general character of a probability sample as a result of a random experiment, i.e.

- that the random experiment could be repeated unlimited times, and
- that the results of the random experiment, i.e. possible samples, may be different, meaning that the result of the random experiment cannot be predicted with certainty for each iteration.

For the $r = 10{,}000$ simulation runs it was never the case that an element was not at all selected. Thus, it can be concluded that the selection probabilities are all positive. The error in the course of the random experiment affected only a part of the whole random experiment.

Ideally, as mentioned above, the selection of the PIAAC sample should have led to equal selection probabilities. This condition is no longer given due to the error in the course of the random experiment. The question is which selection probabilities have been generated by the selection process. Due to the error and the fact that the whole sampling procedure is built on random processes and sort sequences, it is very difficult and time-consuming for either GESIS or the survey institute to reflect this error in formulas in order to exactly calculate the selection probabilities.

The delivery deadline for the sample to the Consortium was dated shortly after the problem was noticed. The calculations needed to be carried out within a short time span. Furthermore, the amount of time that is necessary for one simulation run is not negligible, but cannot be determined exactly. It was thus necessary to find a trade-off between calculation time and an adequate number of simulation runs. The number of 10,000 simulation runs was the highest number which could be achieved under the prevailing circumstances. It was important that in every single simulation run the erroneous algorithm performed like in reality. Therefore, it was neither

possible nor justified to program the algorithm more effectively. If one regards the high sampling fraction, it is clear that a higher number was not necessary in this coherence.

However, the $r = 10{,}000$ samples generated by the simulation, provide an excellent basis for a sufficiently precise estimation of the true selection probabilities. This will be justified as follows:

We observe the event of selecting a person into a sample. The true selection probability given a single iteration of the random experiment is *P*. In *r* independent repetitions of the random experiment the person is selected in, say $p \cdot r$ samples. Thus, according to statistical rules for large *r* and not too small *p* (since *p* is expected to differ not too much from the theoretical inclusion probability)

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{r}};\ p + 1.96\sqrt{\frac{p(1-p)}{r}} \right]$$

includes the true value *P* with a probability of 95%. Due to $1.96\sqrt{\frac{p(1-p)}{r}} < \sqrt{\frac{1}{r}} = 0.01$ with $r = 10{,}000$, the value *p* computed by simulation only deviates *at maximum* in the third decimal place from the true value *P*, most likely even later. This error seems to be negligible in practice.

The experiment is repeated very often and following the law of large numbers the averaged inclusion probability for one element gets asymptotically closer to the true inclusion probability. This principle is commonly used in Monte Carlo simulations. For the statistical properties of the Monte Carlo Estimator, see for example Robert et al. (1999, pp. 20), Rizzo (2008, pp. 153) or Hammersley (1964, pp. 51). Theoretical inclusion probabilities are the result of

$$\pi_{gi} = \begin{cases} \pi_g \cdot \pi^b_{gi} \cdot \pi^{PIAAC}_{gi \mid b} = m\frac{N_g}{N} \cdot \frac{M_g}{N_g} \cdot \frac{32}{M_g} = 32 \cdot \frac{m}{N} \\ \qquad\qquad \text{for community } g \text{ with MOS } \ m\frac{N_g}{N} \le 1 \\ \pi_g \cdot \pi^b_{gi} \cdot \pi^{PIAAC}_{gi \mid b} = \frac{M_g}{N_g} \cdot \frac{32 \cdot smp_g}{M_g} = 32 \cdot \frac{m}{N} \\ \qquad\qquad \text{for community } g \text{ with MOS } \ smp_g = m\frac{N_g}{N} > 1 \end{cases}$$

where MOS is the measure of size and

$$\pi_g = \begin{cases} m\dfrac{N_g}{N} & \text{for community } g \text{ with MOS } \; m\dfrac{N_g}{N} \le 1 \\ 1 & \text{for community } g \text{ with MOS } \; smp_g = m\dfrac{N_g}{N} > 1 \end{cases}$$

is the probability for selecting community g. smp_g is the number of sample points in community g, which were selected using the Cox (1987) algorithm, i.e. $smp_g \in \left\{ \left[m\frac{N_g}{N} \right], \left[m\frac{N_g}{N} \right] + 1 \right\}$ with $[x]$ the largest integer $\le x$. The $m = 277$ communities are selected proportionally to the number N_g of their 16-65 year-old inhabitants. Overall in Germany there are $N = 53{,}989{,}232$ 16-65 year-olds. If $m\frac{N_g}{N} > 1$ then π_g is set to 1.

$\pi_{gi}^b = \dfrac{M_g}{N_g}$ is the probability that unit i is part of the gross sample of persons of size M_g which was provided from the registry of community g.

Under equal probability sampling, $\dfrac{32 \cdot smp_g}{M_g}$ is the probability that unit i is selected from the PIAAC gross sample of size $32 \cdot smp_g$. For community g both the gross sample size M_g of persons and the number of inhabitants that are 16-65 years-old N_g are known; thus $\pi_{gi}^b = \dfrac{M_g}{N_g}$.

Due to the error in the optimization algorithm used by our survey organization for the sample selection, an equal probability sample was not realized. Thus, inclusion probabilities $\pi_{gi|b}^{PIAAC}$ could only be determined by approximation through simulations. For $r = 10{,}000$ simulated samples, the following inclusion probabilities $\pi_{gi}^b \cdot \pi_{gi|b}^{PIAAC}$ for the units i of the PIAAC sample are computed given the erroneous and correct algorithm (see histograms in Figure 4).

The design effect due to unequal selection probabilities is

$$Deff_p = n \frac{\sum_{i=1}^{I} n_i w_i^2}{(\sum_{i=1}^{I} n_i w_i)^2} = 1.22,$$

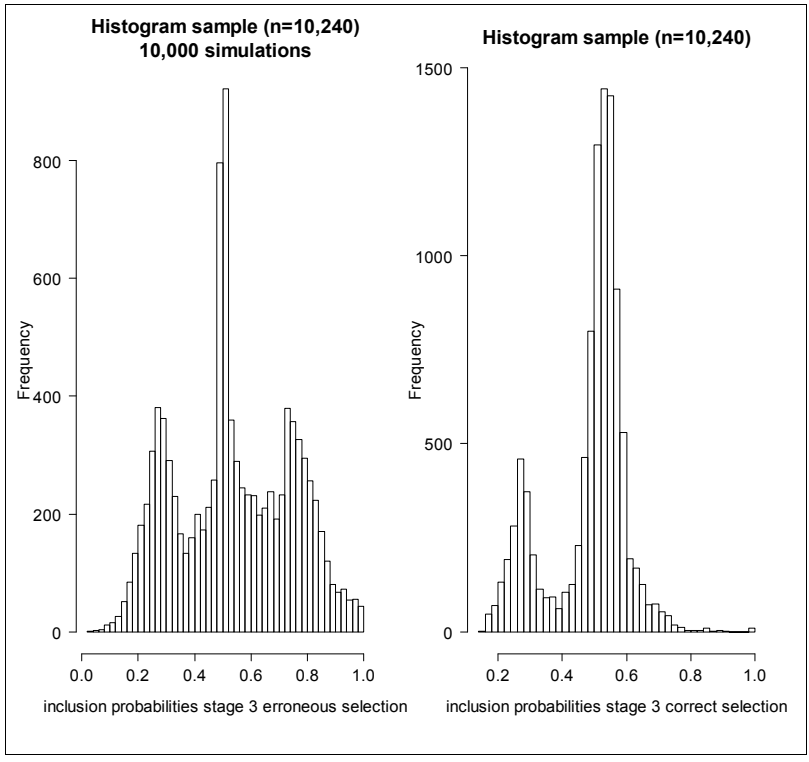


Figure 4: Histograms of $\pi_{gi|b}^{PIAAC}$ for units i of the PIAAC sample given the erroneous and correct algorithm.

where n_i is the number of observations in weighting class i and w_i are the weights in weighting class i. An explanation for the higher design effect given the erroneous optimization algorithm implemented by the survey organization is that this selection process favored certain units while neglecting certain other units. As a consequence, the required equal selection probability was not achieved.

4 Conclusion

Theoretically, the PIAAC sample for Germany should have been selected with equal probabilities for all individuals. However, due to an error in the selection procedure, this target could not be realized. Instead, an erroneous optimization algorithm was applied which led to inclusion probabilities that were too complex to calculate for us in the available time. But since the optimization procedure was

a random procedure, it was possible to determine the probabilities with the help of a simulation. The selection procedure was repeated 10,000 times and the number of times being included in the sample for each individual was reported. This number divided by 10,000 yields a good approximation of the inclusion probabilities. The disadvantage of the incorrect optimization algorithm for our sample is the higher design effect compared to the one based on equal inclusion probabilities. This design effect due to unequal inclusion probabilities was 1.22, i.e. the effective sample size was $n_{eff\_p} = n_{net}/Deff_p = 5{,}319/1.22 = 4{,}360$. In other words: The precision of the estimates is – only because of this error – just as high as if 4,360 interviews of a simple random sample would have been conducted. This is 82% of the original sample size.

Nevertheless, the PIAAC sample is a full probability sample and complies with all requirements of the Consortium. The sample passed the adjudication procedure with the following statements: "Through Consortium review of the preliminary SDIF, an anomaly was detected in the age distribution of the sample, with spikes at ages 30, 40, and 50. Germany investigated the reason for this pattern and discovered an error in the sample selection algorithm at the last stage of selection. Germany provided evidence that the sample remained probability-based despite this error and corrected the selection probabilities to reflect the actual selection algorithm used. However, they were unable to calculate exact selection probabilities, so the probabilities are based on a simulation" (see OECD 2013, Appendix 7, p. 69).

Quite generally, a good approximation for the true inclusion probabilities with 10,000 simulation runs is only meaningful if the sampling fraction f is high enough. In the case of the PIAAC-sample $n = \sum_g n_g = 10{,}240$ out of $N = \sum_g M_g = 23{,}117$ cases had to be selected, so $f = n/N = 0.44$. Otherwise, with a (much) lower sampling fraction the simulation with 10,000 replicates just would have led to white noise and it would have been impossible to determine inclusion probabilities this way.

References

Cox, L. H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association, 82*(398), 520-524. doi: 10.1080/01621459.1987.10478456

Fisher, R. A. & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. London: Oliver and Boyd.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo methods (Vol. 1)*. London: Methuen.

Kish, L. (1994). Multipopulation Survey Designs. *International Statistical Review* 62, 167-186.

Lynn, P., Japec, L., & Lyberg, L. (2006). What's So Special About Cross-National Surveys? In J. Harkness (Ed.), *Conducting Cross-National and Cross-Cultural Surveys.* ZUMA-Nachrichten Spezial Band 12, 7-20.

Lynn, P., Liebig, S., & Weinhardt, M. (2014). *Sampling for the European Social Survey – Round VII. Germany.* Retrieved from http://www.europeansocialsurvey.org.

OECD (2009). *PIAAC Sampling Plan (Main Survey) Part I: Sample Design and Selection Plans.* Revised Report (March 2009). Information for the National Survey Design and Planning Report (NSDPR). Chapter 4.

OECD (2013). *Technical Report of Adult Skills* (PIAAC). Paris: OECD.

Rizzo, M. L. (2007). *Statistical computing with R*. CRC Press.

Robert, C. & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014): *PIAAC Germany 2012: Technical Report*. Münster: Waxmann.

Authors Volume 8, 2014

- Daniela Ackermann-Piek, Mannheim
- Yasemin El-Menouar, Guethersloh
- Siegfried Gabler, Mannheim
- Sabine Häder, Mannheim
- Susanne Helmschrott, Mannheim
- Joost W.S. Kappelhof, The Hague
- Jan-Philipp Kolb, Mannheim
- Débora B. Maehler, Mannheim
- Silke Martin, Mannheim
- Natascha Massing, Mannheim
- Marcel Noack, Duisburg
- Anja Perry, Mannheim
- Beatrice Rammstedt, Mannheim
- Kurt Salentin, Bielefeld
- Rainer Schnell, Duisburg
- Simon Wiederhold, Munich
- Anouk Zabal, Mannheim

Reviewers Volume 8, 2014

We would like to thank the following colleagues for their careful review of the manuscripts published in mda, Volume 8, 2014:

- Johann Bacher, Linz
- Bernad Batinic, Linz
- Inna Becher, Konstanz
- Dorothée Behr, Mannheim
- Constanze Beierlein, Mannheim
- Nicole Biedinger, Mannheim
- Jörg Blasius, Bonn
- Michael Blohm, Mannheim
- Annelies Blom, Mannheim
- Michael Bosnjak, Mannheim
- Michael Braun, Mannheim
- Jan Pablo Burgard, Trier
- Alexandru Cernat, Colchester
- Daniel Danner, Mannheim
- Yasemin El-Menouar, Guethersloh
- Michèle Ernst Stähli, Lausanne
- Cornelia Gresch, Berlin
- Sigrid Haunbacher, Olten
- Franz Höllinger, Graz
- Volker Hüfgen, Düsseldorf
- Andreas Humpert, Duisburg
- Hans Kiesl, Regensburg
- Achim Koch, Mannheim
- Dagmar Krebs, Ludwigshafen
- Frauke Kreuter, Mannheim
- Cornelia Kristen, Bamberg
- Seppo Laaksoonen, Helsinki
- Oliver Lipps, Lausanne
- Geert Loosveldt, Leuven
- Débora B. Maehler, Mannheim
- Jutta von Maurice, Bamberg
- Andrew Mercer, Silver Spring, Maryland
- Yfge P. Ongena, Groningen
- Andreas Quatember, Linz
- Kurt Salentin, Bielefeld
- Evi Scholz, Mannheim
- Dirk Sikkel, Amsterdam
- Christian Spoden, Jena
- Daniel Stegmüller, Mannheim
- Volker Stocke, Kassel
- Mark Trappmann, Nuernberg
- Rolf van der Velden, Maastricht
- Christian Vollmer, Heidelberg
- Stefan Zins, Mannheim
- Cornelia Züll, Mannheim

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).