Lessons from mixed-method evaluations—An example from labor market research

Christopher Osiander*

Department of Research Coordination, Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung; IAB), Regensburger Strasse 104, 90478 Nuremberg, Germany

*Corresponding author. Email: christopher.osiander@iab.de.

Abstract

This article deals with two issues: first, it will analyze a. which approaches in the context of labor market research are generally available and often applied for program evaluation b. why a so-called mixed-method design often is a very useful way to gain knowledge, that is, a combination of different methodological approaches.

Second, the article uses an example from the evaluation of active labor market policy (ALMP) to show how a mixed-method design can look like in research practice and what added value it provides for research and policy advice.

The example is the ‘Initiative zur Flankierung des Strukturwandels’ (IFlaS; analogously in English: ‘Program to Support Structural Change’). It is a program of the German Federal

1. Introduction

Evaluation research can and should help to tackle issues of acceptability of political decisions or programs and their consequences in professional circles and the public to increase their legitimacy or to inform policymakers and give suggestions for program modifications (Eichhorst, Sesselmeier and Yollu-Tok 2008; Head 2010; Fricke, Koch and Kupka 2012). Researchers from different disciplines already have recognized that (1) high-quality data, (2) empirical research based on the state-of-the-art, and (3) evidence-based recommendations derived from (1) and (2) should go hand in hand (Zimmermann 2011). Labor market research in Germany is a good example for a field where the mutual exchange between politics, administration, and science is particularly well-established (Fricke, Koch and Kupka 2012).

The central questions of economic evaluation research are ‘What effects does a political intervention have?’ Or ‘How does a program work?’. These supposedly simple questions are often difficult to answer, because phenomena in the social sciences have at least three important characteristics: first, there is not one causal explanation for them. Moreover, these causes are not independent of each other. Second, as will be explained below, there is no single best method to analyze the impact of a program because programs may have effects on various outcomes. Third, there are several data sources and data sets to answer the questions.

This article deals with two issues: first, it will analyze

a. which approaches in the context of labor market research are generally available and often applied for program evaluation and

b. why a so-called mixed-method design often is a very useful way to gain knowledge, that is, a combination of different methodological approaches.

Second, the article uses an example from the evaluation of active labor market policy (ALMP) to show how a mixed-method design can look like in research practice and what added value it provides for research and policy advice.

The example is the ‘Initiative zur Flankierung des Strukturwandels’ (IFlaS; analogously in English: ‘Program to Support Structural Change’). It is a program of the German Federal
Employment Agency (FEA), where unemployed persons can participate in qualification measures that lead to vocational degrees. This is an important topic due to numerous developments in the labor market. First, it is to be expected that Germany’s labor force potential will decrease significantly in the course of demographic change even in case of substantial net immigration (Fuchs and Kubis 2016; Fuchs et al. 2017). Second, low-skilled workers—i.e. those without a vocational degree from the German ‘dual system’—face a high risk of unemployment and earn significantly less over their life course than high-skilled workers (Carnevale, Rose and Cheah 2011; Schmellen and Stüber 2014; IAB [Institut für Arbeitsmarkt und Berufsforschung] 2016; Stüber 2016). The high political importance of vocational and further training is also reflected in the German Coalition Agreement of 2018, which includes the political aim to support low-skilled workers by using qualification measures (CDU/CSU/SPD [Christlich-Demokratische Union; Christlich-Soziale Union; Sozialdemokratische Partei Deutschlands] 2013: 64). Moreover, the government states that further training is an important strategy to deal with digitization (CDU/CSU/SPD [Christlich-Demokratische Union; Christlich-Soziale Union; Sozialdemokratische Partei Deutschlands] 2013: 41).

The article is structured as follows: in Section 2, I describe some important evaluation approaches used in labor market research and explain why a mixed-method design is often an appropriate strategy to conduct in-depth evaluations. Section 3 describes the qualification program and its evaluation design. Section 4 provides an overview of some selected results and conclusions from the evaluation, and uses this example to show the advantages of closely integrating different forms of evaluations in this context. Section 5 provides a conclusion.

2. Types of evaluation and their significance in the context of labor market research

The term ‘evaluation’ includes processes, results, as well as certain procedures (e.g. Widmer 2000; Kromrey 2001b; Patton 2018). Often an evaluation is described as a comprehensible and systematic analysis of an object, a project or program by which its quality and/or (potential) benefit or utility is to be determined (Stockmann 2007; DeGEval [Deutsche Gesellschaft für Evaluation e. V.], 2008; Fricke, Koch and Kupka 2012). Moreover, uncertainty in decision-making should be reduced (Lee 2000 with reference to Mertens 1998). For example, evaluations are classified according to their purpose, subject, institution, or method (Widmer 2000; Palinkas et al. 2011; Dopp et al. 2019).

An important distinction is the one between summative and formative evaluations or evaluation parts. Summative evaluations analyze the extent to which a program has achieved its intended goals(s). However, they do not answer the question of why programs (do not) work or how they can be improved during the time they exist. Summative analysis happens after the program has ended, or if not, retrospectively, beginning from a fixed date (Scriven 1972; Kromrey 2001b). The majority of summative evaluations are based on quantitative methods although both terms are not synonymous. In the context of labor market research, summative evaluations often try to answer the following question: ‘To what extent does a policy intervention contribute to achieving desired goals?’ An example would be the question of whether and, if so, to what extent qualification measures contribute to increasing the employment chances of the participants and/or to reducing or avoiding unemployment benefit receipt. In economic evaluation research, summative evaluations are very common and quantitative methods have been and still are predominant (Starr 2014).

From a methodological point of view, one faces the so-called fundamental evaluation problem or the problem of causal inference (Roy 1951; Rubin 1974; Holland 1986). Units of analysis—i.e. persons, households, or firms—cannot be observed simultaneously in mutually exclusive states. For example, individuals cannot be participants and non-participants in a qualification measure at the same time. Therefore, one has to construct the counterfactual situation by using (quasi-)experimental designs or to construct it in retrospect by using administrative data. This is to answer the question: what would have happened to a person if, unlike in reality, he or she did not participate in the program?

The selection process—i.e. the allocation of persons to a program—is not random, due to a number of reasons: first, employment agencies exercise discretion within the legal framework and within their financial capacity; job counselors in employment agencies play a crucial role in deciding who receives what services from the ALMP toolbox at what time (for the concept of discretion, see, e.g. the classic study of Lipsky 1980). Programs of ALMP are often tailored to specific groups (e.g. low-skilled workers or single parents), thereby excluding others who are not in the target group. This can also be considered external selection. On the other hand, potential participants often have a choice, e.g. if participation in programs is voluntary and requires so-called co-production—a participant’s willingness to cooperate (Stephan and Wolff 2017). This can be interpreted as self-selection. Both self-selection and external selection result in certain groups of people being overrepresented in programs, such as motivated or competent persons or those with certain labor market-relevant characteristics.

Therefore, a simple comparison of the labor market outcomes of participants and non-participants is not sufficient for identifying program effects, because participants and non-participants differ systematically. Differences in outcome variables may be due to two causes that cannot be clearly distinguished: first, the causal effect (program effect) and second, the effect of different characteristics (selection effect). Therefore, one has to find ways to control for selection bias in order to make both groups comparable.

In the literature, randomized controlled trials (RCTs) are often considered a ‘reference design’ (Kromrey 2001a,b, translation by the author), quoted as ‘gold standard’ (Henry 2015; Torgerson, Torgerson and Taylor 2015; Webber and Prouse 2018) or even called the ‘Rolls-Royce of evaluation approaches’ (Chelimsky 1997: 101). Randomization is used to avoid selection bias and to measure the causal effect of interventions on outcome variables as unbiased and precisely as possible.

RCTs have a long tradition in economic research in general (Harrison and List 2004; Levitt and List 2007, 2009) and labor market research in particular (List and Rasul 2011; Rothstein and von Wachter 2017) and their results have often received much attention [e.g. the classic LaLonde (1986) study]. However, there are practical, methodical, and ethical challenges when conducting RCTs (Head 2010).

First, the practical implementation is usually complex, since politics, administration, and researchers must cooperate closely for successful implementation. There are important questions that have to be answered: is it possible to implement random assignment? If so, how should it be put into practice? What is the organizational and
administrative effort before and during an RCT and to what extent
is this additional work to be done by the available staff?

Second, the notion that RCTs are generally superior to other
evaluation methods has not remained uncontested (e.g. Cartwright
2007, Cartwright and Munro 2010; Cartwright and Hardie 2012;
Deaton and Cartwright 2018: 30). Even in RCTs, there are factors
that can hamper the identification of causal effects. This is true
when researchers have no control over alternative causes or moder-
ator variables that can affect the effect and its size (Cartwright
and Munro 2010; Cowen et al. 2017). The same applies if control
group members are exposed to alternative treatments that are not observed
in the data (Rothstein and von Wachter 2017). Moreover, the labor
market results of an individual must be independent of the treatment
status of other individuals (Rothstein and von Wachter 2017)—the
so-called stable unit treatment value assumption. Furthermore, there
are issues of external validity or generalizability, i.e. the transferability
to other populations or situations (Levitt and List 2007; Cowen
et al. 2017; Rothstein and von Wachter 2017). Or, as Cartwright
and Hardie (2012: 8) put it: ‘No matter how much gold standard
you have for “it worked there,” you cannot pave the road from there
to here with gold bricks’.

Third, ethical issues are discussed (Pearce and Raman 2014), for
example, the question under which conditions it is morally justifi-
able to randomly assign certain groups of people to a treatment
while others are excluded.

Therefore, it is necessary, but not sufficient that a causal effect is
measured. Rather, it is crucial to understand the mechanisms that
lead to a result. In other words, it is not only important to know
whether and, if so, to what extent a program works, but also why
and how (Starr 2014). It is necessary that researchers conduct RCTs
with a clear theoretical foundation. Otherwise, important informa-
tion may be left out.

When RCTs are impossible to implement, other methods are
applied (for a more detailed review, see, e.g. Blundell and Costa
Dias 2009; Imbens and Wooldridge 2009). The most important
methods in labor economics are as follows:

I Natural experiments (Blundell and Costa Dias 2009) exploit
naturally occurring phenomena that lead to randomization.
They combine a comparison of different groups (treated vs.
untreated) before and after exposure to the treatment. Therefore,
they are equivalent to difference-in-differences approaches.

II Instrumental variable (IV) estimations (Angrist and Krueger
2001; Blundell and Costa Dias 2009) are used to control for
unobserved heterogeneity or omitted variable bias. Omitted
variables may influence both outcomes and explanatory fac-
tors. An IV is assumed not to have any direct effect on the
outcomes of interest, but it affects selection into treatment.
Therefore, IVs are a tool for estimating the causal relationship
between explanatory variables and outcome consistently and
unbiased.

III Discontinuity designs (Blundell and Costa Dias 2009; Lee
and Lemieux 2010) use situations where the enrollment into
treatment is not random, but probability of enrollment changes
continuously at some specific and well-known cutoff value or threshold of a single—often continuous—assign-
ment variable.

IV Statistical matching (Heckman, Ichimura and Todd 1997;
Caliendo and Kopeinig 2008; Stuart 2010) uses statistical
twins for comparisons. Those receiving a treatment are com-
pared to those who do not. However, the researcher controls
for those characteristics that are correlated both with the
probability to be treated and the outcome variable. In the
case of a qualification measure, therefore, participants are
compared with non-participants who are similar to the for-
mer in all labor market-related characteristics. After match-
ing, the probability of being treated can be regarded as
quasi-random and not affected by selection bias. Thus, dif-
fences in outcomes of participants can be considered as
caused by the program.

The choice of outcome variables in evaluation research in general
and for impact analyses in particular should be carefully considered:
they should reflect both the goals of the program and the central re-
search questions included in the program evaluation. Ideally, choos-
ing them should be a normative decision made by well-informed
researchers who carefully consider which data they need to answer
their questions and to base their conclusions on. This is closely
related to the question which kind of story researchers want to tell
‘to address the concerns of both the heart and the head’ (Krueger
2015) of readers and practitioners.

Unfortunately, this ideal often cannot be maintained in research
practice: in labor market research, it is—probably similar to other
fields of research—often a practical decision as to which outcome
variables are chosen. To a large extent this choice is determined by
the availability of valid data. For qualification measures, for ex-
ample, taking up employment and avoiding unemployment and/or
benefit receipt are often considered ‘desirable’ outcomes for most
political actors and practitioners in labor administration. But even
results that are supposedly desirable are not always unproblematic:
Economic search models, for example, suggest that longer un-
employment spells may also indicate that individuals are reluctant
to accept low-quality jobs and might pay off in terms of better job
quality (Nekoei and Weber 2017).

However, impact analyses are only possible after a longer time
span and if very detailed (administrative) micro-data are available.
Moreover, these data do not provide information about the mecha-
nisms or organizational context factors that affect an outcome. This
missing contextual knowledge is also referred to as ‘qualitative resi-
dual’ in the literature (Onwuegbuzie and Teddlie 2003).
Therefore, it makes sense to also use formative evaluations. They
are usually carried out during the project and their results influence
the design of the ongoing program—assuming program managers
are willing to modify it (Scriven 1972, 1991; Widmer 2000;
Palinkas et al. 2011; Fricke, Koch and Kupka 2012). For this reason,
the information that the evaluation makes available to program
managers must offer decision-making options (Kromrey 2001b;
Head 2010). Program modifications are to increase the extent to
which the program achieves its goals and/or to correct problematic
elements. In this way, formative evaluations are also integrated into
those processes which pursue a specific change (Chelimsky 1997;
Kromrey 2001b). Often they are carried out in the form of an imple-
mentation study.

Implementation studies analyze different aspects: first, how
employees at the street-level implement the program. Second, to
what extent the program actually reaches its target group(s). Here, it
can provide information on the extent to which implementation
practice supports or undermines program goals, or identify
unintentional side effects. However, a formative evaluation does not provide information about program effects on outcome variables.

The different strengths and weaknesses of summative and formative evaluations clearly suggest that there is no ‘magic bullet’ for evaluation (Head 2010). On the contrary, different approaches can be combined to get a complete picture because summative and formative elements, quantitative and qualitative methods, as well as mixed-method studies are appropriate under different circumstances (Johnson and Onwuegbuzie 2004; Palinkas et al. 2011; Jefferson et al. 2014; Starr 2014; Cronin 2016; Molina-Azorin 2016). Such an approach—the combination of summative and formative elements and of quantitative and qualitative methods—is often called a mixed-method design, method combination or triangulation (Johnson, Onwuegbuzie and Turner 2007; Kuckartz 2014: 29). The key advantage is to combine both the strengths of quantitative and qualitative methods and to avoid their respective weaknesses. The price one has to pay is the additional effort required for an implementation of mixed-method designs.

In our case, we were able to incorporate additional insights from the qualitative part into the design of the summative part. In addition, qualitative and quantitative approaches offer different perspectives on qualification measures and answer different questions: first on the quality of the implementation practice and second on the program effect on the participants. This will be clarified by a case study from the evaluation of ALMP.

3. The case of publicly sponsored qualification measures

3.1 The program: ‘IFlaS’

The FEA promotes qualification measures for low-skilled unemployed persons, i.e. for persons who never completed a vocational degree or whose qualification is no longer useful on the labor market. This is the focus of IFlaS. The program had two objectives: first, it was intended to alleviate regional, occupational, or sectoral skills shortages that are a result of demographic and structural change. Second, it was intended to improve the labor market chances of the unemployed by supporting them on their way to a vocational degree (for a more detailed description of the program, see Fertig and Osiander 2015). The target group were jobseekers and—far less common—workers directly threatened by unemployment, who at the time did not have completed vocational training or who had an ‘obsolete’ vocational degree according to §77(2) No. 1 SGB III (today: §81(2) No. 1 SGB III) (BA [Bundesagentur für Arbeit] 2009).

1 In rare cases, the term ‘mixed methods’ is also used for studies that combine methods that are less common in economics (see, e.g. Marshall 2017, 2018).

2 Other labels for this kind of research are mixed research, integrative research, blended research, triangulated science, etc. Starr (2014) calls the combination of qualitative and quantitative elements ‘Q-squared’ research. For an analysis, see Johnson, Onwuegbuzie and Turner (2007).

3 ‘Obsolete’ vocational degree means that an unemployed person has worked for at least 4 years in a job that requires a different vocational degree than the one he or she once has completed. In the German ‘dual system’, vocational degrees are important certificates that signal specific skills.

In fact, many foreign citizens do not have vocational degrees that are valid in Germany. The program was designed for the unemployment insurance system—the so-called Social Code III (in German: Sozialgesetzbuch III or SGB III).

The potential participants should have had a job recently to signal that they are attached to the labor market. They were given an educational voucher during an interview with their job counselor. They were allowed to redeem the voucher at a licensed provider for qualification measures. The vouchers are usually valid for a course in a pre-specified occupational field within a certain period of time—often 3 months. During the course, the participants received unemployment benefits in the same amount as previously. The employment agency pays for all costs related to the course, e.g. course fees and travel costs. The focus of the program is on qualification measures that lead to a vocational degree, but it was also possible to promote so-called partial qualifications, i.e. short courses that provide specific occupational knowledge.

3.2 Evaluation design

The research design deliberately combined formative and summative elements and related them directly to each other. To the best of our knowledge, this is the first time that these elements were jointly used in an evaluation of qualification measures in Germany. Using Dopp et al.’s (2019) analytical distinction for mixed-method designs, it is an evaluation where the approaches were partly used in sequence, and partly simultaneously. The formative (or qualitative) elements precede the summative (or quantitative) elements (in part) and overlap (in part). In addition, the evaluation is complementary, i.e. the qualitative and the quantitative part answer different but related questions in one single study. According to Palinkas et al. (2011), the design is also

a. unbalanced, i.e. there is one part of the evaluation that is considered the more ‘important’ or dominant part. In this case, the standardized survey and the impact analysis are considered the main part of the study compared to the formative part
b. embedded, i.e. the qualitative case studies are embedded in an impact analysis of the outcomes.

The formative part of the evaluation consists of qualitative case studies in 30 regions (for more information, see ISG [Institut für Sozialforschung und Gesellschaftspolitik GmbH]/Lawaetz [Johann Daniel Lawaetz-Stiftung] 2010, on case studies in economics in general, see Starr 2014). The aim was to understand implementation practice at the street-level and possibly give some advice for re-adjustment where it was necessary. The selection of the employment agency districts was based on the following criteria:

• The labor market situation (good/medium/bad).
• The relative size of the target group (share of low-skilled unemployed on all unemployed: high/low).
• The regional distribution (county/city).

Information on the regional labor market situation was drawn from the labor market typology for Germany by Dauth, Hirschenauer and Rübg (2008). There should be similar as well as contrasting regions in the empirical analysis. The selection is therefore limited to few criteria, which are considered theoretically important for the sampling. There are $3 \times 2 \times 2 = 12$ possible combinations of all dimensions. So it is ensured that every cell of a crosstab is occupied in a sample of 30 cases.
The case studies consisted of interviews with experts for the local implementation of the program. The interviews were conducted in two waves: the first wave was in 2010 and the second wave with a reduced number of 14 regions was in the summer of 2012. In the first wave, there were usually at least three interviews per region: one with a member of the board of employment agencies, one with the manager of the department that is responsible for labor market integration and counseling, and/or one interview with the coordinator for qualification measures as well as a job counselor. Apart from a few justified exceptions, there were also interviews with managers of the Chamber of Commerce and Industry and the Chamber of Crafts. In most cases, these were interviews with the deputy management or some of the department leaders responsible for vocational education and further training. In addition, in each region there were interviews with one or two managers from providers for qualification measures for jobseekers.

The main focus of the case studies was to analyze the way in which the program was implemented: the process itself and its quality in general and the selection of the potential participants for the program in particular. So the central research question is: what is being done regarding the implementation of the initiative in different employment agencies that could lead to unobservable heterogeneity between participants and non-participants or to other relevant (quality) differences? Based on the findings of the first wave of interviews, a standardized survey for the employment agencies was developed. There were three waves of the survey at the end of 2010, in the summer of 2011 and at the end of 2011. The questionnaire was an editable PDF file and was sent by e-mail to the members of the boards of the local employment agencies with a request for a reply.

The summative part of the evaluation tries to find an answer to the following question: to what extent do qualification measures help to improve labor market chances of the participants? We used so-called propensity score matching (Rosenbaum and Rubin 1983; Caliendo and Kopeinig 2008). The basis of the matching is a data set consisting of combined survey and administrative data from the FEA, the so-called Integrated Employment Biographies (IEB).

The combination of survey and administrative data is an innovation from a methodological point of view: to the best of our knowledge, this is the first time that such an approach has been chosen in the context of evaluating qualification measures. We are able to statistically control for factors that are otherwise difficult to observe, such as attitudes toward work and life or locus of control. This is especially important when these factors affect the probability of participation and the outcome variables. General labor market-related skills are usually approximated by the medium- and long-term labor market history of individuals, as well as soft factors such as achievement motivation, perseverance, and other labor market-relevant personality traits. This is a plausible strategy, but unobserved differences between the groups in respect of these factors cannot be ruled out completely. Thus, the survey data can provide an answer to the question whether such data have an additional value in this context and improve the matching.

The survey data were collected in a standardized computer-aided telephone survey. We drew a monthly random sample of unemployed persons in the period from July to November 2010, consisting of participants and non-participants. The aim of this approach was to be able to interview participants as soon as possible after they joined the program. The interviews were conducted between 1 October 2010 and 18 February 2011 by the social research institute ‘Sozialforschung und Kommunikation (SOKO)’ from Bielefeld, Germany. In the relevant first wave of the survey, a net response rate of 42.2% was realized. Table 1 shows the net response rates.

Overall, a full interview was conducted with 40% of the participants and 43% of the non-participants. With the consent of the interviewees, the survey results can be linked to the administrative micro-data, i.e. the IEB. A total of 89% of the respondents agreed to this record linkage. Further analyses were conducted for this group. The administrative data contain detailed personal information about the individual (un-)employment history, such as past employment and periods of unemployment benefit receipt, unemployment spells, and participation in measures of ALMP (qualification measures, publicly subsidized jobs, and the like).

### 4. Empirical results and conclusions from the evaluation

#### 4.1 Results of the formative evaluation

In this section, we show results from the formative evaluation. Particularly, it deals with two topics: first, the allocation logic and second, the framework conditions which may influence the allocation process—in this case, legal specifications and economic developments.

With regard to the allocation logic, a number of program managers criticized the design of the targeting. Some interviewees were critical of the limitation to low-skilled workers, in particular those who worked in employment agencies in Eastern Germany. This is because the share of low-skilled workers is generally lower in Eastern Germany than in Western Germany. Some interviewees added that the program excluded single parents (mostly women) and persons who wanted to get back to work after a longer parental leave. This was due to the fact that their vocational degree was not considered ‘obsolete’ when they were not employed because of parental leave or periods of child-rearing (IGB [Institut für Sozialforschung und Gesellschaftspolitik GmbH]/Lawaetz [Johann Daniel Lawaetz-Stiftung] 2010). In other words, they were not formally regarded as low-skilled. A political reform from April 2012 has changed the law (§77(2) No. 1 SGB III) and altered the definition of low-skilled workers. From this point on, single parents and persons who return from parental leave also have been considered low-skilled workers.

This is a good example of a powerful and successful program modification by using the results of a formative evaluation. In the further course of the evaluation, these modifications were also communicated back to the researchers: program managers and job counselors welcomed the legal change overwhelmingly. Some even stated that single parents have become one of the most important groups for the program over time. The result also indicates that the careful selection of experts for the implementation of a program is an important prerequisite for the success of a formative evaluation. Researchers must have a clear idea of what the potential ‘right’ interviewees are before the field phase begins. If it turns out at the beginning of the field phase that important stakeholders have been overlooked in the research design, this needs to be corrected, whenever possible. Street-level workers like job counselors are obviously important in this kind of research, because they have the most detailed knowledge of the application of the law and its pitfalls and shortcomings. In this case, the situation was communicated either to ...
the management level and from there to the researchers, or directly to the researchers.

From the point of view of the agencies, the amount of financial benefits were problematic for many unemployed persons. Unemployment benefits during a qualification measure are paid at the same rate as regular unemployment benefits, which are 60% of the last net salary and 67% for people with children. During longer qualification measures, participants often have to cope with a very tight financial budget for 2 or more years plus a period of job search following the qualification measure. This is likely to be the case in large cities or regions with above-average living costs. Although the entitlement to unemployment benefits does not expire even during longer courses, the remaining entitlement at the end of the course is only 30 days. So many participants are about to switch to the more rigid unemployment benefit II, i.e. the basic income scheme for needy jobseekers. As a result, there is a relatively strong financial incentive, in the eyes of the program managers, to quit the course in favor of employment in the low-skilled sector. This might be especially true when the circumstances in the household change (birth of another child, unexpected costs, etc.). Others may not even start a course in the first place, because the additional financial benefits of full-time employment are enormous in the short term and the duration of the qualification measure is relatively long. In economic terms, the opportunity costs of participating in the measure are quite high.

Jobseekers who are entitled to unemployment benefit II and participate in publicly sponsored employment (the so-called ‘one-euro jobs’) receive additional compensation on top of unemployment benefit II, which is up to two euros per working hour spent in the measure. The law explicitly states that this is a compensation for ‘extra effort’. Qualification measures are likely to involve considerable extra effort and time, but this is not rewarded by additional compensations.

These findings led to another quantitative research project that dealt with the potential difficulties and obstacles during qualification measures (Dietz and Osiander 2014; Osiander and Dietz 2016). The results from these projects were also discussed when it came to the passing of a new labor market law, which came into force in August 2016 (Deutscher Bundestag 2016). The law includes rules for financial rewards when participants in publicly sponsored qualification measures successfully pass intermediate and final exams. This points out that evaluations can not only be relevant to implementation practice, but in some cases they may also stimulate ideas for future research questions and policy advice.

The good labor market situation was also unexpectedly ‘problematic’. Employment agencies noted that there are less ‘suitable’ jobseekers in the pool of formally eligible persons. In other words, the number of ‘available’ jobseekers has decreased because the ‘better risks’ get a job and are no longer available for qualification measures due to the favorable labor market situation. The e-mail survey revealed that in 2010, employment agencies considered 36% of all eligible jobseekers actually ‘suitable’ for a qualification measure, compared to 28% in 2011. The consequence of this development is that the number of persons entering into measures also has decreased over time, because the identification of suitable candidates has become more difficult for employment agencies.

It becomes clear here that such problems can arise if the program is not adapted to changing conditions over time, i.e. if, for example, the budget is not adjusted or if a central target is to ‘fill’ measures. This reveals a conflict between the requirements of a target system (optimal use of budget, assignment of jobseekers to measures) and the quality of counseling for the unemployed, which should take into account their specific needs. A somewhat careful process of selection and allocation, which ignores preferences and skills of the unemployed, is likely to lead to some frustration among both the jobseekers and the job counselors and/or may increase drop-out rates. If one generalized a bit more, there is some evidence that a theoretically possible qualification of all low-skilled workers will fail in practice. This is because a substantial share of the eligible group probably lacks the essential prerequisites for successful participation in longer courses. It is obvious that qualifications only work under very specific conditions, for example, when funds are available for measures which also provide basic competences.

### 4.2 Results of the summative evaluation

As part of the summative evaluation, we apply propensity score matching. The variables shown in Table 2 were used as observables that may correlate with both participation and outcome variables.

Before matching, there are clear differences between participants and non-participants, which were already pointed out in a similar form by Fertig and Osiander (2015). The participants are systematically less likely to have an upper secondary school degree (‘Abitur’, roughly equivalent to ‘A-levels’) and more often to have a foreign school degree than the members of the comparison group. They also differ in aspects of qualification and the unemployment insurance system.

| Table 1. Gross sample and net response in the CATI survey ("Computer-aided telephone survey") |
|---|---|---|---|---|---|
| | Total | In % | Participants | In % | Non-participants | In % |
| Unadjusted gross sample | 25,282 | 8,601 | 16,681 |
| Quality-neutral losses (invalid address, no match with design) | 3,331 | 634 | 2,697 |
| Adjusted gross sample | 21,951 | 100.0 | 7,967 | 100.0 | 13,984 | 100.0 |
| No successful contact (absent, sick) | 6,294 | 28.7 | 3,173 | 39.8 | 3,121 | 22.3 |
| Refused to be interviewed | 6,392 | 29.1 | 1,588 | 19.9 | 4,804 | 34.4 |
| Complete interviews (= net response) | 9,268 | 42.2 | 3,206 | 40.2 | 6,059 | 43.3 |

Source: SOKO [SOKO Institut für Sozialforschung und Kommunikation] (2011), table also used in Fertig and Osiander (2015).

---

4 A single worker with a full-time job (~160 h of working time per month) who earns the German minimum wage of 9.35 € per hour (as of January 2020) and loses his or her job, is entitled to about 650 euros of unemployment benefits. For a worker with at least one child, the amount increases to about 730 euros.

5 This was the so-called ‘Arbeitslosenversicherungsschutz- und Weiterbildungsstärkungsgesetz (AWStrG)’ that dealt with...
| Independent variable | Coding | Participants | Non-participants | P-values and significance | Participants | Non-participants | P-values and significance |
|----------------------|--------|--------------|-----------------|--------------------------|--------------|-----------------|--------------------------|
|                      |        | Before matching | After matching  |                         |              |                 |                          |
| Socio-demographic variables |        |              |                 |                          |              |                 |                          |
| Male                 | (1 = true) | 58.1% | 57.7% | 0.528 | 58.5% | 57.5% | 0.543 |
| 25–30 years          | (1 = true) | 28.6% | 28.2% | 0.595 | 27.5% | 26.5% | 0.488 |
| 31–36 years          | (1 = true) | 25.3% | 24.4% | 0.000*** | 23.5% | 24.2% | 0.636 |
| 36–40 years          | (1 = true) | 22.1% | 22.2% | 0.832 | 23.9% | 24.2% | 0.812 |
| 41–45 years          | (1 = true) | 24.1% | 27.1% | 0.000*** | 25.1% | 25.1% | 0.993 |
| Foreign nationality  | (1 = true) | 19.8% | 18.2% | 0.000*** | 16.6% | 14.7% | 0.126 |
| Handicap             | (1 = true) | 1.3% | 3.6% | 0.000*** | 4.4% | 5.1% | 0.326 |
| Household size       | # of persons | 2.86 | 2.63 | 0.000*** | 2.75 | 2.83 | 0.048** |
| Number of children in household | (1 = true) | 0.89 | 0.91 | 0.000*** | 0.68 | 0.68 | 0.737 |
| Formal and informal education |        |              |                 |                          |              |                 |                          |
| No school degree     | (1 = true) | 4.3% | 3.9% | 0.379 | 3.8% | 4.6% | 0.250 |
| Low secondary school degree | (1 = true) | 28.1% | 26.5% | 0.125 | 28.9% | 29.3% | 0.805 |
| Medium secondary school degree | (1 = true) | 40.7% | 40.8% | 0.963 | 42.3% | 40.5% | 0.277 |
| Upper secondary school degree | (1 = true) | 17.9% | 21.9% | 0.000*** | 18.3% | 19.6% | 0.318 |
| Foreign school degree | (1 = true) | 9.1% | 6.9% | 0.000*** | 6.7% | 6.1% | 0.417 |
| Number of books in household |        |              |                 |                          |              |                 |                          |
| 0–9 books            | (1 = true) | 7.1% | 11.1% | 0.000*** | 7.7% | 7.7% | 0.997 |
| 10–49 books          | (1 = true) | 32.0% | 33.9% | 0.079* | 33.3% | 32.0% | 0.392 |
| 50–99 books          | (1 = true) | 26.3% | 24.2% | 0.037** | 26.1% | 26.0% | 0.951 |
| 100 books or more    | (1 = true) | 34.5% | 30.7% | 0.000*** | 32.9% | 34.3% | 0.366 |
| PC skills and knowledge of text processing programs | (1 = true) | 78.8% | 70.0% | 0.000*** | 79.1% | 77.9% | 0.358 |
| Attitudes            |        |              |                 |                          |              |                 |                          |
| Professional success is coincidence or luck | Scale 1–5 (4/5 = yes) | 39.2% | 42.0% | 0.013** | 37.7% | 37.4% | 0.812 |
| World is complicated | Scale 1–5 (4/5 = yes) | 20.8% | 29.1% | 0.000*** | 21.1% | 21.9% | 0.568 |
| Achievements in life are fate or luck | Scale 1–5 (4/5 = yes) | 40.9% | 46.1% | 0.000*** | 40.6% | 39.0% | 0.300 |
| Labor market history of the last 5 years |        |              |                 |                          |              |                 |                          |
| Successful vocational degree in the past | (1 = true) | 20.8% | 18.9% | 0.000*** | 20.8% | 19.8% | 0.446 |
| Vocational training terminated in the past | (1 = true) | 22.3% | 14.8% | 0.000*** | 20.5% | 19.7% | 0.537 |
| Regular employment   | # of days | 1,097 | 589 | 0.000*** | 1,183 | 1,180 | 0.844 |
| Unemployment         | # of days | 552 | 1,029 | 0.000*** | 563 | 579 | 0.264 |
| Registered as unemployed | # of times | 4.48 | 4.55 | 0.078* | 4.63 | 4.95 | 0.004** |
| Measures of ALMP     | # of days | 99 | 186 | 0.000*** | 100 | 93 | 0.202 |
| Episodes with measures of ALMP | # of times | 1.16 | 1.65 | 0.000*** | 1.18 | 1.19 | 0.834 |
| Episodes with no registration | # of times | 0.23 | 0.19 | 0.000*** | 0.19 | 0.19 | 0.753 |
| No registration      | # of days | 247 | 230 | 0.002*** | 216 | 226 | 0.397 |
| Unemployment insurance benefit receipt | # of days | 150 | 112 | 0.000*** | 169 | 184 | 0.004*** |
| Unemployment benefit II recepit | # of days | 207 | 830 | 0.000*** | 215 | 214 | 0.960 |
| Sickness during unemployment | # of days | 4.28 | 12.04 | 0.000*** | 5.57 | 4.50 | 0.336 |

(continued)
own more books⁶ and have a better computer literacy than the non-participants: They say they own a PC more often and are also able to name a word processor more often.

There are also substantial differences in employment history between the two groups: participants had been more likely to drop out of vocational training in the past than non-participants. It is plausible that this overrepresentation is due to a deliberate selection of unemployed by the job counselors because of precisely this characteristic. Non-completed vocational training does not necessarily have a negative meaning. On the contrary, it is an indicator that these jobseekers are somewhat closer to the labor market than those who have never started vocational training. In the past, the participants had also been slightly more likely to have completed vocational training than non-participants. Both facts suggest that the participants are ‘good risks’ from the pool of the formally eligible unemployed.

In addition, participants were employed almost twice as long as non-participants in the 5 years before the program start. They were also unemployed for a shorter period of time, participated less in programs of ALMP, and received significantly less often unemployment benefit II, i.e. the basic income scheme for needy jobseekers. Participants received unemployment benefits from the unemployment insurance system for a slightly longer period of time; this is probably a direct consequence of the longer periods of employment and the resulting entitlements. It also might indicate better labor market chances compared to non-participants. The cumulative income of the participants from the past 5 years is also significantly higher than that of the non-participants.

Moreover, there are differences in locus of control and attitudes toward work and life: participants are somewhat less likely to agree that coincidence or luck is responsible for professional success and that fate and luck determine what one achieves in life. In addition, they consider the world to be complicated less often. This also indicates that the participants are a selection of ‘good risks’ among those formally eligible.

This can also be interpreted as an indicator that the survey data offer added value, as they reveal differences that could not be observed with administrative data alone. We compare an unrestricted model with all variables specified in Table 2 with a restricted model that does not include the three variables that reflect personal attitudes and locus of control. A likelihood ratio test rejects the null hypothesis that there is no significant difference between the models (significance at the 1%-level). This indicates that just three variables from the survey data already improve the model.

Matching substantially reduces the systematic differences between the two groups. A nearest-neighbor matching is applied, in which each participant has five statistical twins. After matching, there are still differences at the 5%- or 1%-level for 4 out of 42 characteristics: participants live in slightly smaller households. They also have slightly more episodes where they are registered unemployed or receive unemployment insurance benefit; and they receive unemployment insurance benefits slightly longer. Although these differences are still statistically significant at the 5%- or 1%-level, they are rather minor deviations from the control group. All other characteristics are balanced after matching, i.e. remaining differences are no longer statistically significant from zero. The characteristics no longer explain the program participation after matching well. McFadden’s Pseudo-$R^2$ is 0.01, i.e. only about 1% of the variance is explained by the observed characteristics after matching.

Matching is not successful for every participant; there is no statistical twin for about 5% of the participants. These are predominantly people with a relatively high propensity score, i.e. persons for whom the observed characteristics explain participation very well. These individuals could be included in the analyses if ‘bad’ twins were also taken into account. However, this would be at the price of greater remaining differences between participants and non-participants after matching. This would increase the risk that observed effects could be due to the heterogeneity of groups rather than the program effect. The following analyses are therefore carried out for 5,232 persons, of whom 1,879 are participants and 3,353

---

⁶ The number of books in a household is used in studies on educational achievement like PISA or TIMSS to measure the socioeconomic background of a person (Bos et al. 2003; Bonsen, Frey and Bos 2008). It is assumed that ownership of books is correlated with a good ability to understand their content.
non-participants. For 370 participants, there were no suitable matches, so they are excluded from the impact analysis.

The results of the impact analysis are consistent with various labor market outcomes. The program effect on the probability of being independent of benefits shows a pattern common in the evaluation of qualification measures (see Figure 1): the rate of non-participants receiving unemployment benefits decreases rapidly after the start of observation. After 360 days, about 38% of the non-participants receive unemployment benefits. A total of 720 days after the start of the observation period, about 29% of the non-participants receive benefits, 1,080 days thereafter 26%. At the end of the observation period, the share of non-participants receiving benefits has fallen to around 18%.

Participants leave unemployment benefit receipt more slowly. This is not surprising, as this is due to the participation itself. After 360 days, 75% of the participants receive unemployment benefits. After 2 years, this rate drops when many participants complete their course. Over time, the participants outperform the non-participants. After about 1,080 days, 26% of the beneficiaries receive unemployment benefits, at the end of the observation period 11%.

The yellow curve indicates the program effect comparing participants and non-participants. The difference between the groups initially grows. This is called the lock-in effect in the literature, because participation in the program keeps participants from keeping up employment (e.g. Bosch 2012; Kruppe and Lang 2018). In the recent literature, this period is also referred to as the ‘investment phase’ (Bernhard 2016; Kruppe and Lang 2018), as in the case of qualification measures, the participants are intended to acquire human capital. Again, the effect has no negative meaning, as it is simply an intended consequence of the participation.

In this case, the investment effect is substantial and reaches almost 40 percentage points 300 days after the start of the program. After almost 2 years, the participants are catching up, and almost 900 days after the start, the differences between the groups are no longer statistically significant from zero. Starting from about 1,110 days, the probability of receiving no benefits is significantly higher than that of non-participants. This difference is 10–13 percentage points for a long period and decreases to around six to seven percentage points by the end of the observation period. Starting from 1,110 days after the action has been taken, it is statistically significant at the 1%-level at each further instant until the end of the observation.

Figure 2 shows how the probability of having a job subject to social security contributions develops for participants and non-participants.

At the beginning of the observation period, the share of non-participants taking a job subject to social security contributions increases sharply. After 720 days, almost 60% of the non-participants are employed (again). This share remains stable for a long time and rises again slightly about 1,600 days after the start. By the end of the observation period, it is between 60% and 66%.

Participants take up employment more slowly due to their qualification measure. After 2 years, when many courses end, they catch up and after just over 900 days they have the same probability of being employed as the non-participants. The share of participants in employment increases to about 75%. It remains at this level until the end of the observation period.

There is a longer investment effect, similar to the case of benefit receipt, where participants are less likely to be employed during the measure. This effect is up to almost 30 percentage points. This means that the probability of having employment during the qualification measure is systematically lower for the participants than for the non-participants. Again, this is an expected consequence of the participation. After about 1,100 days, the participants’ probability of employment significantly exceeds that of the non-participants. The difference increases up to 18 percentage points. Subsequently, the effect size decreases somewhat and is around 10 percentage points at the end of the observation period. The latter is not, however, due to the fact that the employment opportunities of the participants decrease, but because the non-participants are catching up due to their slightly increasing share in employment—a phenomenon that hardly can be called undesirable. The differences between the two groups are also statistically significant at the 1%-level at each instant starting from about 1,110 days.

Other impact analyses of qualification measures (Kruppe and Lang 2015, 2018) find also positive program effects of up to 20 percentage points. The effects for IHas are therefore not as pronounced as in other studies, but nevertheless positive, substantial, and statistically significant.

Another indicator of labor market success is the income generated by the participants. Table 3 shows the cumulative gross wages
of participants and non-participants at different times after the start of the observation period.

First of all, participants lag behind the non-participants, who gain wages when they take up a job again. After 720 days, the backlog of the participants amounts to almost 6,000 €. The gap closes in the third year when many of the participants finish the course and take up a job. The reasons cannot be precisely differentiated; it is a combination of the higher employment rate of the beneficiaries, their working time, and the hourly wage. After 1,440 days (= 4 years), the participants have already accumulated 4,800 € more compared to the non-participants. Over the next few years, their advantage increases; after 2,520 days or 7 years, they earned about 25,000 € more than non-participants. Additional calculations (not in the table) show that after almost 1,300 days or three and a half years, the participants also have as many days in employment as their statistical twins. After 1,500 days or about 4 years, the difference to the non-participants becomes statistically significant. At the end of the observation period, the advantage of the participants is about 350 days.

Overall, the quantitative impact analysis shows a clearly positive program effect, which also suggests a general positive assessment of the program implementation.

5. Conclusions

This article analyzes which evaluation designs are generally available and often applied in the context of labor market research for program evaluation, and why a mixed-method design often promises good insights. RCTs are often regarded as the ‘gold standard’ of quantitative research. They require a successful theoretical embedding and their practical implementation is complex and often inapplicable.

Therefore, there is no ‘magic bullet’ that serves all evaluation purposes. It is often useful to set up a mixed-method evaluation, i.e. to combine summative and formative components in order to exploit the strengths of the two approaches and to offset their weaknesses. This usually amounts to a combination of quantitative and—often preceding—qualitative elements. Quantitative impact analyses analyze program effects ex-post from a fixed point in time. Qualitative implementation studies are often carried out during the actual program and are intended to provide information for improvement in implementation practice. The price one has to pay for a mixed-method design is its complex implementation and the high required effort.

Using an example from the evaluation of ALMP, we illustrate this combination of different approaches. For this purpose, we use a qualification program of the FEA, which was tailored to jobseekers in the German unemployment insurance system. We show how formative and summative elements as well as quantitative and qualitative methods of an evaluation can be combined in order to obtain a comprehensive picture of the practical implementation and effects of such a program. The findings from the implementation analysis provide various insights: first, it becomes clear that detailed knowledge of the legal framework by those responsible in the employment agencies is essential for the improvement of such programs. The selection of the ‘right’ expert interviewees should be carried out with great care—the street-level is suitable because it deals with questions of the concrete application of the law.

Second, findings from the implementation part of the evaluation have stimulated further research projects on obstacles before and during participation in further training (Dietz and Osiander 2014; Osiander and Dietz 2016). They have also contributed to an ongoing political discussion on how to improve the financial situation of benefit recipients who participate in qualification measures. This shows how the mutual exchange between program managers and evaluators can, on the one hand, have a positive effect on the gain of scientific knowledge and, on the other hand, influence debates in the respective political field.

The results of the impact analysis show that there are substantial positive program effects: participants are significantly less likely to receive benefits than similar non-participants after three to three and a half years. Participants also have significantly higher probabilities of having a regular job, and they gain significantly higher cumulated wages than non-participants. This is in line with a recent meta-analysis by Card, Kluve and Weber (2018) that finds significantly positive medium and long-term effects for training measures, especially those that emphasize human capital accumulation. Overall, therefore, a positive conclusion can be drawn from the promotion of IFlaS.

Third, we identified some variables from the survey data that affect the participation decision and that are not available in administrative data. This holds true, for example, for the socio-economic background (measured by the number of books in the household) or attitudes toward life and work. Survey data thus offer added value in summative evaluations because they can help to better rule out unobserved heterogeneity. Since such evaluations are time-consuming and costly, it must always be considered to what extent the additional benefit justifies the additional costs.

Acknowledgements

I would like to thank two anonymous reviewers for their helpful comments. This article is dedicated to the memory of Michael Fergit, an excellent
DeGeval [Deutsche Gesellschaft für Evaluation e. V.] (2008) ‘Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments’, Journal of Economic Perspectives, 15: 69–85.

Bernhard, S. (2016) ‘Berufliche Weiterbildung von Arbeitslosengeld-II-Empfängern. Langfristige Wirkungsanalysen’, Soziler Fortschnitt, 65: 153–61.

Blundell, R. and Costa Dias, M. (2009) ‘Alternative Approaches to Evaluation in Empirical Microeconomics’, Journal of Human Resources, 44: 565–640.

Bonsen, M., Frey, K. A., Bos, W. (2008): ‘Soziale Herkunft’, in Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C., Walthier, G. (eds). TIMSS 2007—Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich, pp. 141–56. Münster: Waxmann.

Bosch, G. (2012) ‘Berufliche Weiterbildung in Deutschland 1969 bis 2009: Entwicklung und Reformoptionen’, in Bothfeld, S., Sesselmeier, W., and Bogedan, C. (eds) Arbeitsmarktpolitik in der sozialen Marktwirtschaft, pp. 94–111. Wiesbaden: Springer.

Bos, W., Lanke, E. M., Prenzel, M., Schippert, K., Valtin, R., and Walthier, G. (2003) Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster: Waxmann.

B A [Bundesagentur für Arbeit] (2009): E-Mail-Info SGB III vom 08.12.2009: Geschäftsanweisung SGB II Nr. 45 vom 08.12.2009. Internal Document. Bundesagentur für Arbeit. Nürnberg.

Caliendo, M. and Kopeining, S. (2008) ‘Some Practical Guidance for the Implementation of Propensity Score Matching’, Journal of Economic Surveys, 22: 31–72.

Card, D., Klueve, J., and Weber, A. (2018) ‘What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations’, Journal of the European Economic Association, 16: 894–931.

Carnevale, A. P., Rose, S. J., and Cheah, B. (2011): The College Payoff. Education, Occupations, Lifetime Earnings. Washington, DC: Georgetown University Center on Education and the Workforce.

Cartwright, N. (2007) ‘Are RTIs the Gold Standard?’, BioSocieties, 2: 11–20.

Cartwright, N. and Hardie, J. (2012) Evidence-Based Policy. A Practical Guide for Doing It Better. New York: Oxford University Press.

Cartwright, N. and Munro, E. (2010) ‘The Limitations of Randomized Controlled Trials in Predicting Effectiveness’, Journal of Evaluation in Clinical Practice, 16: 260–6.

CDU/CSU/SPD [Christlich-Demokratische Union; Christlich-Soziale Union; Sozialdemokratische Partei Deutschlands] (2013) Deutschlands Zukunft gestalten. Koalitionsvertrag zwischen CDU, CSU und SPD. 18. Legislaturperiode. <https://www.bundesregierung.de/Content/DE/_Anlagen/2013/2013-12-17-koalitionsvertrag.pdf?__blob=publicationFile> accessed 19 July 2017.

Chelimsky, E. (1997) ‘Thoughts for a New Evaluation Society’, Evaluation, 3: 97–109.

Cowen, N., Virk, B., Mascarenhas-Keyes, S., and Cartwright, N. (2017) ‘Randomized Controlled Trials: How Can We Know “What Works”?’, A Journal of Politics and Society, 29: 265–92.

Cromm, B. (2016) ‘Multiple and Mixed Methods Research for Economics’, in Cromm, B. and Lee, F. S. (eds) Handbook of Research Methods and Applications in Heterodox Economics, pp. 286–308. Cheltenham: Edward Elgar Publishing.

Dauth, W., Hirschmayer, E., and Ruh, F. (2008) Vergleichstypen 2008: Neufassung der SGB-III-Typisierung. IAB-Forschungsbericht. Nürnberg: Institute for Employment Research.

Deaton, A. and Cartwright, N. (2018) ‘Understanding and Misunderstanding Randomized Controlled Trials’, Social Sciences & Medicine, 210: 2–21.

DeGeVal [Deutsche Gesellschaft für Evaluation e. V.] (2008) Standards für Evaluation, 6th edn. Mainz: Deutsche Gesellschaft für Evaluation e. V.

Deutscher Bundestag. (2016) ‘Ausschussdrucksache 18(11)579. Ausschuss für Arbeit und Soziales. 18. Wahlperiode. <https://www.bundestag.de/resourse/blob/404012/90a131792f4e4f138f470867d2f72b67e0dd/materialzesamt_menstellung_arbeitslosenversicherungs-data.pdf> accessed 22 July 2019.

Dietz, M. and Osander, C. (2014) Weiterbildung bei Arbeitslosen: Finanzielle Aspekte Sind nicht zu unterschätzen. IAB-Kurzbericht. Nürnberg: Institute for Employment Research.

Dopp, A. R., Mundy, P., Beasley, L. O., Silovsky, J. F, and Eisenberg, D. (2019) ‘Mixed-Method Approaches to Strengthen Economic Evaluations in Implementation Research’, Implementation Science, 14: 1–9.

Eichhorst, W., Sesselmeier, W., and Yollu-Tok, A. (2008) ‘The Acceptance of Arbeitsmarktrechtformen am Beispiel von Hartz IV’, in Sesselmeier, W. and Schulz-Nieswander, F. (eds) Konstruktion von Sozialpolitik im Wandel. Impulse normative Ebenen, pp. 15–45. Berlin: Duncker & Humboldt.

Ferrig, M. and Osander, C. (2015) ‘Selektivität beim Zugang zu Weiterbildungsmaßnahmen. Die Bedeutung individueller und struktureller Determinanten am Beispiel der Initiative zur Flankierung des Strukturwandelns’, Schmollers Jahrbuch, 135: 279–306.

Fricke, H., Koch, S., and Kupka, P. (2012) ‘Evaluation von Arbeitsmarktpolitik—Methoden, Möglichkeiten, Grenzen’, in Bothfeld, S., Sesselmeier, W., and Bogedan, C. (eds) Arbeitsmarktpolitik in der sozialen Marktwirtschaft, pp. 213–229. Wiesbaden: Springer.

Fuchs, J. and Kubs, A. (2016) ‘Zuwanderungsbudget und Arbeitskräfteangebot bis 2050. Wie viele Zuwanderer benötigt Deutschland für ein konstantes Erwerbspersonenpotenzial?’, Wirtschaft und Statistik, Sonderheft 72016: 103–12.

Fuchs, J., Söhnlein, D., Weber, B., and Weber, E. (2017) ‘Stochastic Forecasting of Labor Supply and Population: An Integrated Model’, Population Research and Policy Review, 37: 33–58.

Harrison, G. W. and List, J. A. (2004) ‘Field Experiments’, Journal of Economic Literature, 42: 1009–55.

Head, B. W. (2010) ‘Reconsidering Evidence-Based Policy: Key Issues and Challenges’, Policy and Society, 29: 77–94.

Heckman, J. J., Ichimura, H., and Todd, P. (1997) ‘Matching an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program’, Review of Economics and Statistics, 64: 605–54.

Henry, G. T. (2015) ‘Comparison Group Designs’, in Newcomer, K. E., Hattery, H. P., and Wholey, J. S. (eds) Handbook of Practical Program Evaluation, pp. 137–157. San Francisco: Wiley.

Holland, P. W. (1986) ‘Statistics and Causal Inference’, Journal of the American Statistical Association, 81: 945–60.

IAB [Institut für Arbeitsmarkt und Berufsforschung] (2016) Aktuelle Daten und Indikatoren: Qualifikationspezifische Arbeitslosenquoten. <http://doku.iab.de/arbeitsmarktdaten/qualo_2016.pdf> accessed 19 July 2017.

Imbens, G. W. and Wooldridge, J. M. (2009) ‘Recent Developments in the Econometrics of Program Evaluation’, Journal of Economic Literature, 47: 5–86.

ISG [Institut für Sozialforschung und Gesellschaftspolitik GmbH/Lawazt Johann Daniel Lawazt-Stiftung] (2010) Implementationsstudie zur „Initiative zur Flankierung des Strukturwandels“—I. Zwischenbericht vom 02. August 2010. Stadtwerke Nürnberg.

Jefferson, T., Austen, S., Sharp, R., Ong, R., Lewin, G., and Adams, V. (2014) ‘Mixed-Methods Research: What’s in It for Economists?’, The Economic and Labour Relations Review, 25: 290–305.

Johnson, R. B., Onwuegbuzie, A. J., and Turner, L. A. (2007) ‘Toward a New Methodology Paradigm Whose Time Has Come’, Educational Researcher, 36: 290–305.

Johnson, R. B. and Onwuegbuzie, A. J. (2004) ‘Mixed Methods Research: A Research Paradigm Whose Time Has Come’, Educational Researcher, 33: 14–26.

Johnson, R. B., Onwuegbuzie, A. J., and Turner, L. A. (2007) ‘Toward a Definition of Mixed Methods Research’, Journal of Mixed Methods Research, 1: 112–33.

Kromrey, H. (2001a) ‘Evaluierung—Ein Überblick’, in Schoich H (ed.) Was ist Qualität. Die Entzauberung eines Mythos. Schriftenreihe Wandel und Kontinuität in Organisationen Band 6, pp. 31–85. Berlin: wvb Verlag.

Kromrey, H. (2001b) ‘Evaluierung—ein vielschichtiges Konzept: Begriff und Methodik von Evaluierung und Evaluationsforschung; Empfehlungen für die Praxis’, Sozialwissenschaften und Berufspraxis, 24: 105–31.
