Which regime works best in social welfare? Comparing outcomes of eight Dutch RCT experiments

Muffels, R.J.A.; Edzes, Arjen; Gramberg, P.J.; Rijnks, Richard; Venhorst, Viktor

Publication date:
2021

Citation for published version (APA):
Muffels, R. J. A. (Photographer), Edzes, A. (Author), Gramberg, P. J. (Photographer), Rijnks, R. (Photographer), & Venhorst, V. (Photographer). (2021). Which regime works best in social welfare? Comparing outcomes of eight Dutch RCT experiments. Web publication/site, Technequality.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 822330

Technequality

FINAL DOCUMENT VERSION 1.0: NOT FOR QUOTATION[1]

Work package 4: Reinventing the Welfare State, Deliverable 4.1

Which Regime Works Best in Social Welfare?
Comparing Outcomes of eight Dutch RCT Experiments

April 20, 2021

©Ruud Muffels, Arjen Edzes, Peter Gramberg, Richard Rijnks & Viktor Venhorst

Abstract

Current technological innovations (automation, robotization, digitization, AI, big data) may have adverse employment effects notably for the low skilled welfare recipients. They face reduced chances for getting access to secure and fairly paid jobs also while two in three lack the basic qualifications needed to acquire the lowest level jobs, let alone that also more than one third consider themselves unfit to work due to serious physical or mental health issues. Therefore, eight Dutch municipalities (Deventer, Groningen, Nijmegen, Tilburg, Utrecht, Wageningen, Apeldoorn-Epe, Oss) started in the fall of 2017 and early 2018 a two-year long unique randomized control trial (RCT) to test three alternative regimes for people on welfare in which more than 5,000 recipients participated[2]. The treatments set up were (1) exemption/self-management, that is exemption of the application obligations and rendering more trust and autonomy to the recipient for self-management, (2) intensive or tailored support, that is providing tailored and intensified counselling support to improve claimants' work and social participation opportunities (e.g., in education, training or volunteer work) and (3) earnings release, that is rewarding welfare claimants for finding work by allowing participants to keep a larger part of their earnings on top of their benefit (work bonus). The experiments share some features of participation and basic income approaches even though their design and implementation are rather different. We found no evidence that the alternative welfare regimes have reduced employment effects compared to 'workfare' regimes. In some municipalities we find small positive significant effects on parttime and fulltime employment and on people's self-efficacy, social trust and trust in caseworker' support. No significant positive effects were found on health and wellbeing. The use of field experiments for testing the outcomes of alternative welfare regimes provides new avenues for welfare state policy in an era of rapid technological and economic change.

[1] We are grateful to Marcel de Kruijk and Stefan van Wanrooij for providing outstanding research assistance. We thank Kirsten Blom-Stam for excellent research input on the survey outcome indices. We are obliged to all project leaders and representative officers of the municipalities for their collaboration and support. We are also grateful to the excellent collaboration with the Bureau for Economic Policy (CPB) and ZonMw, notably the supervisory committee, for their helpful comments and suggestions. We thank the Ministry of Social Affairs and Employment for financial support. The report is written for the Technequality project which got funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 822330. The experiment has been registered at the AEA-RCT Registry. The pre-analysis plan is available at this register. Information on questionnaires can be obtained from the authors.

[2] A similar ninth experiment in Amsterdam started much later in 2018 and will be ended only in 2022 for which reason this report does not include the results of the experiment in Amsterdam.

Key words: RCT, welfare, participation income, register and panel data, treatment models, social policy

CONTENT

1. Introduction: design, reason and research question

In the fall of 2017 and spring 2018 eight Dutch municipalities started a rather unique randomized control trial (RCT) experiment in which more than 5,000 social assistance beneficiaries participated. The experiment was aimed at testing the effectiveness of alternative support regimes for people on social assistance. Six of them (Deventer, Groningen, Nijmegen, Tilburg, Utrecht and Wageningen) were officially rewarded as experiment cities (after Government approval) and made use of experimentation article 83 in the Participation Act (former Social Assistance Act) permitting municipalities to implement a two-year lasting experiment from October 1, 2017 to October 1, 2019. The other two (Apeldoorn-Epe and Oss) were informally acknowledged as experiment cities while making use of the room in the existing act to launch similar field experiments but without being legally permitted to design an extra earnings release treatment during the experimenting period. Much later in 2018, a ninth also not formally acknowledged experiment was started by the municipality of Amsterdam that will last for three years up to 2022. In this paper we report therefore on eight out of the nine Dutch experiments.

The reason for the municipalities to start thinking about the experiments was, among other things, the need to improve the effectiveness of social assistance for reintegrating recipients into employment because of social welfare's poor records in getting people back into paid work (5 to 10% pro year). The aim was to compare the effectiveness of current activation policies induced by 'moral hazard' concerns with strict conditionality and tight monitoring and control ('workfare'), with a more relaxed compliance regime with less enforcement, extra support and positive financial incentives for people on social assistance (SA). The choice for the experimental treatments or interventions was partly based on recent theoretical insights from motivational psychology and behavioral economics about how to motivate and activate welfare recipients notably those with long stays in welfare vis à vis their job search behavior[3]. Basically, three treatments were distinguished, aimed at: (1) improving the self-management by relaxing the application obligations and rendering more trust and autonomy to the recipient (exemption group), (2) providing tailored counselling support to improve claimants' work and social participation opportunities, e.g. in education, training or volunteer work (tailored support group) and (3) rewarding welfare claimants for their active job search by allowing participants to keep a larger part of their earnings on top of the benefit through an extra work bonus (earnings release group). The researchers, were responsible for the evaluation of the tests, they conducted the experiments jointly with the municipalities and researched the effects on the participants' (re)employment chances but notably also on their perceived autonomy and self-management capacities and their subjective health and wellbeing.

The main research question was if and to what extent the three carefully designed treatments or social assistance regimes work better than the 'care-as-usual' treatment with respect to employment and health-

[3] In section 3, the theoretical underpinnings of the various treatments are briefly explained.

wellbeing outcomes and for what reason? The answers to this question are likely to give insight on which 'ingredients of the various treatments' work best and what we can learn from it for reform of existing welfare policies at local and national level to better deal with the upcoming challenges caused by health and economic crises, such as Covid-19 and GFC (Global Financial Crisis), but notably also the rapid pace of technological progress.

From the start of the experiment in 2017 the researchers have organized themselves in LOEP, the "National Consultation Group Experiments Participation Law", so as to closely collaborate and learn from each other but also to harmonize data collection and data handling, to implement the three experimental treatments in a more or less similar way and to conduct collaborative research. The Technequality project was one of the research projects conducted by the LOEP researchers. The various Universities were responsible for researching the outcomes of the local experiments but the municipalities, in their role as owner of the experiments, had a large say in decisions on the way the experiments were designed and implemented. Specifically, the municipalities had a strong say in the exclusion criteria for the target population, the choice and content of the various treatments, the access conditions for the participants and how the treatments will be implemented within the local bureaucracy. Nothwithstanding these differences in implementation, the local experiments had a lot in common such as their theoretical foundation, the RCT design, the linkage to the national register data, the survey questionnaires, the research methodology and the empirical analyses (cf. section 5.1). The CPB, The Bureau for Economic Policy Analysis, was asked by the Government to report on the overall outcomes up to October, 1 2019, when most experiments had finished, but only insofar it concerns the outflow to work based on administrative data from the national registers[4]. The LOEP researchers subsequently reported on the overall findings on outflow to work based on the same administrative data as well as on the so-called 'soft outcomes', that is on their self-efficacy, job search and health and wellbeing, using the three-wave panel survey data collected with the participants. They also worked closely together with the CPB to fine-tune the data handling and analyses and to be able to timely finalize the local reports which were published on May 31, 2020 (cf. list of local reports in the references summarized by Sanders et al. 2020), together with the CPB report (De Boer et al. 2020). After the release of the reports in June 2020, the researchers continued to work on the administrative and survey data. In this paper we report on the preliminary results from our first analyses up to June 2020 for the eight cities. For Tilburg, we updated the data to extend the time horizon from 16 months as in the CPB report to 24 months. We also linked the data for the non-official experimental cities Apeldoorn-Epe and Oss to the administrative data of the CBS[5].

[4] The experimental data containing information on the assignment of the participants to the various treatments were linked to the administrative microdata of Statistics Netherlands containing information on demographics, education, social assistance income, employment and earnings while covering the entire Dutch population (CBS).

[5] The reported employment effects in the local reports of Apeldoorn-Epe and Oss were based on local data on the administered reason of exit out of social assistance (to paid employment), whereas the researched employment effects in the CBS microdata were based on the level of earned wage income.

Relationship with Technequality[6]

Welfare is a social contract: workers may count on it to soften the consequences of unemployment, assuming hat they do whatever they can according to their ability to return to work. The social contract presumes employability, and as a consequence, reintegration into the labor market is usually pursued by providing counselling and employment support (including training, protected or subsidized employment), but also close monitoring of compliance behavior and monetary incentives for avoiding 'moral hazard' behavior. Theory, but also empirical and simulation studies predict that current technological innovations (automation, robotization, digitization, AI, big data) may for various reasons require us to rethink these assumptions. Simulation studies differ to the extent by which destruction of existing and creation of new jobs is forecasted. In the literature estimates of job losses vary rather depending on the assumptions and methods used but are between 47% (Frey &Osborne 2017) and 9% (Arntz et al. 2016). However, technological change might not just displace jobs but also generate economic growth and therewith employment (e.g in the service sector). Jobs and tasks change, because they become automated, robotic or digitized and therefore skills demand change. Due to 'skills-biased-technical change', the demand for high skills increases and for low-skills diminishes (Acemoglu 2006; Autor 2015). However, polarisation of employment might cause the share of mid-level jobs to decline and of jobs at a lower or higher level to increase (Autor 2015; OECD 2020). One would expect that because of automation the lowest-level jobs which are easily automated would be displaced but that seems not the case in the Netherlands (Fouarge 2016). Because of flexibilization and a growing service sector the share of low-skilled jobs remains high although many of these jobs lack security and fair pay. Reskilling might be the royal way for access to secure and fairly paid jobs but may not be an obvious response for people on social welfare. If potential workers cannot be reskilled, or there are no jobs such as in a severe economic or health crisis, the assumptions underpinning the social contract turn out wrong. This might then result in increased labor market segregation, that is polarization and dualization on the labor market and rising unemployment and inequality. Whatever the views are on the consequences of technological change for the employment of low-skilled people, either displacement or polarization and dualization, it may lead to pleas for a reconceptualization of the social contract and welfare state reform (Hemerijck, 2017; Morel et al. 2012; Groot et al. 2019). This can be pursued through reducing welfare conditionality and loosening the income-work linkage or through tailored employment support to ease access of welfare recipients to existing regular or newly created jobs. In the first scenario of weakening welfare conditionality, a participation or unconditional basic income (PI, UBI), or a negative income tax (NIT) has been proposed. In the second

[6] This section is derived from the written text of Work package 4 of the Technequality proposal (cf. Levels et al. 2018)

scenario of tailored employment support, options for providing a 'basic job' or a work-participation guarantee such as for youth come into the picture (Van Parijs 2004; Atkinson, 1996)[7].

Relation with participation and basic income approaches

Even though the work-benefit linkage is loosened, the local experiments are very different from PI/UBI approaches because they cover a very specific group instead of the entire population, and benefits are only partial unconditional while some reciprocity is still presumed. Participants in return for the wavering of the liabilities are expected to be committed to and put effort into the treatment to make it work. In some municipalities the participants also signed an agreement in which rights and duties were mentioned. Moreover, it is still presumed that a participant is seriously looking for work and that if a proper job is offered, he or she will also seriously consider it. However, there are also similarities with a basic income approach. First of all, the relaxation of the job search obligations will allegedly put less stress on people and allow them to search for better job matches and sustainable employment. Secondly, the shift from negative incentives through sanctions and penalties to trusting and rewarding people's efforts by means of a work bonus on top for people in the 'earnings release' condition will motivate people to increase their working hours. Thirdly, and partly for similar reasons, municipalities envisaged that the intensified and trust-oriented way of counselling in the various treatments will improve the motivation, health and well-being of the participants resembling again the expected positive outcomes as shown in some basic income experiments (Widerquist et al. 2013). After the preliminary results of the eight completed field experiments were published, some municipalities started to rethink their reintegration policies based on the findings which in their view provide clues for redesigning local welfare policies in the spirit of income or job guarantee approaches (see note 3).

Outline

The main focus of this report is on presenting the overall first results of the eight completed experiments, as reported in a joint preliminary report and some magazine and newspaper articles in May-June 2020. For a good understanding we will briefly explain the design and content of these local experiments in the main text but for a more detailed overview we refer to the local reports and the methodological Annex 1 where we will explain the content of and differences between the nine experiments in more detail. The focus of the main text will therefore be on the overall design and the outcomes with respect to outflow to work and health and wellbeing. Our task is threefold: 1. Compare the evidence on work and health and wellbeing outcomes across these eight field experiments during the experimental period, 2. Update the information on work outcomes, derived from the administrative data, from October 1, 2019 to March, 1 2020, and

[7] For further reading on European policies to deal with the challenges of the technology transition and the GFC and COVID-19 crises, see the Commission's documents: "A Bridge to Jobs - Reinforcing the Youth Guarantee" (11320/20), the European Skills Agenda for Sustainable Competitiveness, Social Fairness and Resilience (COM(2016)81 final) and the recently issued Recovery Plan (COM(2020)442 final).

therewith widening the observation window from 16-24 months to 22-29 months[8] and 3. Formulate some conclusions with respect to the research findings and provide some insights and lessons learned for labor market and social policies.

2. Policy context

2.1 Dutch social assistance and experimentation law article 83[9]

The Dutch social assistance (SA) scheme is a universal, 'last-resort' benefit scheme to which everybody who is not eligible anymore (or never was) for other income-replacement benefits such as unemployment insurance is entitled to. It provides a safety-net to people with insufficient income below subsistence level but who still might have some rest-capacity to work. People fully unfit to work are eligible for a disability benefit and those reaching pension age are eligible for a retirement pension. Social assistance is funded by taxes and provide a guaranteed minimum income for people without any other sources of income. Access to social assistance in the Netherlands is conditional on a number of specific obligations and compliance requirements to which recipients need to abide. The SA benefit is means-tested and access to it is subject to several 'willingness-to-work' tests. Recipients must provide the required information needed to assess the right on a benefit, attend invited gatherings with the caseworker, write a specified number of application letters at regular intervals, apply or register for work with temp-agencies and need to accept any available job offer instead of only those that fit their experience and skills. The level of the SA benefit is linked to the level of the statutory minimum wage: a single person household on welfare receives 70 per cent of the minimum wage, whereas a couple, due to economics of scale in consumption, is entitled to 100 per cent of the minimum wage[10]. The statutory minimum wage can be considered as the Archimedean point in the social security system because all social security benefits, ranging from unconditional child allowances, unemployment insurance, social assistance to public old-age benefits are linked to the minimum wage. As of 2001, the payment of SA benefits became decentralized from the central to the local government with the aim of enhancing efficiency in the administration and implementation. Pursuing a strategy of activation appear to have raised efficiency significantly (Broersma et al., 2013). From January 1, 2015 on, the implementation of the Social Assistance Act (from then on called the Participation Act) became the full responsibility of the municipalities. One of the accompanying changes is that people on welfare now have an even more strict duty to accept work, even if it does not fit their skills or occupational background and that recipients must be willing to commute up to three hours. Municipalities got however more leeway to determine the requirements for access to and receipt of a full social assistance benefit, varying from

[8] In this draft report updated results are only presented for five cities Tilburg, Wageningen, Oss, Apeldoorn and Epe. Updated results for the other cities were not yet available.

[9] This section is partly derived from Groot, Muffels & Verlaat (2019). For further information see also Verlaat et al. (2020).

[10] For people living in larger households with five or more adults, the maximum benefit can be at most 190 percent of the minimum wage.

engaging in volunteer work to insertion into all sorts of unpaid societal useful activities. The regular way of supervision is rather strict and hinges upon so-called 'workfare' principles, meaning the use of benefit penalties if the strict application obligations are not met [the 'stick'] combined with benefits at some distance of the minimum wage to provide financial incentives [the 'carrot'] when people move quickly into paid work. Failure to fulfil the rather strict obligations may result in sanctions or benefit cuts at the discretion of the municipality that may curtail benefits from 30 to 100% for a maximum of three months. In day-to-day practices, municipalities make in only 10 to 15% of the cases use of their authority to implement benefit cuts or sanctions.

In addition, caseworkers have leeway to exempt recipients from the strict application obligations if there are good reasons for it such as having serious social or (physical or mental) illness problems because of which a person is (temporally) unfit for work. Since municipalities are fully responsible for the payment of SA, they for budgetary reasons and to a varying degree became stricter in testing the legal requirements for (full) eligibility to SA benefits and detecting signs of benefit fraud ('legality test'). When beneficiaries start to work and earn additional income, they may keep (not withdrawn from the benefit) up to 25% of the extra earnings for a period of six months after enrolment in the current scheme (depending on the local regulations of the municipality). After these six months 100% of the extra earnings are taxed away by reducing the benefit with the same amount. This provide little incentives for people longer on SA to be engaged in part-time work. Only when they find (nearly) fulltime work for at least 27 hours a week at the level of the minimum wage, because of which their earnings exceed the level of the SA benefit for a single person (that is 70% of the minimum wage), the payment of SA is stopped and they may of course then keep 100% of their earnings. One of the treatments in the experiment is called the 'earnings release' group because according to the experimentation law article 83, the earnings withdrawal rate may be reduced from 100% after the first 6 months of welfare residence to 50% so that participants are allowed to keep 50% of the extra earnings which are not withdrawn from the benefit during the entire two year-experimenting period. However, the existing maximum of 202 euros per month is retained.

The monitoring and testing of the recipients' compliance behavior to these requirements has been put into the hands of the local welfare offices and notably the caseworker responsible for the implementation of the social assistance scheme. With respect to the monitoring and testing of welfare recipients' behavior, a distinction needs to be made between the 'legality test' and the 'expediency test'. The legality test concerns the assessment of the right of access to the welfare benefit and is in the majority of the cased laid into the hands of specialized benefit payment managers. The expediency test concerns the assessment of which intervention is likely to work best for the recipient of welfare in the current labor market context which task is handed over to the caseworker. He or she also acts as the day-to-day contact person for the recipient. For successful reintegration of the beneficiary into paid work the caseworker has within the Participation Act access to a variety of tools or trajectories which can be put in place to speed up the return to parttime or fulltime paid work. This ranges from activation tools such as following language or application courses, education or training programs and volunteer work to employment supervision or support and employment subsidies (e.g., for partially disabled persons). These tools can be used only insofar they put no limit on the availability of the recipient for the labor market when a job offer would come by, making it in practice very difficult to let people for example follow education or training courses for a longer period of time. This will allegedly put less stress on people and allow them to search for better job matches and sustainable employment. The idea of rewarding instead of penalizing is reflected in the reduced deduction or withdrawal rate of extra earnings. The municipalities expected that the more relaxed and rewarding way of treatment (reduced conditionality) will improve the motivation, health and well-being of the participants.

2.2 Local experiments: target population and sample sizes

In the Netherlands, at the time of the start of the experiment, 4% of the population of 15-65 years of age lives on social assistance of which 20% (92,000) were inhabitants of the ten experimenting cities. The number of participants in the ten social assistance experiments was about 5,200 as shown in Table 1, which is about 6.6% of the target population of people on welfare in these cities. Table 1 shows the number of participants but also the size of the total and the target population of social assistance recipients in each city, of which the latter is lower than the first because some specific groups are excluded from the experiment (such as disabled, nearly-retired and homeless people). Taking the experiments on a nation-wide scale they resemble one of the biggest experiments in welfare, worldwide, even slightly larger than the basic income experiments in the US or Canada in the 1970s.

Table 1: Social assistance (SA) total and target population, projected and realized number of participants, ultimo 2017

Municipalities			Participants	
RCT Experiments	SA Total population	SA Target population	Projected	Realized
Official (article 83)	N=	N=	N=	N=
Deventer	3117	3117	1000	383 (1584)[1]
Groningen	11000	8744	700	890 (8192)[1]
Nijmegen	8000	5000	400	304
Tilburg	8200	6000	800	780
Utrecht	12250	8338	900	752
Wageningen	800	800	300	410
Subtotal	43617	31761	4100	3519
Informal				
Oss	2225	1500	300	344
Apeldoorn-Epe	4300	3425	540	559
Amsterdam	42000	40000	5250	808
Subtotal	48525	44925	6090	1711
Total	92142	76686	10190	5230

Note: 1) Groningen and Deventer used pre-randomization of the target group after which people were invited. In the end 890 and 383 welfare recipients respectively were registered as participant.

Source: Sanders et al. (2020) for Article 83 experiments and Muffels & Gielens (2020) for the informal experiments.

3. RCT-experiment: implementation of the treatments[11]

3.1 Sample and recruitment procedure

The experiment in all eight cities was set up as an RCT (random control trial) experiment (Athey & Imbens 2017; Deaton, 2018)[12]. Only social assistance recipients fulfilling certain criteria were selected. The criteria to exclude people varied somewhat across the experiments. In most cities people entitled to a disability benefit or who during the experiment become retired were excluded, but also people living in care institutions or without a home address and migrants with insufficient proficiency in Dutch except in Utrecht where questionnaires were translated into various languages. In five cities (Tilburg, Utrecht, Nijmegen, Groningen and Deventer) youngsters up to 27 years of age were excluded because they have a distinct support regime. Further details are provided in Annex 1A. Overall, about 20% to 30% of the social assistance population was excluded. In Groningen and Utrecht, the sampling was done in one round, but in quite a few cities, such as in Deventer, Nijmegen, Tilburg, Wageningen, Oss, Apeldoorn and Epe the sampling was split in two rounds. In the first round, current recipients of social assistance fulfilling the inclusion criteria were selected whereas in the second round, new inflow of welfare recipients was invited to participate. It means that in the cities with two rounds of sampling, the maximum duration of the experiment for those belonging to the new inflow was not 24 months but 14 or 24 months. In Table 1 we show the timing of the experiments and the various inflow moments across the 8 experiments.

Table 1: Inflow and timing of the 8 local experiments

City	Deventer	Groningen	Nijmegen	Tilburg	Utrecht	Wageningen	Apeldoorn-Epe	Oss
Inflow moment	01-Oct-17	01-Nov-17	01-Dec-17	01-Dec-17	01-Jun-18	01-Oct-17	01-Oct-17	01-Oct-17
	01-Feb-18		01-Apr-18	01-Dec-17 to		02-Oct-17 to	02-Oct-17 to	02-Oct-17 to
	01-Jul-18			01-Jul-18		01-Sep-18	01-Jul-18	01-Jul-18
End of experiment	01-Oct-19	31-Oct-19	31-Dec-19	01-Oct-19	31-Dec-19	31-Dec-19	01-Oct-19	31-Dec-19
Duration	14-24 mths	24 mths	21-25 mths	22 mths	19 mths	27 mths	24 mths	27 mths

[11] This part of the paper is derived from Sanders et al., 2020 in which the researchers organized in LOEP have summarized the results for the six article 83 experiment cities.

[12] In Groningen and Deventer, the experiment was designed using a Zelen's design (Adamson, J. et al., 2006). In a Zelen's design experiment, the target population is randomized prior to the invitation to participate, the reverse of a more conventional RCT design.

Source: LOEP, 2020. Partly derived from De Boer et al. (2020).

Recruitment of participants

In some cities the invitation to participate went out to a limited number of people on social assistance whereas in other cities all people meeting the inclusion criteria got an invitation. In all cities, except in Groningen and Deventer, potential participants were randomly assigned to the three intervention groups after being registered and admitted to the experiment. In Deventer and Groningen, the random sampling was done before the invitations went out, but only in Groningen welfare recipients were invited for a specific intervention group, in a so-called Zelen's design (Adamson et al., 2006). In Deventer people had to register first after which they were informed about the assignment to a specific group. The implications of the various designs will be discussed later. People are free to withdraw from the experiment after they have been admitted and registered. The withdrawal rate differs between the cities and ranges from 5% (Groningen) to 30% (Tilburg)[13]. The people who did not accept the invitation to participate belong explicitly or implicitly to the group of refusers to the request to partake in the experiment. The group of participants that in the end was admitted to the experiment might be a selective group of people but also the group of non-participants that explicitly or implicitly refused to partake might be a selective group compared to the target population of social assistance recipients. Balance tests were performed to compare the composition of the various experiment groups with the comparison groups, that is the control group and the (randomized) group of non-participants (see methods section).

3.2 Implementation of the treatments

Apart from 'care-as-usual', that is the control group with the regular treatment (CG), in which people are subject to the ruling application and re-integration obligations at the start of the experiment in 2017 (in social policy labelled as the "stick and carrot' or 'workfare' approach), there are three different treatments which are also distinctly implemented in the nine cities (see also Annex 1)[14]:

A. *The exemption or self-management group.* In this exemption regime, participants are expected to help themselves in finding paid or unpaid work (volunteer work) or other participation opportunities such as education or social and health support. Two distinct regimes can be discerned. In four of the nine experiments people could decide themselves whether they want to receive support and have contact with a caseworker (Groningen, Nijmegen, Utrecht, Deventer). They got more or less full autonomy in deciding on the content of the support, even when this means that they don't request for support at all. In two of these cities (Deventer, Nijmegen) this group was additionally entitled to the extra earnings release scheme of article 83 (reduced withdrawal rate of 50% up to the maximum of 202 euros a month). The number of contacts with the caseworker was therefore

[13] Table 2 in section 5, provides more information on these withdrawal rates.

[14] In Groningen in addition, a so-called customized treatment group was designed in which people could choose in which of the three treatments they want to partake.

low and mostly once a year. In the remaining five cities (Tilburg, Wageningen, Apeldoorn, Epe, Oss), participants in this group were expected to learn how to help themselves: that is through self-management. The caseworkers received a training from a training Centre for supporting and learning clients on self-management. In both regimes, participants were exempted from the existing application and re-integration obligations. In Tilburg, they got on top of the earnings release an additional work bonus of nearly 200 euros per month (paid out half-yearly or yearly) when they find fulltime work (granted twice at maximum).

B. The *counselling or intensive support grou*p. Participants in this treatment group get extra support through tailor-made supervision and intensive mediation. The intervention is designed in all nine cities, but in Nijmegen, Deventer and Tilburg combined with earnings release. The treatment is tailored to the personal needs of the participants, meaning that participants have a say in co-deciding on the content of the treatment and the tools or trajectories available for support. The range of tools and trajectories has been extended and customized to the needs of the participant. The participants not only have more frequent contact with their caseworker (5/6 to 10/11 times a year) but the focus has also been on improving the quality of contact and support through putting more trust in people and paying more attention to the personal situation and needs. In Nijmegen 'group-coaching' was provided and in Tilburg and Groningen dedicated caseworkers were hired on the external labor market. In Groningen also people having experience as benefit recipient became part of the team of caseworkers. The leading principle also changed as framed by the question put to the participants "What do you need?" instead of what the caseworker perceives as needed. This also means that the 'intrinsic motivation' of the participant got a larger weight in the offered support. In this way the recipient has more say and autonomy about the treatment that is more tailored to his or her needs.

C. *The earnings release group (only permitted for the six article 83 experiment cities).* In three cities (Groningen, Utrecht, Wageningen) the intervention consists of a single treatment whereas in the other three cities (Deventer, Nijmegen, Tilburg) it was combined with intensive or tailored counselling and support. For participants in the six cities a more generous earnings release scheme is granted when they find paid work or work more hours; instead of the standard withdrawal rate of 75 per cent (during the first 6 months after enrolment in welfare) or 100 per cent (after 6 months of stay in welfare) of the extra earnings, a reduced rate of 50 per cent was granted, however, up to the same maximum of 202 euros per month. The participants in this group are to some extent exempted from the application and re-integration obligations except in Wageningen where this group got the 'care-as-usual' treatment. The number of personal contacts with the caseworker is increased up to 5 or 6 times a year when it is combined with intensified support. In Tilburg, the participants get the earnings release but on top of that an extra work bonus of about 200 euros per month for people moving to fulltime paid work and therefore moving out of social assistance.

Differences in design of the treatments

Table 2 below shows the differences in the composition of the various treatments across the eight experiments. Some cities such as Groningen, Utrecht and Wageningen but also Oss and Apeldoorn-Epe have only single treatments whereas Deventer, Nijmegen and Tilburg have combination treatments in which exemption and a work bonus is combined. This makes it more complicated, though not impossible, to disentangle and assess the pure effect of the separate components of the combined treatment. The Table does however not show the differences in the content of the treatment of people in the control group providing "care as usual" across the various cities. The information provided to the researchers by the municipalities about the 'care as usual' treatment shows quite some differences in the content, work methods and support tools offered. The content of the support in the control group is dependent on the way the municipality has interpreted the implementation room in the current Participation Law, the way the municipalities deal with the local labor market conditions and the way of supervision and support by each individual caseworker. With respect to the experimental groups, Table 2 shows that there are quite some differences in the design of the experiments across the cities but also quite some overlaps.

Table 2. Differences in design of the treatment groups across the eight experiments

Treatment group / city	Deventer	Groningen	Nijmegen	Tilburg	Utrecht	Wage-ningen	Oss	Apeldoorn-Epe
T1	A + C	A	A + C	A + C	A	A	A	A
T2	B + C	B	B + C	B	B	B	B	B
T3	D + C	C		B + C	C	C		
T4		E (A,B of C)						

Note: A=Exemption/Self-management; B=Intensive and tailored support; C=Earnings Release or work bonus; D=e-support with an APP; E=customized support (A, B of C). T1 to T4 = treatment 1 to 4
Source: LOEP, 2020; Adjusted from De Boer et al. (2020).

Groningen, Utrecht and Wageningen, all have implemented only single treatments A, B and C. With Oss, Apeldoorn and Epe they have in common to implement the single treatments A and B. Deventer, Nijmegen and Tilburg share the combination treatment of exemption A with a work bonus C although the level of the work bonus differs across the three cities. Tilburg provides an extra work bonus on top of the bonus permitted in the experiment. For these reasons the researchers decided to first analyze the effects of each treatment cities separately instead of jointly, even though it will limit the power of the analyses.

Reference groups 1 and 2

In all cities there is at least one additional comparison group of people receiving social assistance and receiving the regular treatment but who are not participating in the experiment. The first comparison group consists of people belonging to a randomized sample of non-participants that is drawn from the target population before the sampling started of the experimental groups (Refgroup 1: Groningen, Deventer, Utrecht). The second group consists of all non-participating respondents belonging to the target population, but who either did not receive an invitation, did not respond to it or did not accept the invitation to take part in the experiment (Refgroup 2: all other cities). This second reference group is not randomized. As a result, it might be a selective group of beneficiaries and thus it is not very well suited for causal inference. However, after correction for selectivity, it is possible to use this group to check for differences with the control group and therewith provide evidence on the external validity of the results.

Control group

In most municipalities the *regular treatment* ("care as usual") is executed by a team of caseworkers who are jointly responsible for the monitoring and counselling support of the stock of recipients. Before the experiment started the caseload of each fulltime caseworker was rather high, and ranges in practice between 100-400 people, dependent on the size of the city and welfare population, the labor market conditions and the local organizational context. In the experimental treatments the caseload has in most municipalities been reduced compared to the caseload the caseworker had before in the regular treatment, dependent on which city and which treatment. The caseload for the exemption group and earnings release group was generally much less reduced than for the intensive support group[15].

The *experimental treatments* were conducted by dedicated caseworkers who were assigned to one specific treatment. The caseworkers were for practical reasons not randomly selected out of the entire pool of caseworkers in a city although that would have been the preferred way in an RCT design. In Tilburg, all caseworkers were externally hired to support the participants in the two-year experiment, whereas in Groningen only for the intensive support group. In other cities caseworkers were selected after the group of caseworkers was asked to express their interest in partaking the experiment. In Wageningen, because half of the SA population took part in the experiment, caseworkers were as much as possible dedicated to one treatment but also had to support SA recipients who were not participating. The use of dedicated caseworkers assures that the way the specific treatment is implemented remains the same over time and will not be mixed up with another treatment blurring the content of the treatment and results.

4. Theory and hypotheses

Review of literature

To arrive at plausible and testable hypotheses a brief literature review, covering the economic and psychological literature, was conducted to explore the size range of the effects of the treatments implemented in the experiment (cf. Muffels & Gielens, 2019). The review studies by Card et al. (2010, 2015) show that the effects of activating labour market policies averaged 11.5% over all medium-term experimental and non-experimental studies (12 to 24 months) whereas the average percentage for

[15] To give a few examples, in Tilburg, Wageningen, Oss and Apeldoorn the caseload was reduced to 40-60% for the intensive support groups and 60-75% for the exemption or earnings release group.

experimental studies is found to be much lower, 5.6%. The effects pertain to the average effects of assistance in job search, work creation in the public sector, monitoring and control or sanctions, subsidies for private employment and training. The review concerns both, people entitled to unemployment insurance and people entitled to social assistance. The review also shows that there are considerable differences in the calculated effects of counselling programmes depending on the type of research, but also depending on the type of support offered. Active counselling, intensive employment support and job coaching towards reintegration into employment emerge to be the most effective way in bringing people back into paid work, *i.e.*, more effective than training or support in searching for work (see also Bolhaar et al. 2019). Especially for people with a large distance to the labour market, intensive mediation or support is needed which appears effective. Intensive supervision or coaching of people with a psychiatric disability works effectively to lead these people to work, according to international research (Marshall *et al.*, 2014; Bond *et al.*, 2015). The effects of this intervention are greater than the effects of 'workfare-based' interventions and vary between 10 and 25% (on average 15%). It was therefore assumed that the experiments will yield an absolute increase of the employment rate between 5 and 10%-points (for the exemption and earnings release treatment) to 15-25%-points (for the treatment of tailor-intensive supervision). This would imply that the treatments would on average create an additional effect on the exit chances into work of about 10%-points (7.5%-points for two groups and 20%-points for one group). It means that if the current exit rate is on average 20% over a period of two years, the rate will rise due to the interventions to 30% (50% increase). For the other outcome measures (physical and mental health and wellbeing) no clear effect estimates could be deducted from the literature because the evidence was overall rather mixed. A recent scoping review study in the Lancet focused interestingly on interventions similar to basic income and concluded on the basis of 27 studies (RCT and quasi- experimental studies) that: *"Evidence on health effects was mixed, with strong positive effects on some outcomes, such as birthweight and mental health, but no effect on others. Employment effects were inconsistent, although mostly small for men and larger for women with young children. In conclusion, little evidence exists of large reductions in employment, and some evidence suggests positive effects on some other outcomes, including health outcomes"* (Gibson et al. 2020, p. 165). In the conclusion of the article the authors noted that: "*Many studies reported stronger effects on health and educational outcomes in more disadvantaged groups"* (p.173*).* On this note, no further evidence was given. Recent research has been conducted by Bigotta *et al.* (2018) on a similar Swiss RCT-experiment also covering people in social assistance (Lausanne) and also lasting for two years (2015-2017) with intensive tailor-made mediation. That study shows an effect of 9% over the two years period, indicating that the presumed 10% for the Dutch experiments is in a similar range.

Theoretical background

The theoretical background of the experiments was formed by a combination of behavioral economic insights and insights from social-psychological cognition and motivation theories. The first was focusing on the impact of positive financial incentives, trust, autonomy and freedom of choice and the second on the impact of reciprocity, self-management and intrinsic motivation. In both strands of the literature the posited impact of these behavioral factors was on job search behavior and people's health and wellbeing. Four insights, which will be briefly explained below, have stood at the basis of the experiment.[16]

- The first insight concerns recent research findings about the influence of poverty on the 'mindset' or mental state of people. Research in this relatively new area of research shows that (financial) scarcity and stress due to poverty reduce people's cognitive resources (Mani et al. 2013; Mullainathan & Shafir, (2013). If financial scarcity and fulfilling social assistance obligations take up a large part of people's cognitive resources, there is little room or autonomy left for important and cognitively challenging tasks, such as retraining for another job, maintaining one's social network or actively searching for paid work.
- The second insight comes from behavioral economics and deals with the influence of labor market and welfare institutions and the implicit values they represent on people's job search and participation behavior. Underlying values are, for example, reciprocity or 'tit for tat' and trust. Reciprocity means that individuals reward personal support and trustworthy treatment, for example by making an extra effort (positive reciprocity), while doing the opposite if they are treated badly or treated on the basis of mistrust (negative reciprocity) (Fehr and Schmidt, 2003). Negative incentives, such as financial penalties and benefit sanctions, are not necessarily the best way to encourage people to cooperate and to comply to the rules. Economic and sociological research shows that although negative financial incentives have significant positive short-term re-employment effects (van Abbring et al. 2005; van der Klaauw & van Ours 2013; Hullegie & van Ours 2014), the longer-term employment effects are rather unclear, for example because of job entries with low job security or a poor job match (Koning, 2009; Knoef & van Ours, 2016). Findings from experimental economy also shows that, in exchange for the conveyed trust, people are extra motivated and do their best for their task, and thus reward those who trust them (the trustor) with extra effort and performance. In this way, trust leads to feelings of positive reciprocity and therefore to sustained commitment and increased job search effort, employability and productivity (Bohnet et al. 2001).
- The third insight comes from psychological motivation theory, which teaches that extrinsic stimuli can crowd-out intrinsic motivation. Research also showed that intrinsic motivation can be enhanced by offering an activity as a choice (autonomy) rather than a means of control (Frey and Jegen 2001). Self-determination theory (Deci and Ryan, 1985) states that intrinsically motivated people engage in an activity because they find it enjoyable and interesting, demonstrating greater effectiveness and persistence in their behavior and improved well-being (Ryan et al., 1997). This theory also states that trusting or putting confidence in people creates a sense of self-determination, which in turn has a positive effect on job-seeking behavior and sustainable employment (Fishbein & Ajzen 2010).

[16] This theoretical section is for a large part derived from Groot, Muffels & Verlaat (2019).

- The fourth and last insight is about capacitating people and providing 'freedom of choice' which stems from Sen's 'capability theory'. Within this theory, capabilities are the options people have or are offered, to be or do the things they have reasons to value most for their own lives. In this way, people have or are given opportunities or options that enhance their well-being (Sen, 1999; 2004). Both treatments, the exemption/self-management but also the tailored support treatment, through learning people to be or become self-reliant, might render people freedom of action, autonomy and choice while increasing their capability set (set of opportunities or choices). According to Sen’s framework, if people have different starting positions and therefore unequal ‘freedom of choice’, it is justified and necessary to treat them unequally and provide for extra support (uneven help in uneven conditions).

These insights have shaped the design of the experiments and the definition of the treatment groups. The self-management treatment was inspired by ideas on the role of autonomy and self-determination and intrinsic motivation for behavior, whereas intensified tailored support was inspired by ideas on the role of reciprocity in rendering trust and confidence and creating opportunities for people or freedom of choice. Lastly, the design of the earnings release group was inspired by ideas on the impact of rewarding and providing positive (financial) incentives on job search behavior. Inspiration for these behavioral insights as a tool for social policymaking has further be found in Nobel laureate Richard Thaler’s ideas on nudging: that is, encouraging people to behave in their broad self-interest by providing the appropriate positive reinforcements and by doing so shaping their choice architecture and therewith indirectly influencing their behavior (Thaler and Sunstein, 2008). By providing alternative behavioral choices or ‘nudges’ to people that might serve their own interests better, irrational choices may be avoided because of which it might be a more effective way to evoke preferred behavior. Learning people to become more self-reliant in the exemption treatment, a work bonus as a stimulus for searching part-time employment and tailored support with a focus on intrinsic motivation for job search and providing tailored choice options to people such as volunteer work, education and training opportunities, and (mental) health and social support, might provide alternative choices and prevent irrational choices by the participant.

Hypotheses and policy expectations

These insights have resulted in five main hypotheses dealing with the effects of the various treatments on employment and health and wellbeing:

- H1: The relaxation of compliance rules and of imposing benefit sanctions will reduce the level of stress and therewith free recipients’ mindset for active job search. Because of increased job search efforts more people on welfare will in the end find paid work (exemption/self-management);

- H2. Giving people more autonomy while putting more trust in them might evoke feelings of positive reciprocity through which extra effort will be put in searching for (un)paid work (exemption; intensive support);
- H3. Providing more tailored support to let people be engaged in activities which fit their intrinsic motivation best, are likely to increase exit into (un)paid work (intensive support);
- H4. Putting more trust in people by providing more autonomy and free choice, or providing intensified, tailored support, will lead to increased levels of subjective health and wellbeing (exemption; intensive support).
- H5. The earnings disregard or work bonus in the earnings release treatment might increase the outflow to part-time work because of the positive financial incentives associated with working more hours (earnings release or work bonus).

Policy expectations

The municipalities wanted to test whether alternative regimes would work better than the current rather strict monitoring and control ('workfare') policies who were believed to be not very effective nor in bringing people back to paid work or to unpaid work, nor in improving the level of social participation and of health and wellbeing. More relaxed alternative regimes (exemption) and intensified support were expected to be more effective for people with long welfare spells and low employment chances because there is need for in-between steps between work and non-work before they will be able to move into paid employment. In their eyes, research is needed to test alternative support regimes which might be better equipped to activate people by focusing on people's 'intrinsic motivation', less enforcement, more autonomy (exemption), rewarding financial incentives (earnings release) and intensified tailored support.

5. Data, outcome measures and methods

5.1 Data

The researchers designed and set-up the local experiments in close collaboration with the municipality. With respect to data collection: (1) they collected the local data on the design and implementation of the experiment within each municipality, that means the target population of welfare beneficiaries; the selection and assignment of each registered participant to the experimental groups; the identification of the withdrawals after registration by treatment group and the way of implementation of the treatments within the municipal bureaucracy; (2) they designed the three-wave panel survey data with the participants containing the information on the various 'health and wellbeing' outcome indicators and (3) they linked the experiment and survey data to the administrative register employment data, as contained in the microdata register files on welfare, income and employment of Statistics Netherlands (CBS)[17]. The linkage of both, the experimental and survey data, to the register data permitted to analyze the employment and health and wellbeing outcomes in a similar way by estimating identical empirical treatment models. It also allows to compare the results across the various municipalities.

Sampling data

The target population consisted of the stock of recipients of social assistance recipients being registered as such in a particular time frame just before the start of the experiment. These sampling time frames differ across the various cities. It ranges from some ten weeks in Utrecht to one month in cities such as Tilburg, Nijmegen or Groningen.[18] The number of participants in each experimental group ranges across the various cities from 93 to just more than 200. Reference group 1, the randomized comparison group of non-participants, is more or less of similar size (205 in Deventer and 146 in Groningen). But, reference group 2, consisting of all non-participating social assistance recipients such as in Nijmegen, Tilburg, Wageningen, Oss and Apeldoorn-Epe, is much larger than any of the three treatment groups. In Table 2 we show some sampling data characteristics, that is the number of participants in the experimental treatments in the various cities, the number of people partaking in each wave of the survey and the number of people that responded (experiment and surveys) or has withdrawn after registration. The Table shows that the percentage of people that dropped from the survey because of withdrawal, moving to another city, detention, retirement or death varies between 5% (Nijmegen) to 30% (Tilburg). In Tilburg quite some people have said to refuse to partake in the end because of their current bad health situation. A rather small fraction of people in all cities seem to have withdrawn for reasons of being assigned to the control group or a treatment group they did not like (see the local reports for more information). Most people withdrew shortly after being registered and within one year after the start of the experiment. People who left the scheme for having found a paid job remain participants and under supervision of the coaches. In Deventer they got a special questionnaire to be filled in after they had left for work. The Deventer response rate was on average lower for this special questionnaire than for the standard one (45% instead of 68% for the last survey). It makes clear why response rates drop in some cities for the second and third wave whereas for the first wave response rates turn out to be rather high for these kind of surveys among low-income groups and ranging between 80% to close to 100% (Nijmegen, Deventer, Utrecht, Groningen). In Nijmegen, partaking in the survey was conditional for partaking in the experiment.

[17] The local experiment data were first linked to the local population registry file (*GBAPERSOONTAB*). It contains information on age, gender, and migration background. Next, the data were linked to the education registry to obtain the highest education level (*HOOGSTEOPLTAB)*. For spell duration at the start, we used the social assistance benefit file (*BIJSTANDUITKERINGTAB)*. Household composition data were available in *GBAHUISHOUDENSBUS*. Finally, information on hours of work and earnings were obtained from the social security registers (*SPOLISBUS)*. Data were used for three cities on the first three quarters of 2019. For five cities we used updated data to June 2020.

[18] More detail can be found in the various local reports (cf. References: list of local reports)

Table 3. Sampling data characteristics: number of participants and withdrawals by treatment, withdrawal and response rate by survey (WR/RR)

Cities / Treatment number	T0	T1	T2	T3	T4	T5	T6	T7	R1	R2	WR/RR
Treatment type	CG	A	A+C	B	B+C	B+D	C	E	Ref1	Ref2	%
Official experiments											
Deventer (Dev): started N=383	**93**		**113**		**93**	**84**					
Invited	397		376		384	427			205		
Stopped	0		15		14	13					*11.0%*
Survey response - Baseline survey	95		105		88	79					95.8%
Survey 1	44		78		42	39					53.0%
Survey 2 (excl. stopped)	72		79		42	41					68.6%
Groningen (Gro) N=890	**222**	**183**		**144**			**153**	**188**	**146**		
Withdrawal/stopped during experiment	25	19		40			24	5			*12.7%*
Survey response - Baseline survey	222	183		144			153	188			100.0%
Survey 1	202	168		118			136	164			88.5%
Survey 2 (excl. stopped)	200	160		107			130	163			97.8%
Nijmegen (Nij) N=304	**94**		**110**		**100**					**5664**	
Withdrawal/stopped during experiment	8		5		2						*4.9%*
Survey response - Baseline survey	122		122		122						100.0%
Survey 1	122		122		122						100.0%
Survey 2 (excl. stopped)	122		122		122						100.0%
Tilburg (Til) N=780	**202**		**193**	**191**	**194**					**4606**	
Withdrawal/stopped during experiment	61		46	75	53						*30.1%*
Survey response - Baseline survey	142		143	88	134						84.4%
Survey 1	102		139	115	133						81.4%
Survey 2 (excl. stopped)	74		122	86	123						74.3%
Utrecht (Utr) N=752	**188**	**189**		**188**			**187**		**198**		
Withdrawal/stopped during experiment	8	13		15			12				*6.4%*
Survey response - Baseline survey	173	175		170			168				91.2%
Survey 1	155	151		141			147				79.0%
Survey 2 (excl. stopped)	147	148		133			142				81.0%
Wageningen (Wag) N=410	**93**	**106**		**98**			**113**			**456**	
Withdrawal/stopped during experiment	13	20		15			16				*15.6%*
Survey response - Baseline survey	59	90		68			88				74.4%
Survey 1	58	56		53			72				58.3%
Survey 2 (excl. stopped)	53	57		46			64				63.6%
Informal experiments											
Oss (Oss) N=344	**119**	**110**		**115**						**1496**	
Withdrawal before experiment	17	13		14							*12.8%*
Survey response - Baseline survey	75	89		89							84.1%
Survey 1 (excl. stopped)	41	80		70							63.5%
Survey 2 (excl. stopped)	49	64		67							59.8%
Apeldoorn/Epe (Apd-Epe) N=580	**190**	**188**		**202**						**4629**	
Withdrawal/stoped during experiment	20	38		30							*15.2%*
Survey response - Baseline survey	174	153		188							88.8%
Survey 1	91	78		112							48.4%
Survey 2 (excl. stopped)	77	85		107							54.7%

Note: CG=control group; A=exemption/self-management; B=intensive support; C= earnings release; D=e-support with APP (Dev); E=customized support (Gro), Ref1/Ref2 = reference group 1/2.

Source: LOEP, 2020; De Boer et al. (2020).

5.2. Outcome measures

Article 83 of the Participation Act set the requirements for the various local experiments as well as for the research that is considered an essential part of the test. With respect to outcome measures, a distinction is made between primary and secondary outcome measures. Primary outcome measures concern outflow to work because of which recipients become less or non-dependent on welfare while earning at least part of their income with paid work. The national Government was particularly interested in these employment outcomes because the primary aim of the Participation Act is to improve or speed-up the transition from welfare into paid work. The municipalities saw in recent years a very small minority of recipients being able to transit to paid work each year (5 to 10%) even though their 'carrot and stick' or 'workfare' approach in implementation had become stricter with a view to monitoring peoples' compliance behavior. For that reason, they argued that for them the test would be also a success when people on welfare because of a more relaxed treatment would feel happier and healthier and therefore become more self-confident and self-reliant to be able to move into paid work in the future. These indirect effects of the alternative treatments on health, wellbeing, trust and self-reliance were by the Ministry of Social Affairs and Employment labelled as secondary outcome measures. For the municipalities it also implied that when in-between steps are needed to bridge the gap between work and non-work, there should be room in the implementation practices to use a variety of appropriate tools such as education and training, volunteer work, sheltered employment or social and psychological support. In consultation with the municipalities, researchers agreed on some commonly defined secondary outcome measures. These are subjective wellbeing (SWB), health (self-rated health and mental health: SMHEALTH), self-efficacy in finding work (SE), job search efforts (JSI), perceived capabilities (CAP), volunteering and social networking (SPART), trust (social and institutional trust: TRUST) and financial situation (FINSIT). Researchers in each municipality could opt for applying additional outcome measures and some did such as Tilburg, Wageningen, Oss and Apeldoorn on income and deprivation poverty (INCPOV; DEPPOV). See Annex 1 for a summary of the exact operationalizations. Below we give some further detail on each of these outcome measures.

A. Employment outcome measures

For the employment outcome measures we assessed the outflow to fulltime, parttime or temporary paid work during the time of the experiment. These measures have been assessed in three ways.

- First, local administrative data were used providing information on the reason of termination of the registration as a social assistance beneficiary. When the reason was outflow to fulltime work (work with at least 27 hours a week at the minimum wage), the wage income of the recipient equals at least the benefit and the registration of the welfare recipient was ended. The registration is in most of the cases based on a careful investigation of the employment status and wage income of the former recipient[19]. We assume the registration is reliable even though the underlying evidence is based on

[19] The local register information on employment exit from the welfare benefit includes exit due to subsidized sheltered employment but also to self-employment if the earned income exceeds the 70% of the minimum wage threshold.

reported information from the beneficiary. The local register evidence on welfare and employment status has been used in some of the local reports of Tilburg, Nijmegen, Wageningen, Oss and Apeldoorn-Epe, (see list of reports in Annex 1).

- Second, survey data were used in which participants were asked in wave two and three to report their work status and changes since therein since the last survey with a view to fulltime, parttime, temporary or volunteer work. This survey evidence on employment status changes has been reported in the local reports. Due to substantial survey non-response especially in later waves, the evidence cannot provide a full and reliable picture of the employment outcomes.
- Third, register data of Statistics Netherlands (CBS) were used by linking the experiment data to the microdata files containing information on benefit recipiency, the person's number of working hours in the jobs they occupy after the experiment, type of contract (temporary versus open-ended) and the monthly earnings in these jobs. The information stems from the social security administrations and the job files of the employers who delivers it to the administration of the unemployment offices (UWV). The register information is also used in the Netherlands Bureau for Economic Policy (CPB) report on the six official experiments (De Boer et al., 2020). Also, some of the local reports were based on these register data (Groningen, Deventer, Nijmegen, Utrecht) and the joint research report by the researchers (Sanders et al. 2020). The information on jobs and earnings in the first data analyses covered January 2015 to October 2019. Later on, we updated the information to December 2019 and recently to June 2020. This means that the observation window was extended to a period of 24 months before the experiment to 3 to 9 months after. For the work outcome indicators, we used the register data on whether and when participants move into part-time, fulltime or temporary paid work after participating in the experiment. Contrary to the local data no information was available at that time on exit into self-employment creating some differences in outcomes between the two sources. Two sets of indices have been defined based on either the level of monthly earnings or the number of working hours per week. The earnings measures view the proportion of people in the various treatments earning 50, 70 or 100% of the minimum wage whereas the hours measures look at the proportion of participants working 8 hours, 12, 27 or 32 hours or more a week. The Dutch social assistance benefit is linked to the minimum wage and amounts for an individual to 70% of the minimum wage. When people earn at least 70% of the minimum wage or when they work at least 27 hours a week, they receive a wage income at least equal to the social assistance benefit because of which the recipient moves out of the benefit scheme. When people earn less than 70% or work less than 27 hours a week, they seemed to have moved to parttime work because of which they stay in welfare and get a partial benefit.

B. *Secondary outcome measures.*

For the secondary outcome measures we use the information collected in the three surveys.

Wellbeing and health

Subjective wellbeing is used as a measure for people's subjective feelings about the quality of life and has been assessed using a set of two in academic work commonly used scales on life-satisfaction (Diener et al. 1985; Veenhoven 1984; Lyubomirski & Lepper 1999) and a meaningful live (Ryan & Deci 2001; Seligman et al. 2000). For life satisfaction and meaningful live, we use a one-question scale ranging from zero to ten. The life satisfaction question was also used in the World Value Survey[20]. The 'meaningful live' question is derived from the European Social Survey.

Subjective health is based on a single item question (a 1-5 Likert scale ranging from very good to very bad health) on self-reported general health derived from the SF-36 health survey (Ware & Sherbourne, 1992). Our mental health index is a 5-item subscale of the same MOS 36-item short-form health survey that we used for subjective health. The MHI-5 asks respondents to rate their mental health during the past month on a 6-point Likert scales ranging from 0, never to 5, continuously, such as 'This past month I felt very anxious' and 'I felt depressed and gloomy'. The two scales are next aggregated into a subjective and mental health scale (SMH).

Self-efficacy and job search intensity

For the impact on job search two indices were defined on people's self-reported job search abilities and their job search efforts. First, a self-efficacy scale was constructed consisting of four items extracted from the 10-point scale developed by Ellis and Taylor (1983) indicating people's confidence or self-esteem in finding work oneself now or in the near future (Saks et al. 2015). We selected four items about people's self-reported self-efficacy in finding work using a five-point Likert scale ranging from 1, fully disagree, to 5, fully agree (SE)[21]. Second, we constructed a so-called job search efforts index consisting of three items: people's job search intensity measured by the weekly number of hours spent on job search in the last four weeks as well as the outcomes of job search with a view to the number of applications and invitations for an interview in the last four weeks (JSI)[22].

Sen's Capabilities

The impact of the experimental treatments on people's perceived level of freedom of choice is measured by the participants' perceived level of 'capabilities' as measured with a self-reported capability index as constructed by the group based on Sen's capability approach (Sen, 1999; 2004). The index is composed using a seven-items long question on people's capabilities partly derived from Klink et al. (2016) who used similar items in their questionnaire for people at work. These were: a. to do things one is qualified for; b. to learn and to do new things in life; c. to co-decide on important decisions in work; d. to have good

[20] For combining the two indices into one scale we rescaled both into a 0-10 scale.

[21] These self-efficacy items were: 1. Confident in finding work oneself if one really makes an effort; 2. Confident in finding work in the near future; 3. Confident in making a good impression in applications and 4. Confident in finding a job that fits one's skills and experience.

[22] The three job search intensity items were normalized and aggregated into an index ranging from 0 to 10.

contacts with other people in work; e. to set one's own targets; f. to have a decent income and g. to contribute to the life of others. In some cities (Deventer, Nijmegen) only the first 6 items were asked for. Two questions were asked: 1. Whether one considers the capability item important for their own life and 2. Whether one thinks one can achieve this in the current situation. A five-point Likert scale has been applied for the second question ranging from 0, not at all to 5, always. After rescaling and weighting the scores with the importance attached to each item (ranging from 1 'not important' to 5 'very important') we created an index of perceived capabilities (CAP) ranging from 0 to 10. The capabilities questions were included in the surveys of 6 of the 8 municipalities (not in Groningen and Utrecht).

Social participation

One of the goals of the municipalities for the experiments was to improve people's level of social participation by offering in-between steps between work and non-work when a paid job was impossible on short notice. For that reason, three indices were applied, the first one on perceived social integration in society, the second one on the weekly number of hours spent on volunteering and informal care and the third one on the frequency of people's contacts in their social network (SNETW). The social integration measure was based on a single-question, derived from the Panel Study Labour Market and Social Security (PASS), to what extent people feel welcome or integrated in society on a scale ranging from 1 to 10. The scale was translated and adapted to range from 0 to 10. The volunteering question asked for the number of weekly hours spent on volunteer work or informal care whereas the social network measure was based on the number of monthly contacts people have with their friends and acquaintances, neighbors and family. The frequency of monthly contacts scores on this last measure ranges from seldom or never, less than once a month, to once a week or more. By assigning a value ranging from 0 to 4 times a month and normalizing these scores, we created a social network index on a scale from 0 to 10.

Trust

Trust refers to two different components of trust, social trust or the trust people put in others and institutional trust, the trust people put in institutions such as the government, the parliament or politicians (Glaeser et al. 2000; Uslaner, 2002). These two components of trust were also asked in the European Social Survey. The ESS question on whether people can be trusted or not is translated and included in the three survey questionnaires. The ESS question on institutional trust, consisting of four items, is rephrased to include the institutions the beneficiaries of social assistance have to deal with such as the municipality, the social assistance office and the caseworker[23]. The social trust variable is measured on a scale ranging from 0 to 10. The institutional trust variable consisting of the four institutional trust items (government,

[23] One of the three ESS questions on social trust is used and reads as follows: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.". The question on institutional trust reads as follows: "How much trust do you put in the following institutions: the national government, the municipality, the social assistance office and the caseworker?". The answer categories have a 4-point Likert scale ranging from "not at all" to "full trust".

municipality, social assistance office, case worker) uses a 4-point Likert scale ranging from 1, not at all, to 4, full and rescaled from 0 to 10. The two indices were in the final step combined into an aggregated generalized trust index (TRUST).

Financial stress and poverty

Finally, we employed some measures for perceived financial strain and poverty. The financial strain index was based on a five-item question on people's current financial situation derived from Elo et al. (2003), in which people were asked to rate their situation on a five-points scale from 1, "I need to make debts to make ends meet" to 5. "I can save a fair amount of my income"[24]. For income poverty a cut-off point was defined, so that people scoring 1 (need to make debts) or 2 (need to dissave) were considered income poor. Some cities used in addition a measure for deprivation poverty. For material deprivation a so-called material deprivation scale has been extracted from a more extended deprivation scale as applied and used in the EU-SILC questionnaires (European Statistics on Income and Living Conditions). Instead of the full list we used a restricted list of 14 items which in our view were most relevant in the Dutch context (see Annex 1 for the list of items). For deprivation poverty, out of these 14 items we selected 5 items which were considered basic and essential for making a minimum standard of living: once a day fruit or vegetables; once a day, a meal with meat, poultry or fish; replace worn-out furniture; replace worn clothes and repair damaged equipment in the house. When people lack three or more items out of this list of five, provided each item was also considered (very) important for their lives, we count them as deprivation poor (DEPPOV).

Not all measures will be reported for the purpose of this report. Here, we will focus on a limited set of measures having a direct or indirect effect on job search and employment opportunities: job search intensity, self-efficacy in finding work, subjective health and wellbeing, social trust and trust in the caseworker, and financial stress.

5.3 Methods

In this section we discuss in 5.2.1 first the empirical model that has been used for testing the linear effects of the various treatments on the various employment and non-employment measures using the administrative and survey panel data. In 5.2.2 we discuss some methodological issues which might complicate the measurement of the causal effects and we briefly explain the research strategy followed.

5.3.1. Empirical models

[24] The scale was actually a seven-item long scale because people could refuse to answer (6) or answer that they cannot tell (7).

Partaking in the treatments in these local experiments is voluntary. People can withdraw after assignment to the treatment. This *failure-to-treat* means that some of the people assigned to the treatment has not received the treatment. The consequence is that identification of the ATE, that is the average treatment effect for the treated and non-treated, of a particular intervention becomes impossible. To circumvent this problem, two alternative estimation procedures have been proposed which are also used here: ITT (intention-to-treat) and LATE (local average treatment effect). For the survey data we can only estimate the ITT-effects since no survey-data were collected for the non-treated.

ITT-analysis

The purpose of the local experiments is to estimate the causal effect of each treatment compared to the standard treatment in the control group. To be able to arrive at reliable estimates of the causal treatment effects some methodological requirements need to be met (see section 5.2.2 below) such as the absence of (self)-selection into the treatments, accidental differences that arise in the randomization procedure, the withdrawal of people causing a low compliance rate because of which the statistical effects 'water down' or are hard to find and behavioral effects at the side of the participants just because of being part of an experiment (De Boer et al. 2020; Angrist & Pischke 2009). The ITT-analysis or *intention-to-treat* analysis is based on all people assigned to the treatment even when they withdraw or stop at a later stage. The real participation of the people assigned to a particular treatment is therefore not taken into account. All people are assumed to have participated and be compliers with the assignment to the treatment. The effects of each treatment are compared with those for the randomized control group based on the assignment to a particular treatment. The LATE or *local-average-treatment effect*-analysis estimates the effects taking account of the share of compliers to the treatment. The LATE analysis calculates the effect for the compliers to the treatment. The ITT and LATE effects are related to each other. If all people assigned to a treatment comply to the treatment, the share of compliers is one and the ITT-effects are similar to the LATE-effects. If there are non-compliers the LATE effects are equal to the ITT-effects divided by the share of compliers. The LATE-effects are in that case larger than the ITT-effects which can be regarded as a conservative or lower bound for the true causal effect of the treatment. The ITT effects are the average of the LATE effects for the compliers and the zero effect for the non-compliers. When the compliance rates are high (as is the case in most of the local experiments with a post-randomization procedure) the ITT-effects give a good estimate of the true effects. The ITT-effects can be estimated with a linear probability model. The LATE effects can under certain assumptions be estimated with a 2-Stage Least Squares (2SLS) regression model. The ITT-model for the administrative data is given by equation (1).

$$Y_{it} = \alpha_{0t} + \sum_{t=1}^{T} \beta_t T_i + X_i' \gamma_t + q_i + \varepsilon_{it} \qquad (1)$$

where Y_{it} is the outcome of interest for individual *i*, measured at *t* time periods after the start of the intervention. For the *administrative data* that is used for the employment outcome measures, *t* is measured in months. $\sum_{t=1}^{T}\beta_t T_i$ represents a series of dummies for the set of *T* treatments, where the omitted category (*T=0*) is the control group treatment and $X_i'\gamma_t$, a set of individual and household variables: gender, age, migration background (western, non-western migrant) living situation (single, single parent, couple no kids, couple with kids, other), highest education level (low, intermediate, high), spell duration and cumulative earnings 3 to 24 months before the start (earnings history), to control for selection effects. These are fixed over time and defined at three months before the start of the experiment. The variable q_i represents a fixed variable as well indicating the quarter in which the individual started the treatment. ε_{it} is the standard error term used in this linear-effects model.[25] The model does not include interaction effects between the controls (X_i') and the treatments (T_i) even though research into the heterogeneity of the treatment effects by gender, age, skill-level and duration might add to the interpretation of the results. This is left for further scrutiny. Eventually, robustness checks were performed by estimating various models with different specifications of the model including 6,12 and 24 months of earnings history, age and age squared, and log duration and log duration squared. The specification explained here is called the 'main' model.

For the survey data which are used for measuring the health and wellbeing effects of each treatment, the model used is more or less similar except that we add the baseline level of the outcome variable in the equation (Y_0) and that the time *t* refers to the three survey waves. For the purpose of this report, we only consider the outcomes for the second wave of the panel survey at *t* =2.

$$Y_{it} = \alpha_{0t} + \textstyle\sum_{t=1}^{T}\beta_t T_i + X_i'\gamma_t + Y_{i0} + q_i + \varepsilon_{it} \qquad (2)$$

The parameter of interest is β_t, that describes the effect of a treatment *T* in the *t* survey wave after the start of the intervention controlling for the initial outcome in the baseline survey. Again, no interaction effects were included.

LATE-analysis

To obtain estimates for the average treatment effect of the treated (LATE) we can either divide the ITT by the estimated proportion of compliers in the assigned treatment condition or instrumenting the random assignment to a treatment in a two-stage least squares (2SLS) regression. The last option is used here.

[25] Several models with a different set of covariates were tested to examine the sensitivity of the results for the main model specification. We also estimated a logit specification for the main model instead of OLS. In two separate models we included age squared and replaced the duration variable by the log of it. The results can be obtained at request. The results differ only slightly and do not change the general picture.

The specification for the first stage is identical to the ITT model except that the dependent variable T_{it} is a binary variable indicating the actual treatment status of the individual.

$$\hat{T}_{it} = \alpha_{0t} + \sum_{t=1}^{T} \beta_t T_i + X_i' \gamma_t + Y_{i0} + q_i + \varepsilon_{it} \qquad (3)$$

Then in the second stage the treatment effects are calculated including the estimated treatment effects for the compliers in the first stage. For the remainder the variables in the model are identical to the ITT-model in equation (1).

$$Y_{it} = \alpha_{0t} + \sum_{t=1}^{T} \beta_t \hat{T}_{it} + X_i' \gamma_t + Y_{i0} + q_i + \varepsilon_{it} \qquad (4)$$

The LATE effect for month *t* is represented by the β_t's. Recall, that the outcome measure Y_{it} can only be assessed with the administrative data because no survey data are available for the non-compliers. It means that for the survey data we can only perform the ITT-analyses.

5.3.2 Some methodological issues and research strategy

To be able to conduct a pure and unbiased estimation of the causal treatment effects of each group compared to the comparison group, some *methodological issues* are needed to take into account and to address in the analyses.

- *Pre -and post-randomization.* Through randomization the experimental groups are allegedly comparable in observed (education, duration) and unobserved characteristics (motivation, effort). When randomized before registration such as in Deventer and Groningen (pre-randomization), the results for the experiment groups are then valid for the target population of beneficiaries (external validity). When randomized after registration such as in the other cities (post-randomization), the results are then only valid for the experimental groups (internal validity). In the case of Groningen, the so-called Zelen's design (ZEL) with pre-randomisation is used, that is, the entire population is randomly assigned to a treatment, control, and reference group, and subsequently invited to participate in the trial (so-called post-randomization consent). In a standard RCT design, all the participants that refuse to participate in a trial are excluded before randomisation, whereas in the ZEL they are included. This means that ZEL designs have substantially lower compliance rates than the RCT design. This low compliance rate is problematic in the assessment of treatment effects. Smaller effects will likely become unidentifiable, and larger amounts of noise may produce counterintuitive results.
- *Random allocation.* Notwithstanding the randomization procedure for the experimental groups, in reality, there may be accidental differences in the composition of the groups that may affect the outcomes. For Nijmegen the local report mentions the possibility that the randomized control group behaves different because of selection in observed (earnings before start; health conditions) and unobserved characteristics (motivation; effort, expectations)

- *Experiment effects*. The behavior of the participants and the caseworkers in the control group should remain the same during the experiment, otherwise a true comparison with the treatment groups is not possible. However, experiment effects might change the behavior of the participants and/or the caseworker in the experimental groups (so-called Hawthorne effects) or in the control group (so-called John Henry effects; Duflo et al. 2008; Athey & Imbens 2017; De Boer et al. 2020). In the CPB report possible John Henry effects were suggested for Groningen, but not for Deventer. For Tilburg, no formal test was possible, but a graphical inspection of the estimated ITT-employment effects over time indicated that the control group showed much higher employment outcomes already shortly after the start which continued over time. For that reason, the CPB-report concluded that the found effects cannot be attributed with security to a causal treatment effect. In the results section we provide some more evidence.
- *Equal starting position*. Before or at the start (baseline), the randomized experimental groups should have an equal starting position on the chosen outcome measures. If this is not the case, the effect found is not the result of the intervention but of differences in starting position (selection). The researchers have therefore employed *placebo regressions* using some of the employment outcome measures to check whether the equal starting position supposition is met or not by formally testing and/or graphically inspecting the estimated employment effects of each treatment group compared to the control group for each month already from 25 months before the start of the experiment. If the condition is met, the experimental groups should be comparable before and at the start with a view to the chosen employment outcomes. In some cities this condition was not fully met, such as in Utrecht, Nijmegen and Wageningen.
- *Sample selection*. Before and at the start or baseline, the experimental groups should be more or less equally composed compared to the (randomized) reference group of non-participants. If this is the case, there is no selection among participants (because of withdrawal) and non-participants (because of submission). Before, we argued that if there is selection, the comparison of the control group with a randomized reference group of non-participants at the start and shortly after can show the existence of selection in the control group. A similar comparison is possible between the treatment groups and the randomized reference group showing the existence of selection, if any, in the treatment groups.

In the various local reports presenting the research outcomes but also in the overall report and the report of the CPB, these methodological issues were discussed in more or less detail. For three cities (Tilburg, Groningen, Nijmegen), the CPB-report concluded that the measured effects on fulltime paid work cannot, causally, be fully attributed to the treatments due to the finding that the control group behaves rather differently concerning employment already shortly after the start in Groningen. In two other cities, Utrecht and Wageningen, the control group behaved slightly different before the start with respect to parttime work. These findings will be further discussed in the results section but show that selection and experiment effects in the control groups (John Henry) need to be taken into account and further investigated.

Research strategy

Taking these issues into account, what then should be our research strategy? We have chosen for four different steps: *1.* Performing representativity and balance tests to investigate whether there are significant differences found in background characteristics between the various groups; *2.* Performing ITT (intention-to-treat) and LATE (local average treatment effect) regression models for estimating the linear effects of the various treatments *3.* Performing placebo regressions to investigate selection effects by comparing the various groups on pre-treatment outcomes already 24, 12 or 6 months before the start and 4. Investigating and testing for experiment effects (notably John Henry effects for the control group). Testing for Hawthorne effects (behavioral change in the treatment groups) is difficult because there is no true comparison group but John Henry effects can be investigated by comparing the employment outcomes over time for the control group with those for the treatment groups. If the employment outcomes evolve already shortly after the start in a very different direction in the control group compared to the treatment groups it might indicate the existence of behavioral effects. A formal test of behavioral effects in the control group is possible by comparing the outcome for the control group with a randomized comparison group of non-participants who also received the regular treatment (John Henry effects). The latter test could be done for Groningen and Deventer only. Before we present the results on these four steps, we first report on the empirical model we have used for calculating the treatment effects of each treatment in the various cities.

6. Results on employment outcome measures

In this section, the results of the four steps in the research strategy will be subsequently discussed. In 6.1 we discuss the representativity and balance tests. Then in 6.2 we present some descriptive results by mapping the outcomes for fulltime and parttime work for each month from 25 months before the start to 16 to 24 months after the start[26]. In 6.3 and 6.4 the ITT-analysis and LATE-results are presented and in 6.5 the results of the placebo regressions and experiment effects analyses.

6.1. Representativity and balance tests

In the first step, to check for selectivity in the submissions while using the register data, we performed representativity and balance tests, first, by comparing the characteristics of the target population versus the

[26] Here we present the evidence on parttime and fulltime employment only. The results for the 70% minimum wage measure are similar to the ones for the fulltime measure. For Deventer, Groningen, Nijmegen and Utrecht we present the results for 25 months before the start to 16-24 months after the start as provided in the CPB report (De Boer et al. 2020).

sampled groups (representativity), and second, by comparing the various treatment groups with the control group (balance test). The balance tests show to what extent our randomization procedure has been successful regarding the composition of the groups. In the representativity tests we compare the target population and sampled groups by regressing each background characteristic on the two groups (OLS)[27]. The results on the representativity tests are shown in Table 4 below (for the precise coefficients we refer to the CPB-report and the local reports). The results are more or less as we expected beforehand, showing that more older persons tend to be willing to participate, more singles and less couples, less low educated, less migrants, notably of non-Western origin, and less people with longer durations in welfare at the start. This picture is found in all cities, though in some cities we find some noticeable deviations from this pattern.

Table 4: Representativity tests: significant differences between target and sample group by city (number of people in target population and sampled groups between brackets; p<0.10)

City	**Deventer (N=1789/387)**	**Groningen (N=8338/890)**	**Nijmegen (N=6030/366)**	**Tilburg (N=5386/780)**
Age	Older	Older	**Younger (less >55 yrs)**	Older (less <34 yrs)
Gender	More males			**More females**
Living situation	**More couples, no kids**	More singles, less couples	More singles	More singles, less couples
Country of birth		Fewer migrants (non-Western)	Fewer migrants	Fewer migrants (non-Western)
Education	Higher educated	Higher educated	Higher educated	Higher educated
Spell duration	Shorter durations (less >3 yrs)	Shorter durations (more <1 yr; less >3 yr)	Shorter durations (less >3yrs)	**Longer** durations (more >1 yr)
Wages 3-9 mnths bef. start		Higher wages	Higher wages	**Lower** wages
City	**Utrecht (N=8338/752)**	**Wageningen (N=932/410)**	**Apeldoorn (N=3837/486)**	**Oss (N=1849/345)**
Age	Older	**Younger (less >55; more 35-44)**	Older (less <34 yr)	Younger (more <34 yr; less >55 yr)
Gender		More males		
Living situation	More singles	**More couples with kids**	More single parents, less couples	Less singles; more single parents
Country of birth	Fewer migrants (non-Western)	**More migrants (non-Western)**	Less migrants (non-Western)	Fewer migrants (non-Western)
Education	Higher educated	Higher educated (less low)	Higher educated (less low)	Higher education (less low)
Spell duration	Shorter durations	Shorter durations (more 1-3 yr; less >3 yr)	Shorter durations (more <1 yr; less >3 yr)	**Shorter durations (more <1 yr; less >3 yr)**
Wages 3-9 mnths before start			Higher wages	Higher wages

Source: LOEP, CBS-microdata analyses, February-June 2020 (noticeable differences in bold).

In Deventer and Wageningen, more couples without kids have participated. In Tilburg more females and people with longer durations in welfare and lower earnings at the start of the experiment were registered. In Wageningen also more younger people and more migrants with a non-Western origin partake. In Oss we found that people with rather short welfare spell durations (<1 year) were much more overrepresented than in other cities possibly associated with the sample design to have 50% of the sample consisting of

[27] The detailed results of the balance tests for the various cities can be obtained at request from the authors.

newcomers and 50% from people longer than one year in welfare. A similar sample design was adopted in Apeldoorn-Epe showing also there that shorter durations were overrepresented. In the local report of Nijmegen, a separate study on sample selectivity was discussed, showing that people who before the start of the experiment were already engaged in part-time or temporary work were strongly overrepresented. Underrepresented were people who already were exempted from the obligation to work, or were reluctant to register due to being very stressed (Betkó et. al., 2019). This suggests that the participants in Nijmegen are a relative well-off group compared to the target population as also shown in Table 4. The message from the representativity checks is that selectivity due to submission and withdrawal (after registration) cannot be ignored when in the next step, we want to estimate the true causal treatment effects. For that reason, we correct for selectivity by including the variables of interest, as far as available, in our regression models[28].

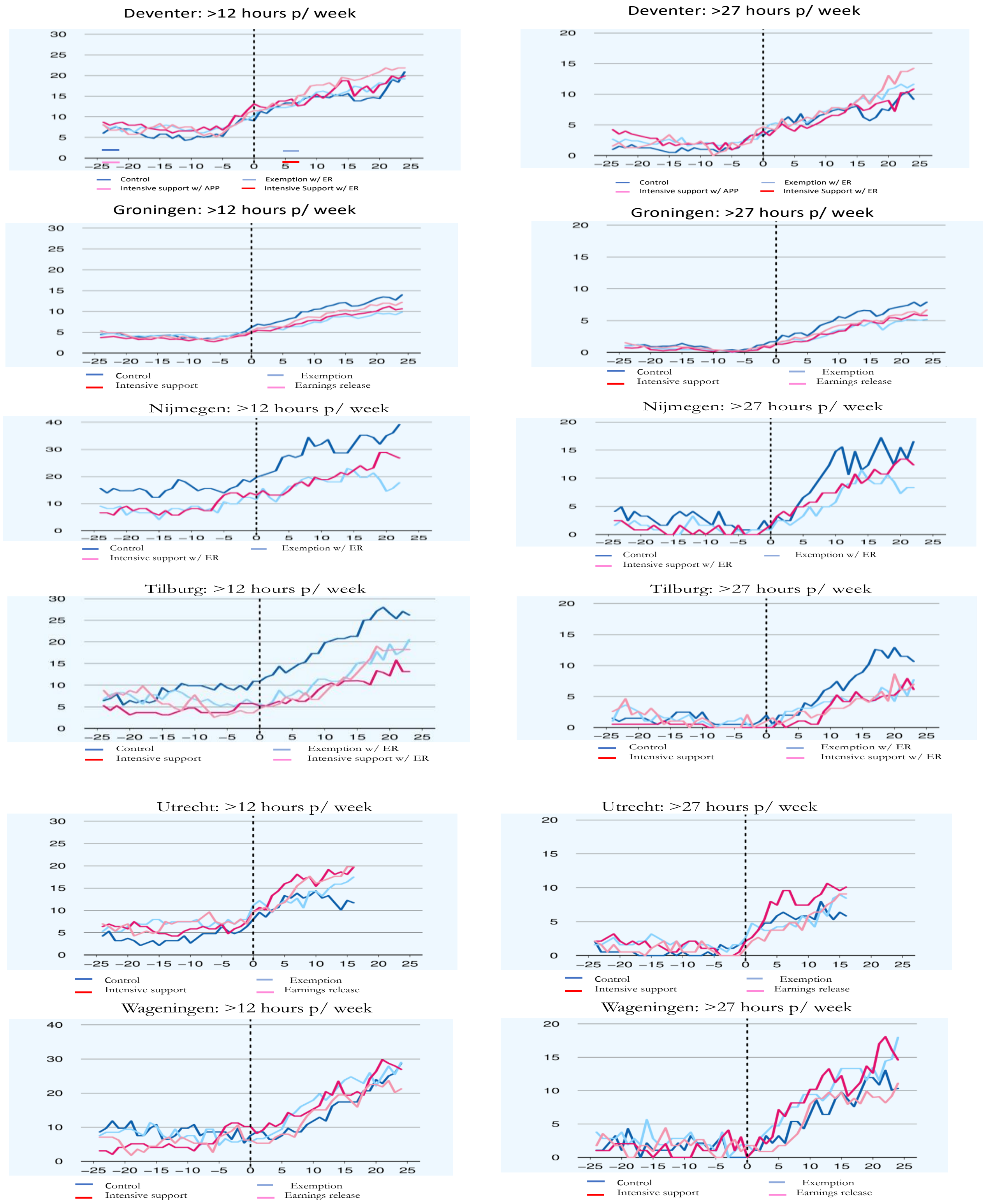
Balance tests

The results of the balance tests show that the randomization of the participants either before or after invitation has worked out rather well. We found very little significant differences in background characteristics between the treatments and control group in the various cities but slightly larger differences with the randomized reference group 1 in Groningen notably on migration background and education. In Utrecht lower educated were underrepresented in the exemption group and the earnings release group. In Nijmegen more intermediate educates were registered in the intensive support with earnings release group and people in the exemption group had lower earnings 6 months before the start. In Tilburg the two intensive support groups registered more older people, with longer durations and lower earnings at the start than in the control group. Wageningen had younger people in the exemption group. In Oss there were more singles registered and in Apeldoorn more middle aged (45-54) people in the intensive support groups.

6.2 Descriptive results

In the CPB-report (De Boer et al. 2020), descriptive evidence on the employment outcomes for the six official experiments is given in the form of graphs showing for each city the evolution of exit into parttime including fulltime (12 hours or more) and fulltime employment (27 hours or more) over time (in percent points) from 25 months before the start of the experiment to 16 to 24 months thereafter. The results are presented in Fig. 1. The graphs are based on the raw data and subsequent regression analyses (see section 6.3) must show to what extent the differences will be statistically significant after controlling for background characteristics.

[28] Another way to correct for selection is to use matching techniques (e.g., Propensity Score Matching) to compare the outcomes for people in the various treatment groups with those of similar non-treated others, based on some observed characteristics (see the local reports of Tilburg, Wageningen, Oss and Apeldoorn-Epe). However, it cannot be ruled out that there remains selectivity in unobserved characteristics.


Deventer: >12 hours p/ week

Deventer: >27 hours p/ week

Groningen: >12 hours p/ week

Groningen: >27 hours p/ week

Nijmegen: >12 hours p/ week

Nijmegen: >27 hours p/ week

Tilburg: >12 hours p/ week

Tilburg: >27 hours p/ week

Utrecht: >12 hours p/ week

Utrecht: >27 hours p/ week

Wageningen: >12 hours p/ week

Wageningen: >27 hours p/ week

Source: De Boer et al. (2020)

Figure 1: Evolution of exit into parttime and fulltime employment over time for the various treatment groups in each of the six official local experiments (raw data)

They show what happens with the employment exit rates directly before and after the start of the experiment for the various treatment groups compared to the control group (the blue line). In three cities, Groningen, Nijmegen and Tilburg, the control group behaves differently already at the start and shortly after the start for both, exit into parttime (including fulltime) and fulltime employment. In Groningen, the control group not only behaves differently compared to the treatment groups but also rather differently compared to the randomized reference group of non-participants (not shown in Fig. 1).

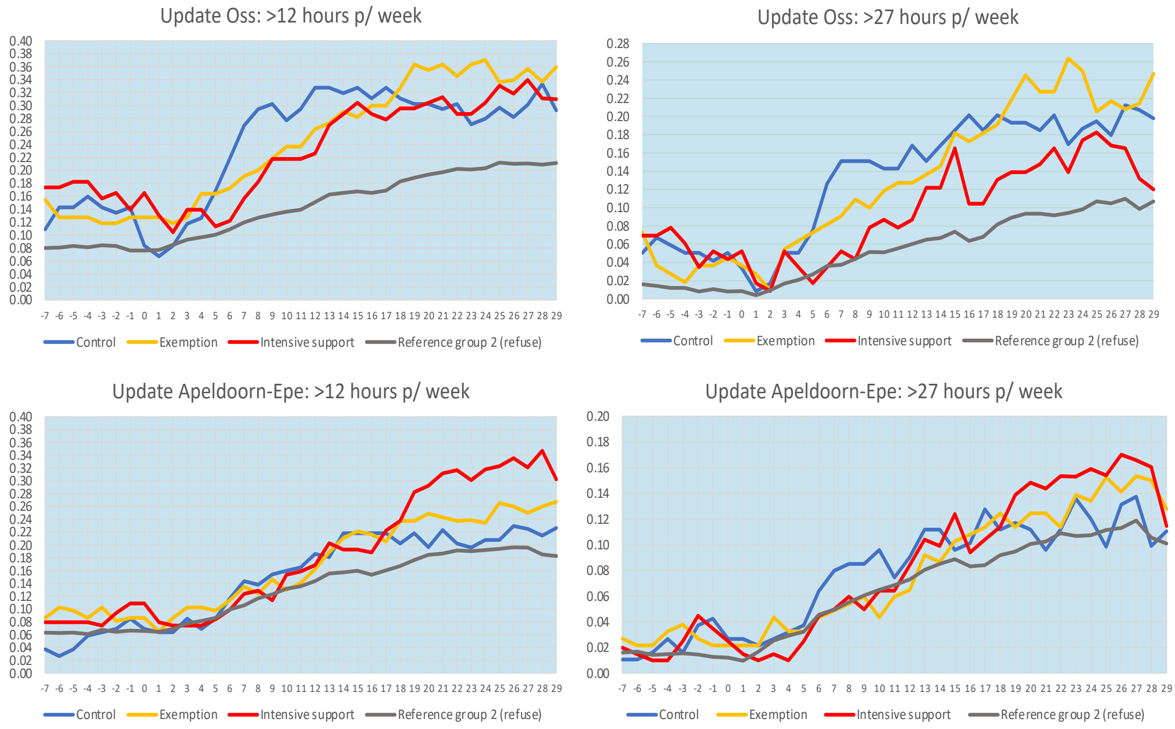
For Groningen, it has been argued that the negative differences with the control group are due to a large number of non-participants in the control group who exit more to employment than participants in this group whereas in all treatment groups and in the random reference group the non-participants exit less to employment than participants. The researchers argue that only differences in motivation (motivated to perform better with a view to exit to work) can explain these implausible results (Edzes et al. 2020). This suggests that experiment effects in the control group (John Henry effects) might have influenced the results in Groningen[29].

In the local reports of Nijmegen (Betkó et al. 2020) and of Tilburg (Muffels et al. 2020) the researchers also point to implausible differences in employment exit in the control group which already start before (Nijmegen), at or shortly after the start of the experiment (Tilburg). This might be attributed to the emergence of 'accidental differences' in observed or unobserved characteristics from the randomization procedure (Nijmegen) or to selection effects (before the start) or behavioral (John Henry) effects in the control group (after the start such as in Tilburg)[30]. The largest negative differences are found in Tilburg between the intensive support group and the control group amounting to 11 percent points (measured at 16 months of observation). Also, in Nijmegen we found substantial negative differences notably for parttime work for the two treatment groups already before and shortly after the start. The results further show that the strongest positive employment differences are found in Utrecht, both for parttime and fulltime exit. The earnings release and the intensive support group seem to perform best with a view to exit into employment which differences with the control group becomes stronger the longer the experimental treatments last. Also, in Deventer and Wageningen the results show positive differences for notably exit into fulltime employment for the intensive support (in Deventer with an App) group and remarkably also the exemption group. The differences tend again to become stronger the longer the experiment lasts. For the two informal experiments of Oss and Apeldoorn-Epe, in Figure 2, the same results are presented, but now up to 29 months after the start instead of 24 months[31].

For Oss the employment results for the control group seem to be much higher for both, the 12 hours and the 27 hours measure already shortly after the start. However, after month 18 the gradient of the control group line becomes much flatter and the initially negative differences for the exemption and the intensive support group disappears and turn into positive differences. The strong positive employment results for the control group in Oss shortly after the start suggest the existence of experiment effects.

[29] For the regression analyses we examined therefore for Groningen also the effects in comparison with the randomized reference group.

[30] Experiment effects in the control group can arise with the participants but also with the caseworkers who because of participating in an experiment might behave differently.



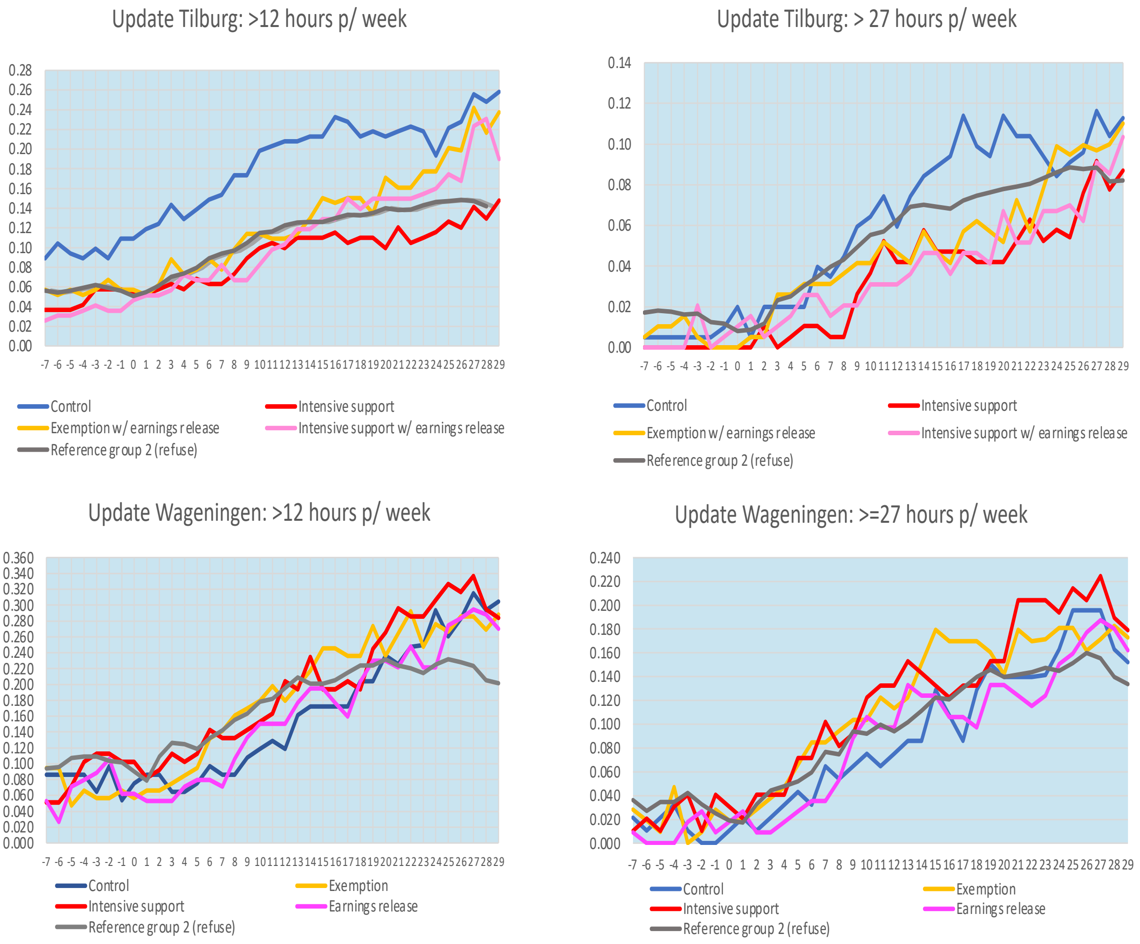
Source: Updated CBS-microdata from January2015-June 2020.

Figure 2: Evolution of exit into parttime and fulltime employment over time for the various treatment groups in each of the three informal local experiments (raw data)

Also, the large differences found between the control group and the reference group of non-participants already shortly after the start, even though this group was not randomized, suggest that experiment effects might have blurred the results[32]. In the local report for Oss, it is argued that the caseworker dedicated to the control group was part of the team of caseworkers participating in the experiment and implementing the treatments. The caseworkers were in tandem dedicated to a particular treatment and operated in close teamwork. The caseworkers for the treatments got a training before the start and discussed occasionally as a team the casuistry of the various participants. This might have led to "cross-over effects" in Oss already at the start which have affected the content of the regular treatment (see Muffels et al. 2020c). For Apeldoorn-Epe there is little evidence that experiment effects were affecting the regular treatment also because the regular treatment was executed by a separate independent office contracted by the municipality. The caseworkers of the municipality and the external office nevertheless operated as a team and also discussed the casuistry as a team. In the local report 'cross-over' effects were discussed but the researchers found no strong evidence for it (see also Muffels et al. 2020d). The findings for Apeldoorn-Epe show -

[32] In the local reports for Oss, Apeldoorn, Tilburg and Wageningen matching techniques were used to control for selection of the reference group. After correction, the differences between the control group and the reference group became a bit smaller in Oss but did not disappear also suggesting that experiment effects might have blurred the results in Oss.

notably for parttime work - positive differences for the exemption and the intensive support group up to month 27. After month 27 the employment results in Apeldoorn-Epe for the two treatment groups intensive support and exemption seem to decline suddenly but remarkably also for the control group and the reference group. This negative result might for the experimental groups be explained by the ending of the experiment at the end of 2019 and the support-change ('back to standard') and motivation-loss of the people involved. In Oss, the same decline can be observed for the treatment groups including the control group but not for the reference group of non-participants. This finding adds to the observation that in Oss John Henry effects might have affected the employment outcomes of the control group although more scrutiny is needed. Finally, we show the updated results for Tilburg and Wageningen (up to 29 months after the start) in Figure 3.



Source: LOEP, Updated results

Figure 3. Evolution of exit into parttime and fulltime employment over time: updated results for Tilburg and Wageningen.

Again, the findings confirm the earlier findings for the control group (affected by John Henry effects) and the treatment groups in Tilburg. The employment results for the reference group are worse over the entire experimenting period compared to the control group but much better compared to the treatment groups.

This is due to selectivity of the treatment groups. Further inspection of the data shows that the treatment groups have more women, more older people and longer spell durations compared to the reference group[33]. The figures also show the decline in employment exit in both cities after month 27, especially in Wageningen, possibly caused by the ending of the experiment. Notice also the decline in parttime and fulltime employment outcomes for the reference group of non-participants after month 27 in Wageningen which appears very unlikely. The employment outcomes of the earnings release group in Wageningen appear worse of all for the three treatments but only up to month 24, after which the employment record of this group improves strongly up to month 27, the month in which the experiment has ended.

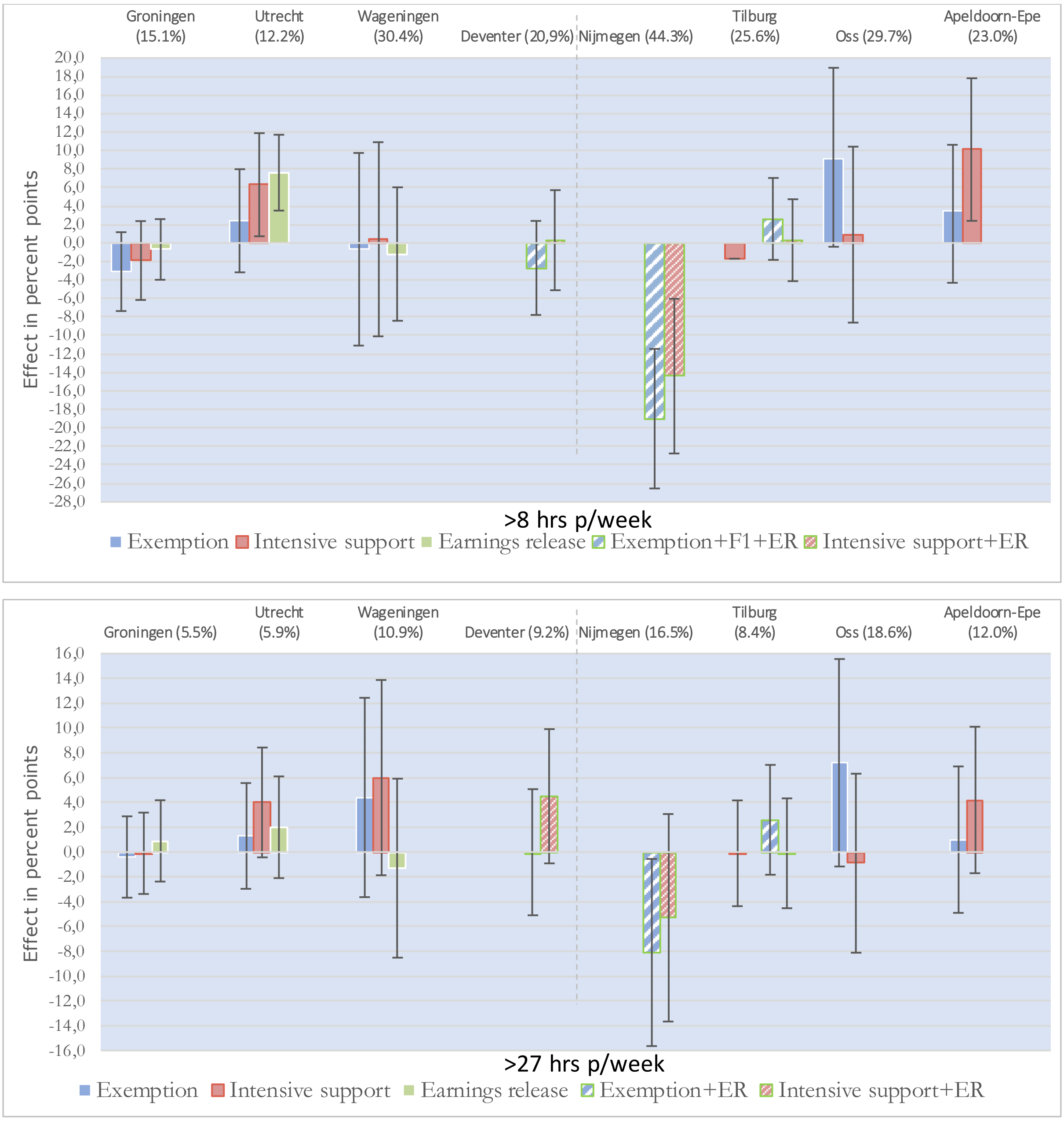
6.3 ITT-regression results

In this section the ITT regression results based on our main empirical model are presented for two employment measures, exit into fulltime employment or 27 hours a week or more and exit into parttime employment or 8 hours or more (the effects are presented in Table A3.1 and A3.2 in Annex 3). In the graphs, we first (left part) show the outcomes for the official experiments and then (right part) the outcomes for the informal experiments (Apeldoorn-Epe and Oss). For the official experiments, first the outcomes for the experiments are shown with only single treatments such as in Groningen, Utrecht and Wageningen and then the ones with combination treatments such as in Deventer, Nijmegen and Tilburg.

The first graph in Figure 4 showing the effects on exit into jobs of 8 hours or more reveals a mixed picture of positive and significant effects in Utrecht and Oss for exemption and in Utrecht and Apeldoorn-Epe also for intensive support. Significant negative effects are found in Nijmegen and Groningen for exemption and intensive support. In Tilburg negative significant effects were found for intensive support[34]. Also, in Deventer and Nijmegen where exemption is combined with a work bonus, the effects on this work measure are negative, but only significant in Nijmegen. However, in Utrecht, Deventer, Tilburg, Apeldoorn and Oss small but insignificant positive effect are found for the exemption group. Notice finally the large positive but (just) insignificant effect of exit into these jobs in Fig. 4 for the exemption group in Oss (9.1%).

[33] After controlling for selection effects of the reference group through applying PSM-matching techniques, the two treatments exemption and intensive support, both combined with earnings release, show slightly higher exit rates into fulltime employment compared to the reference group.

[34] The balance tests showed that this intensive support group in Tilburg has compared to the control group more older people, longer welfare spell durations and lower earnings already at the start. Part of this negative effect is hence, possibly not a treatment but a selection effect.



Note: Treatment effects compared to the control group in percentage points and 90% confidence interval. Results for the control group (means) are given in the heading. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. For Groningen, the comparison group is not the control group, but a randomly composed reference group (see Edzes et al. 2020). The figures for Tilburg and Wageningen are updated figures for 24 months after the start because of which they are slightly different compared to the ones presented by De Boer et al. (2020) and Sanders et al. (2020). For Utrecht the effects are assessed at 16 months and for Nijmegen at 22 months after the start. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. The effect estimates for the treatments in Groningen, Tilburg and Nijmegen cannot be fully attributed to the treatment. Two treatments which are not comparable to any other elsewhere are not shown here: the intensive support with app group in Deventer and the customized support group in Groningen.
Source: Sanders et al. (2020), LOEP calculations 2020.

Figure 4. Effect sizes and confidence intervals for exit into parttime (including fulltime) and fulltime employment

The second graph in Figure 4 on fulltime exit into employment shows that the effects for exit into larger jobs are in most cities for both exemption and extra support (except for Nijmegen and Oss) positive but of course smaller than for the 8-hours measure. The effects become however insignificant in all cities except for Nijmegen. The small numbers (power) of participants in each group notably in the smaller cities (Wageningen, Nijmegen and Oss) might have affected the significance level. The negative effect for the intensive support group in Tilburg becomes much smaller now compared to the 8 hours measure and its effect becomes stronger positive in Deventer. The effects for the exemption group are now, except for Groningen and Nijmegen, also positive though still insignificant. Earnings release exerts a positive effect now in Groningen and its effect is still positive in Utrecht. In Wageningen the effect for the earnings release or work bonus group is more negative showing that the negative effect is caused by exit into fulltime work and not by parttime (9-26 hours) work. Further analyses in Utrecht into more heterogeneous effects (Verlaat et al. 2020) show that the positive effect of intensive support occurs especially with people with low labor market chances or a large distance to the labor market. The barriers they face to find work might be associated with insufficient skills or a bad physical and mental health which according to more than half of the welfare population in these cities makes them unfit to work (cf. Muffels, 2020e).

However, the 8-hours measure includes exit into fulltime jobs of 27 hours or more. To obtain the net effects for parttime work only (9-26 hours a week) we need to subtract the effects for the 8-hours measure with the effects for the 27 hours measure (cf. Fig. 5).

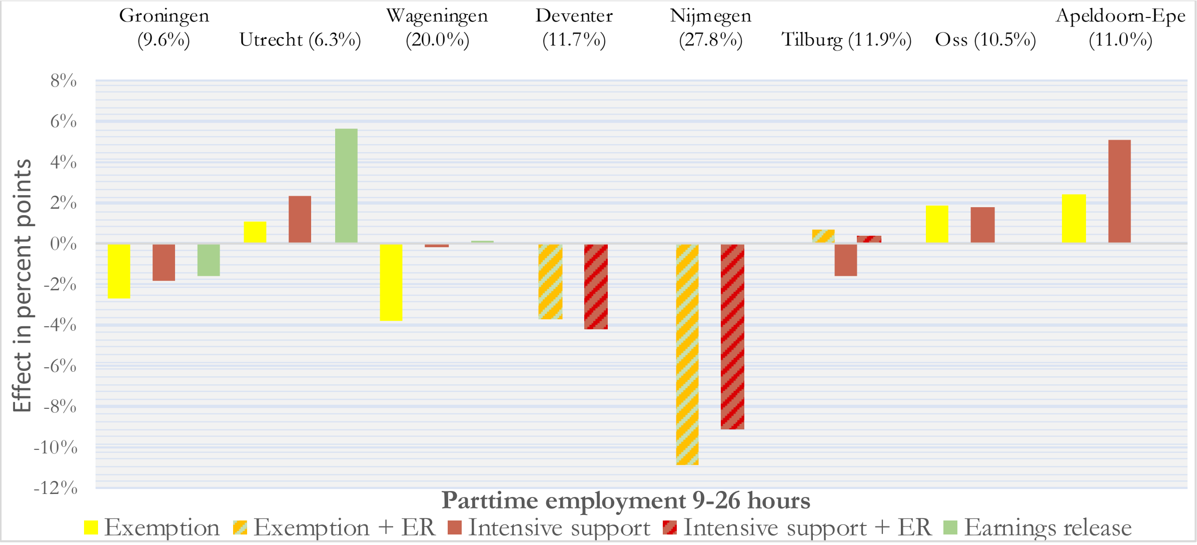


Figure 5. Net effects of exit into parttime employment, 9-26 hours a week

Figure 5 shows a larger negative pure parttime effect now for the exemption group in Wageningen. The positive effects for the exemption group in Oss and Apeldoorn-Epe are also strongly reduced. It appears that the exemption treatment is more likely to harm exit into parttime jobs than that it stimulates working in these parttime jobs. It might mean that if people have to search for jobs themselves, they accept or search less for parttime jobs but prefer jobs with longer hours and more security probably because the majority of small parttime jobs is temporary agency work lasting rather short. The earnings release group show an insignificant and small negative effect on parttime employment in Groningen and Wageningen but a positive and significant effect in Utrecht. Finally, intensive support show positive and significant effects for exit into parttime jobs in Utrecht and Apeldoorn-Epe but small negative effects in Groningen, Nijmegen and Tilburg.[35]

These findings provide very dissimilar and mixed evidence on the hypotheses about the expected positive employment effects of tailored support and earnings release (H3 and H5). Earnings release or an extra work bonus does not increase exit into fulltime work but seem to increase exit into parttime work in some cities. Tailored support seem notably for people with longer stays in welfare and weak employment chances raise their employment opportunities on parttime as well as fulltime jobs but again only in some cities.

6.4 LATE-regression results

LATE-regression models were estimated for the eight cities to calculate the average treatment effects for the people who actually got the treatment instead of all people who were assigned to the treatment as is done in the ITT-analyses presented before. The researchers calculated lower and upper bounds for the estimated LATE-effects[36]. Also, in the CPB-report ITT and LATE-results are presented but only for the six official experiments[37]. In most municipalities the LATE-effects are very similar to the ITT-results because of high compliance rates (most people assigned also got the treatment). Low compliance rates are only found in Groningen and Deventer where randomization took place before submission to the treatment. The CPB report has indicated that due to experiment effects in the control group (John Henry), the results for Groningen, cannot be attributed to the treatment only. But only in Groningen the treatment effects can be compared with those for the randomized reference group of non-participants instead of the control group. Compared to that randomized comparison group no significant ITT-effects were found for Groningen. For Tilburg and Wageningen we have used the data for, respectively, 23 and 24 months after the start. The overall picture is very similar to the ITT-results with significant and positive effects for the 8 hours employment measure in Utrecht but not for fulltime exit (27 hours or more). Negative significant effects are again found in Nijmegen (for exemption and intensive support) whereas the negative significant ITT-effect for Tilburg for intensive support turned insignificant. In both cities, the effects could not be

[35] For these three cities the negative effects could not be assessed with security possibly due to selection (Nijmegen) or experiment effects (Groningen and Tilburg).

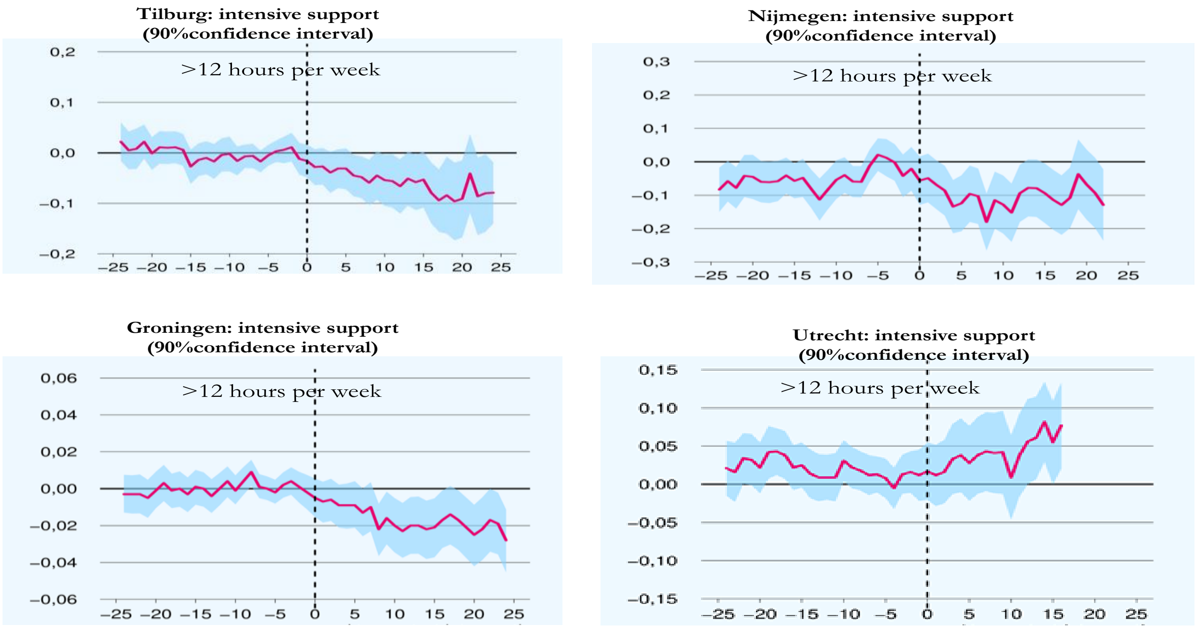
[36] The analyses give a lower and upper bound for the LATE-effects but only the lower bound is given (cf. Gerber & Green, 2012). This is obtained by assuming that everyone who is assigned to the treatment always receive the treatment whereas the upper bound assumes that everyone who drops out did never receive the treatment. Also, in the CPB-report only the lower bound is presented.

[37] The evidence for Tilburg and Wageningen in the CPB-report is however based on register data covering a period up to 1 October 2019. For Tilburg effects are calculated for 16 months and for Wageningen 24 months from the start. In the ITT-results, presented earlier, we used the updated results for Tilburg and Wageningen, both at 24 months after the start.

assessed with security due to accidental randomization differences (Nijmegen) or implausible outcomes for the control group compared to the treatment groups (Tilburg) already at or shortly after the start of the experiment. The observed effects can therefore not causally be attributed to the treatments in these two cities. In Annex 2 (Table A2.2) we therefore present the LATE effects for seven cities only, leaving out Groningen.

6.5 ITT-analyses: placebo regressions and experiment effects

In the third and fourth step selection and experiment effects were further investigated. For investigating selection effects placebo regressions were performed which show the employment effects of the participants in each treatment group from already 25 months before the start of the experiment (see also De Boer et al. 2020). Experiment (John Henry) effects, were investigated by inspecting the employment effects from the start up to month 24 of the experiment for the control group compared with the treatment groups. For two cities, the CPB performed a formal test of the existence of John Henry effects by comparing the employment outcomes of the control group with a randomized reference group (Deventer, Groningen). The employment outcome measures for fulltime (27 or more hours a week), parttime (8 or 12 hours or more a week) and 70% of the minimum wage were used for the analyses. Here we only show the figures for the 12 hours measure (including hence also fulltime work) but the results for the pure fulltime measure are more or less similar.



Source: De Boer et al. 2020

Figure 6: Placebo regressions on "outflow to paid work of 12 hours or more", for the ***intensive support*** *group, compared to the control group in Tilburg, Nijmegen, Groningen and Utrecht (the blue cloud depicts the 90% confidence interval).*

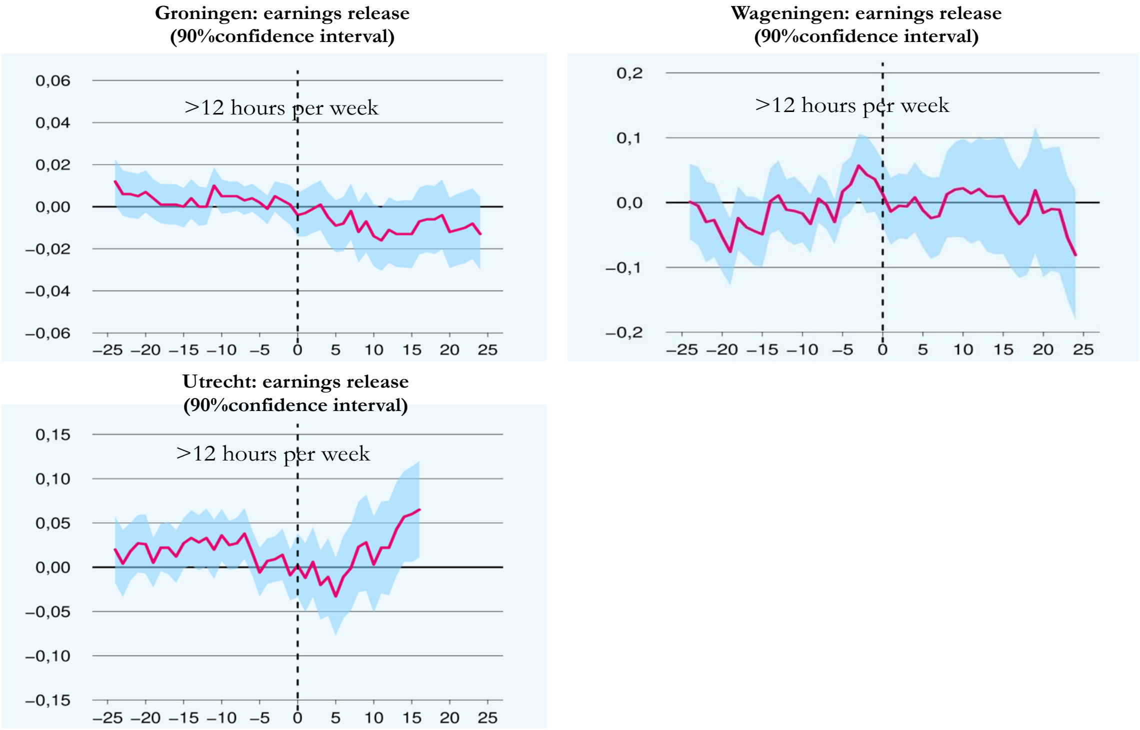
The set of covariates used in the regressions is similar to the ones for the representativity and balancing tests: age, gender, living situation, country of birth, education, spell duration and wage income measured at 3-6 months before the start. We only present the results for the intensive support treatment in four of the six cities for which the effects were calculated. Further evidence is found in the CPB report (DeBoer et al. 2020). This analysis reveals to what extent the treatment group compared to the control group already behaves differently before the start of the experiment due to selection effects caused by a different employment history. This is especially relevant for the cities for which we found negative employment effects in month 24 such as Groningen, Tilburg and Nijmegen. The results for these cities are compared with those for Utrecht where we found positive employment effects for the 12 hours measure. The findings suggest that notably in Nijmegen negative selection effects caused by accidental randomization differences might partly explain the negative employment effects found for the intensive support group. For Utrecht, on the other hand, we found small positive selection effects. For Groningen and Tilburg no selection effects were found.

Experiment effects

The figures also show the employment outcomes in the period from the start of the experiment up to 24 months later. If the employment effects already shortly after the start (when no treatment effect can be expected) moves in a negative or positive direction compared to the control group, it might indicate the existence of experiment effects (John Henry) in the control group provided there are no selection effects. For Nijmegen, we observed selection effects but in Tilburg and Groningen experiment effects might be at stake because the intensive support group shows already at the start significantly lower exit rates into employment of 12 hours or more than the control group. These negative effects tend to become more significantly negative over time. In Utrecht the picture is just the opposite. Already at the start the intensive support group performs better than the control group and the significant positive effects tend to increase over time. The Utrecht picture is the one that one would expect from providing more intensive counselling support to welfare beneficiaries.

The same analysis was done for the other treatments of exemption and earnings release. In Figure 7 we show the same ITT-regression analysis but now for the "earnings release" treatment and only in three cities, Groningen, Wageningen and Utrecht. In these cities the earnings release treatment is a single treatment, not combined with exemption or extra support, as it is in Nijmegen, Deventer and Tilburg. Again, the results show that for the 12-hours measure notably in Groningen the earnings release group exhibits negative employment effects at the start which tend to become more negative over time. In Wageningen they were only slightly negative at the start, alternating negative and positive up to month 20, after which the effects turn strongly negative up to month 24. The evidence for Utrecht shows a slightly negative effect at the start that turns more negative in the first 6 months after the start, but then showing rising positive effects after month 9 up to month 16.

These results for both treatments indicate that for Tilburg and Groningen John Henry effects might be at stake because of which the treatment effects cannot fully be causally attributed to the treatment.

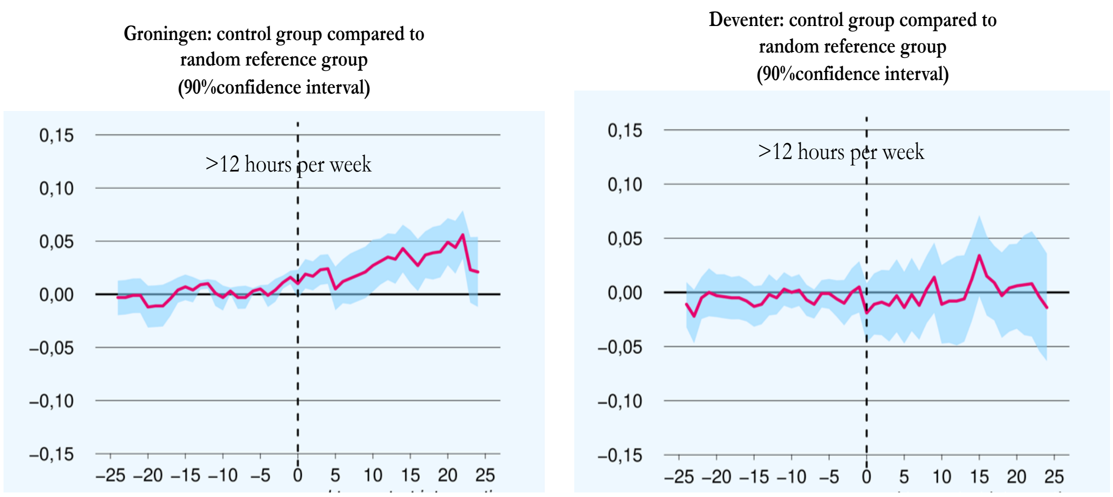


Source: De Boer et al. 2020

Figure 7: Placebo regressions on outcome measure: "outflow to paid work of 12 hours or more", for the ***earnings release*** *group, compared to the control group in Groningen, Wageningen and Utrecht (the blue cloud depicts the 90% confidence interval)*

A formal test

In the CPB-report a formal test is conducted on these John Henry effects in Deventer and Groningen, because only in these two cities a comparison was possible between the outcomes for the control group and a randomized comparison group of non-participants. The results are presented in Figure 8. The comparison between the control group and the randomized reference group shows that John Henry effects are likely to exist in Groningen but not in Deventer.



Source: De Boer et al. 2020

Figure 8: Graphical test for John Henry effects on "outflow to paid work of 12 hours or more", for the ***control group****, compared to the randomized reference group in Groningen and Deventer (the blue cloud depicts the 90% confidence interval)*

7. Results on survey outcome measures

In this section we present the results on some but not all of the survey outcome measures. The focus is on measures which are most relevant for the Technequality project: job search, self-efficacy in finding work, health and wellbeing and social and institutional trust (trust in caseworker) and financial stress. They are selected because they either directly (job search, self-efficacy) or indirectly (health and wellbeing, trust, financial stress) affect people's labor market behavior. The indirect measures influence people's behavior through their impact on their perceived quality of life (subjective health and wellbeing), their social capital (social and institutional trust) and their mindset (financial stress). Trust in other people is a commonly accepted measure of social capital because it is associated with the broadness and quality of people's contacts in their social network, that is found to impact people's job opportunities (Coleman 1988; Knack & Keefer 1997; Uslaner 2002). But also trust in the supporting institutions notably the municipality and the caseworker impact people's mindset and their job search behavior. Much of our expectations with respect to the effects of the various treatments on job search and employment are based on behavioral economics and motivational psychology frameworks as explained in the theoretical section. In this section we present the first results as presented in June 2020 in the local reports, the joint summary report and the CPB report.

7.1. Survey data: some methodological concerns

Before, we explained that to correct for selection, we control for possible differences in baseline outcomes by including the first wave outcome in each model for the various outcome measures (cf. Eq. 2). The empirical model is now tested on the second and third wave with the baseline outcome level in wave one included in the model. Here, we only present the estimation results using the last, third wave. However, the response rates in the second and third wave are in various cities (Deventer, Utrecht, Nijmegen, Tilburg, Wageningen, Apeldoorn, Oss) lower than those in the first wave. This is partly due to dropping out of the experiment (moving to other city, retirement, death) but also to non-response notably for those who found a paid job and moved out of the welfare registers[38]. In some of the local reports (Nijmegen, Tilburg, Wageningen, Oss) wave two and three results are estimated showing that with various outcome measures (wellbeing, health, social participation, self-efficacy) the gradient of the lines showing the effects of the treatment for the three waves have a downward kink after the second wave. Further scrutiny is needed how we can correct for this selection effect. A second issue to be mentioned is that in some cities the baseline or wave one survey is not filled in by the respondents directly at the time of registration but at a later moment, in some cases (such as in Tilburg, Oss and Apeldoorn-Epe) one or even some months later. This means that part of the effect of the treatment is already taken up in the baseline level and does then not show up in the measured effect at wave three[39]. More scrutiny is needed to what extent the time span between the start of the experiment and the response to the questionnaires might have affected the results.

7.2. Results on job search, health, wellbeing, trust and financial situation

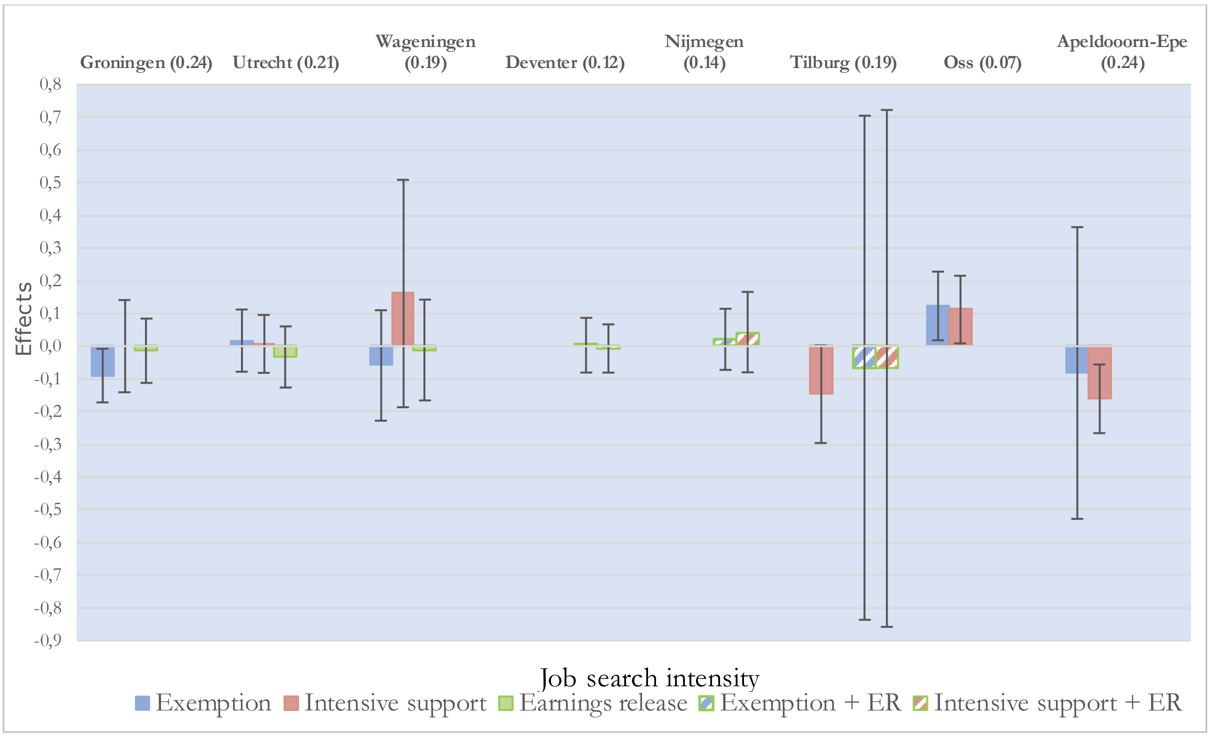
Detailed estimation results of the ITT-effects regressions on these set of outcome measures are given in Annex 4 (Tables A4.1. to A4.3). Below, we present the outcomes for the selection of outcome measures mentioned earlier. First, results are given for job search intensity and self-efficacy in finding work.

Job search intensity

The average score ranges between 0.19 and 0.24 except for Oss showing a rather low score (0.07). Most effects are insignificant except for the exemption group in Groningen and the intensive support group in Tilburg and Apeldoorn-Epe showing significantly negative effects. Also, for Wageningen and Apeldoorn-Epe the results for the exemption group are negative but insignificant. However, the effects are positive for the exemption group, as well as for the intensive support group, in Utrecht, Oss and Nijmegen. In Nijmegen the exemption and intensive support treatment is combined with earnings release. The earnings release group show in three cities where it is a single treatment, Groningen, Utrecht and Wageningen, negative but insignificant effects on job search. The results appear hence, rather mixed while the effects vary widely between the cities and groups but also within the groups indicating that the content of the treatment is rather different across the experiments.

[38] In some local reports for Tilburg, Wageningen, Oss, Apeldoorn and Deventer it is shown that exit into paid work have lowered the response rates in the second and third wave. In Deventer, a separate survey was held with the people who have found work showing a response rate of 45% instead of 75% on average for the three waves.

[39] For that reason, also an alternative model specification is tested in the local reports for Tilburg, Wageningen, Oss and Apeldoorn in which the estimates are based on the second and third wave controlling for wave and the baseline outcome. The results are not very different from the ones presented here.



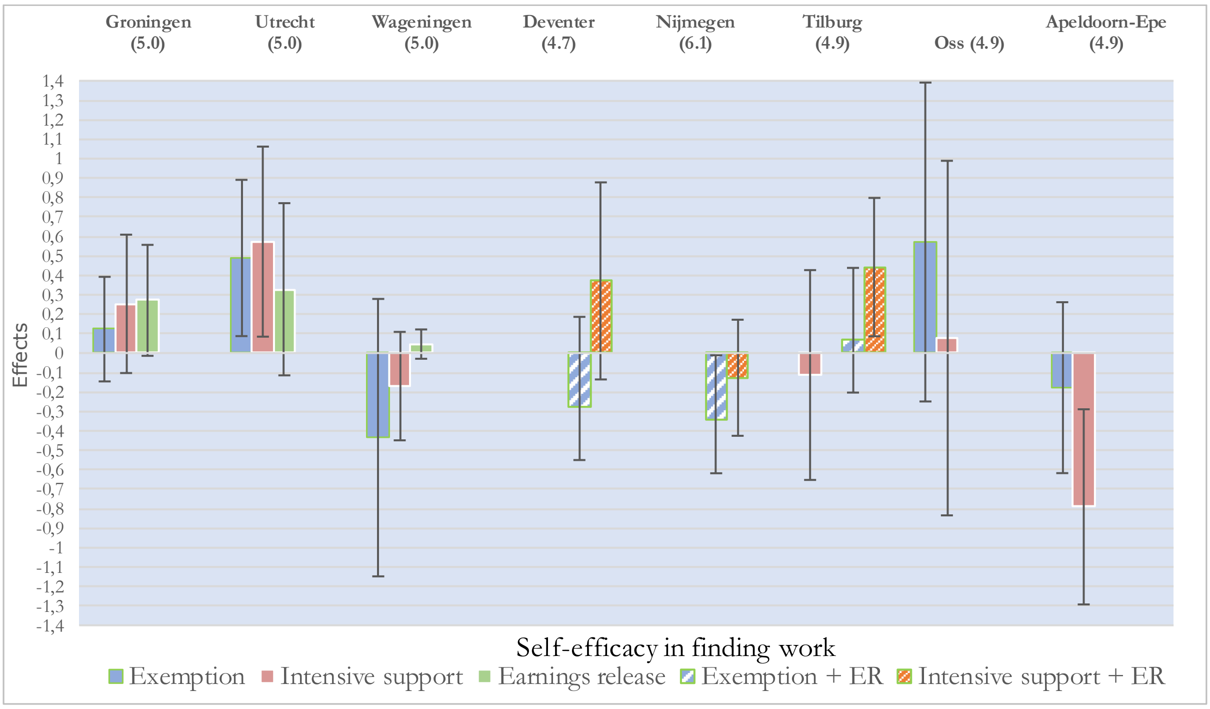
Note: Treatment effects compared to the mean scores for the control group in and 90% confidence interval. Results for the control group (means) are given in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. For Groningen the comparison group is the control group.

Figure 9. Effects on job search intensity

With respect to exemption, recall, that the exemption treatment in Groningen comprises a relaxation of the obligations while leaving people more or less alone. In Utrecht and Nijmegen there is only limited contact with the people and mostly when he or she requests for it. In Tilburg, Wageningen, Oss and Apeldoorn-Epe the treatment in the exemption group is somewhat different. Caseworkers got a training how to support beneficiaries to become more self-reliant. That means that in the beginning there is more frequent contact with the participants that over time is reduced. But also the intensive support treatment unveil a large variation in the content of the treatment, partly because it is combined with earnings release in Deventer, Nijmegen and Tilburg and partly because municipalities implemented the treatment differently. It emerges that the actual number of contacts with the participant varies, that in some cities because of the experiment the caseload of the caseworker was more reduced than in other cities and that the quality and intensity of the support provided by the caseworker vary. Viewing the results for the exemption group, it appears that in five cities the results are negative and in Groningen even significantly negative. In these cities there is in this treatment little contact with the people which appears to reduce job search efforts instead of improving it. The negative effects for the single earnings release treatment where apart from earnings disregard ‘care-as-usual’ is provided with scarce contact, might also be associated with the absence of counselling support. Combined with the positive effects for the intensive support groups in three cities, both results suggests that counselling support might help for effective job search. The evidence on job search intensity therefore provides mixed evidence on the first and second hypotheses according to which positive effects of reducing stress and freeing people's mindset (H1) and increasing people's autonomy and freedom (H2) are expected on job search when the obligations are relaxed as in the exemption group. The evidence shows, cautiously, that in addition to giving people with weak employment prospects more autonomy and freedom, counselling support is needed. In day-to-day practices people with longer residence in welfare have little contact with their caseworker and are more or less left alone. That seems not the best way to reintegrate people in employment.

Self-efficacy in job search

A related concept concerns people's perception about their self-efficacy in finding a job. The average scores are low and very close, ranging from 4.8 (Utrecht) to 6.1 (Nijmegen) on a scale from zero to ten. For perceived self-efficacy in finding work, positive significant effects are only found in Utrecht, for both, the exemption group and the intensive support group.



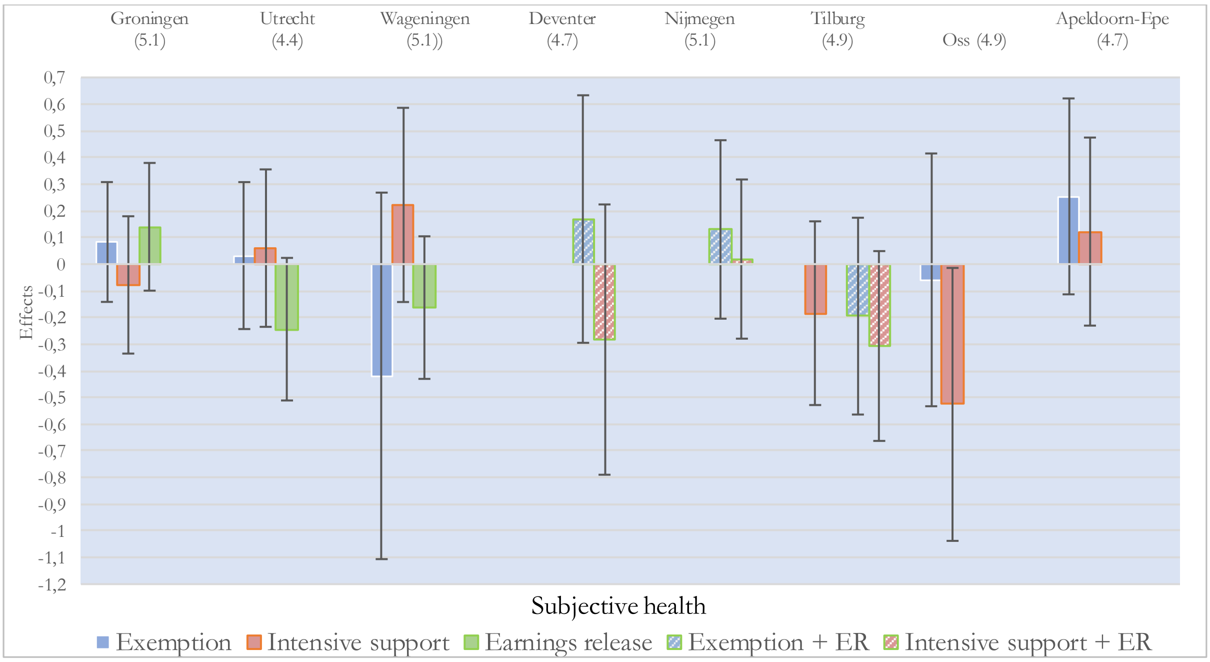
Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) are given in the heading . For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. For Groningen the comparison group is the control group.

Figure 10. Effects on self-efficacy in finding work

Also, in Groningen and Oss positive, but insignificant effects are found for both groups whereas in Wageningen and Nijmegen negative but insignificant effects were found. One other city, Apeldoorn-Epe, shows however negative significant effects for the intensive support group. For the earnings release group we found positive but insignificant effects on self-efficacy in Utrecht, Groningen and Wageningen. When intensive support is combined with earnings release such as in Deventer, Tilburg and Nijmegen, positive but insignificant effects are found, but only in the first two cities. In Deventer and Nijmegen where exemption is combined with earnings release, negative but insignificant effects are observed. The reason for the positive effects of the single treatment of earnings release in the three cities might then be that if people are supported and rewarded for their employment search through the extra work bonus, they become more confident about their own opportunities in finding a job. The positive effects on self-efficacy when earnings release is combined with intensive support also suggests that counselling support combined with earnings release might improve people's job search efficacy.

Subjective and mental health

In the next two graphs we present the results for subjective health and wellbeing. First, we look at subjective and mental health. The average health scores on a scale from zero to ten vary between 4.4 in Utrecht to 5.1 in Nijmegen and are rather low. Overall, the results show no significant health effects for any of the treatments except for intensive support in Oss which effect is negative. Moreover, the health effects appear rather mixed across the cities showing positive effects in Utrecht, Nijmegen and Apeldoorn-Epe for the treatments, exemption and intensive support, but negative effects for these treatments in Oss and Tilburg.



Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. Also, for Groningen the comparison group is the control group.

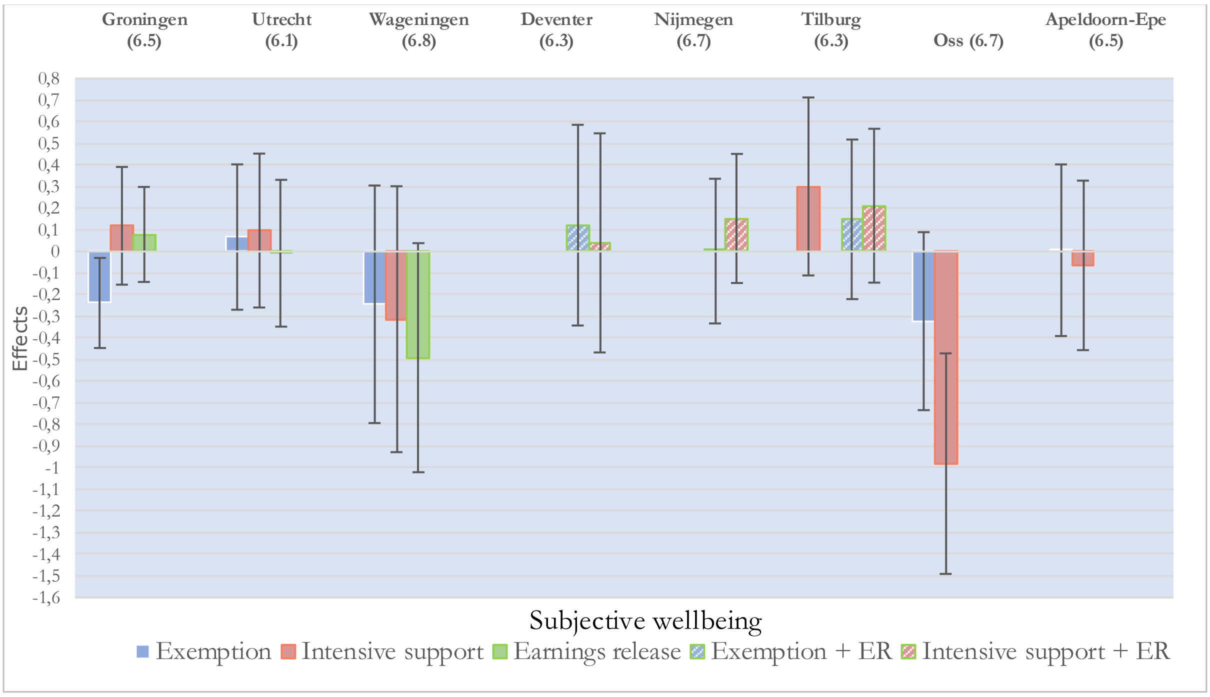
Figure 11. Effects on subjective and mental health

For Oss, the local report suggests that this is not only a treatment effect because part of the effect might be attributed to selection effects (worse work and health conditions of the control group already at baseline and survey non-response of people who have exited to paid work). In Tilburg experiment effects with the control group, being a group with better work chances, might have affected the results. In Deventer and Groningen positive effects are found for the exemption group but negative ones for intensive support. In Wageningen positive effects show up for intensive support whereas the exemption and earnings release group show negative effects. The extra financial support might only have a small positive health effect for the earnings release group because the work bonus is rather modest. Insignificant but negative health effects for this group were found in Utrecht and Wageningen but small positive effects in Groningen.

The overall conclusion might then be that no health effects are found in any of the cities and that any of the treatments perform clearly better on health outcomes than the other. The large variation in health outcomes across the cities might be related to the different way of implementation and content of notably the exemption and intensive support treatment. These results contradict the expectations we have formulated in hypothesis 4 on the positive effects of exemption and notably tailored support on levels of perceived health and wellbeing. The reason might be that the interventions were not aimed at improving people's health and that two in three beneficiaries on welfare perceive their health as rather poor. It is therefore unlikely that a small intervention as provided in the treatments might have a strong effect on people's subjective health. It might also be that it takes a longer time before health effects show up. The literature further suggests that specific employment support tailored to the specific health impairments of the person but also to the workplace and the needs of the employer might be more effective than a general and small intervention as performed in this experiment (Marshall et al. 2014).

Subjective wellbeing

The average subjective wellbeing level for these beneficiaries is rather low with scores ranging from 6.1 in Utrecht to 6.8 in Wageningen in the control group whereas the average level (for life satisfaction) is about 7 to 8 for the entire population (CBS 2017). The picture for the wellbeing effects of the various treatments is slightly more positive then for subjective health. In two cities, a negative significant effect was found. That is Groningen for exemption and Oss for intensive support. Likewise, to what is found with subjective health, the effect in Oss might not be attributed to the treatment but to selection effects. Viewing the results for the other cities it appears that the wellbeing effects for exemption and intensive support are insignificant but positive in Utrecht, Nijmegen and Tilburg. In Wageningen all treatments show a negative but insignificant effect. Generally, the positive effects are larger for intensive support than for exemption. For earnings release the picture is mixed with positive effects in Groningen, no effect in Utrecht and negative effects in Wageningen. Again, the results contradict hypothesis 4 but the results come not as a surprise. From the literature we know that there is substantial variation in subjective wellbeing but also that the level of wellbeing is rather stable over time. It is only because of serious life events like the death of a child, an employment shock or a serious illness that wellbeing levels drop for a longer time (Kunzmann et al. 2000; Helliwell and Huang; Headey et al. 2010). Likewise, to what we concluded for subjective health, it is therefore very unlikely that a small intervention like the ones we tested here will have a strong effect on people's subjective wellbeing.

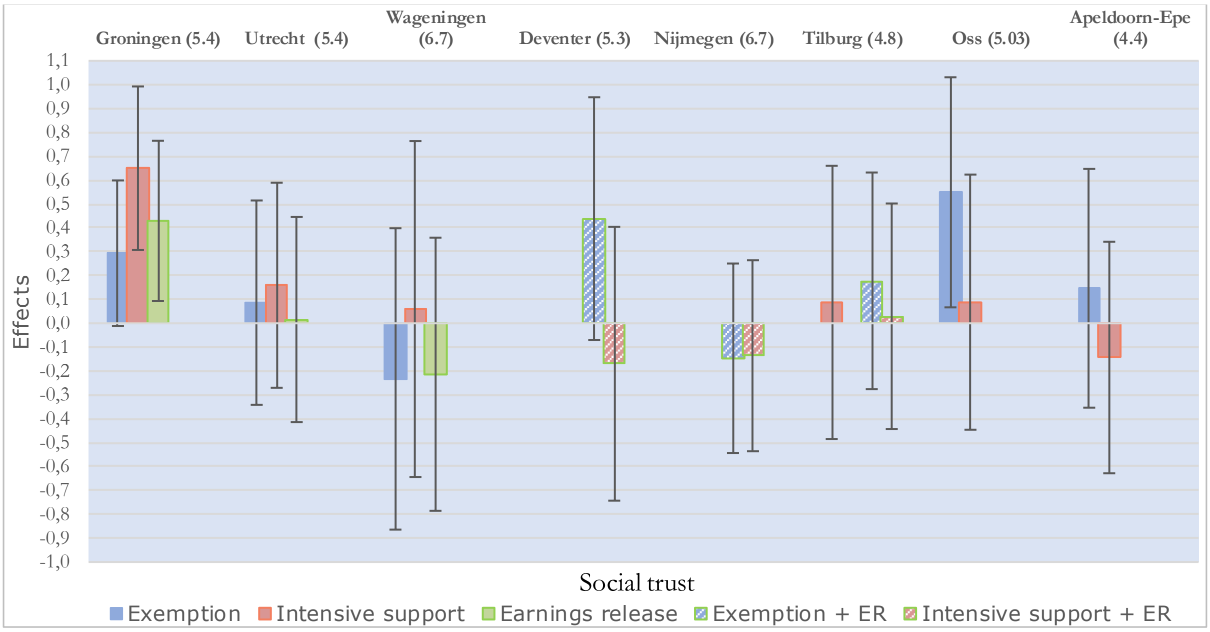


Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. Also, for Groningen the comparison group is the control group.

Fig. 12. Effects on subjective wellbeing

Social trust

Before, we argued that social trust is a yardstick for people's social capital and therefore important for re-integration into employment. Figure 13 shows the average scores on social trust that ranges between 4.4. in Apeldoorn-Epe and 6.7 in Wageningen and Nijmegen. In most cities, except for Wageningen and Nijmegen, positive effects are found for all treatments but the strongest positive and significant effect was found for intensive support in Groningen. For Oss a positive and significant effect was found for the exemption treatment. Positive, but insignificant effects on all treatments were found in Utrecht, Tilburg and Oss. In Nijmegen, Deventer and Apeldoorn-Epe, negative but insignificant effects were found for intensive support. Note also the positive but insignificant effects for the earnings release group in Groningen and Utrecht. For the combination treatments of exemption and extra support with earnings release, the results are mixed with negative insignificant results in Nijmegen (both treatments) and in Deventer (intensive support with earnings release) whereas positive effects are found in Tilburg for both treatments. Overall, we find positive effects on social trust (the trust put in other people) for most treatments, and in most but not all cities. If people convey more trust in others and become more active in their social networks, their social capital increases which can have positive effects on employment but also on wellbeing and health (Coleman 1988; Knack & Keefer, 1997). It is therefore an important yardstick also for the judgement of these experiments on their contribution to social participation.



Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. Also, for Groningen the comparison group is the control group.

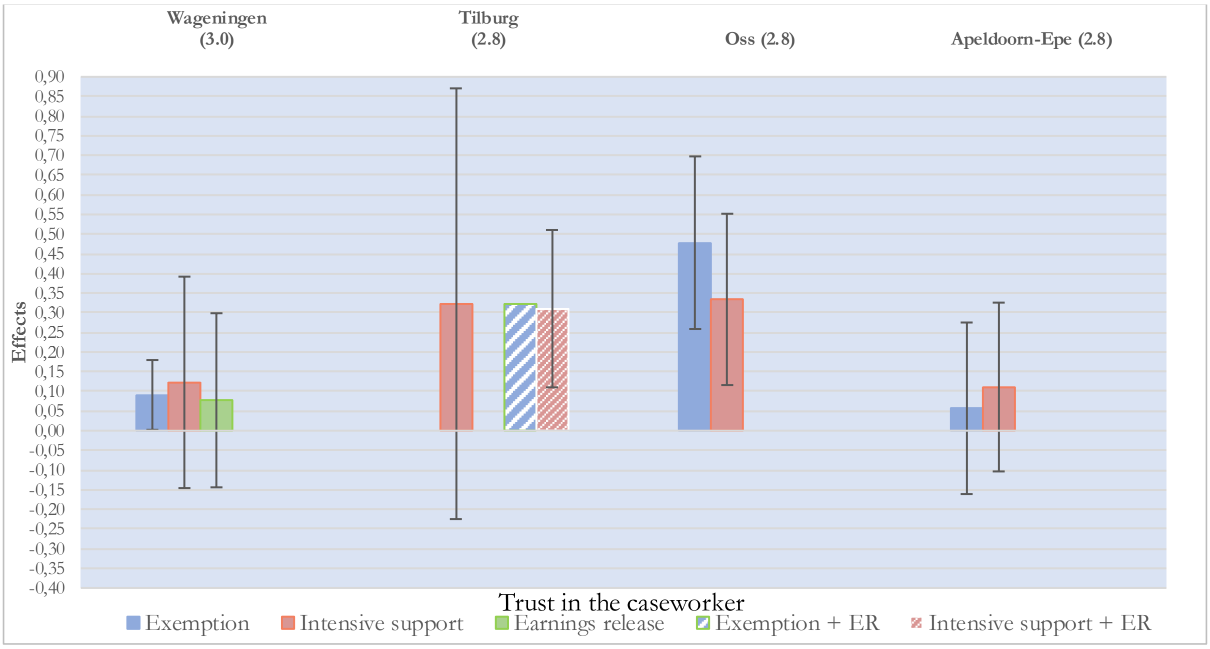
Figure 13. Effects on social trust

Trust in the caseworker

In some cities (Tilburg, Wageningen, Oss, Apeldoorn-Epe), the experiments were labelled 'trust' experiments to express that providing people with more autonomy and hence putting trust in their self-management capacities (exemption) might in the end work better than regular 'workfare' practices unveiling distrust (based on 'strict conditionality and tight monitoring and control'). A second trust measure used in the research is therefore the trust people put in the caseworker of the welfare department. It is one of the components of institutional trust[40]. We expect that a 'trust-focused' approach will result in higher levels of trust of the participants in the caseworker and that this might pay-off in terms of more active job search (exemption and extra support) and higher levels of subjective health and wellbeing. The results are presented in Figure 4.6. The average caseworker' trust scores appear more or less the same in the four cities under scrutiny. The findings show positive and strongly significant effects in Tilburg for intensive support

[40] The question is also asked in two other cities, Utrecht and Nijmegen, but not reported in their local reports and therefore not presented here.

and intensive support combined with earnings release and in Oss for both treatments, exemption/self-management and intensive support. In the other two cities positive but insignificant effects show up. The conclusion can therefore be that the more relaxed treatment and the extra attention (more frequent contacts) and support provided improves the trust-relationship of the participant with the caseworker at least in two of the four cities because of which the supervision might become more effective.



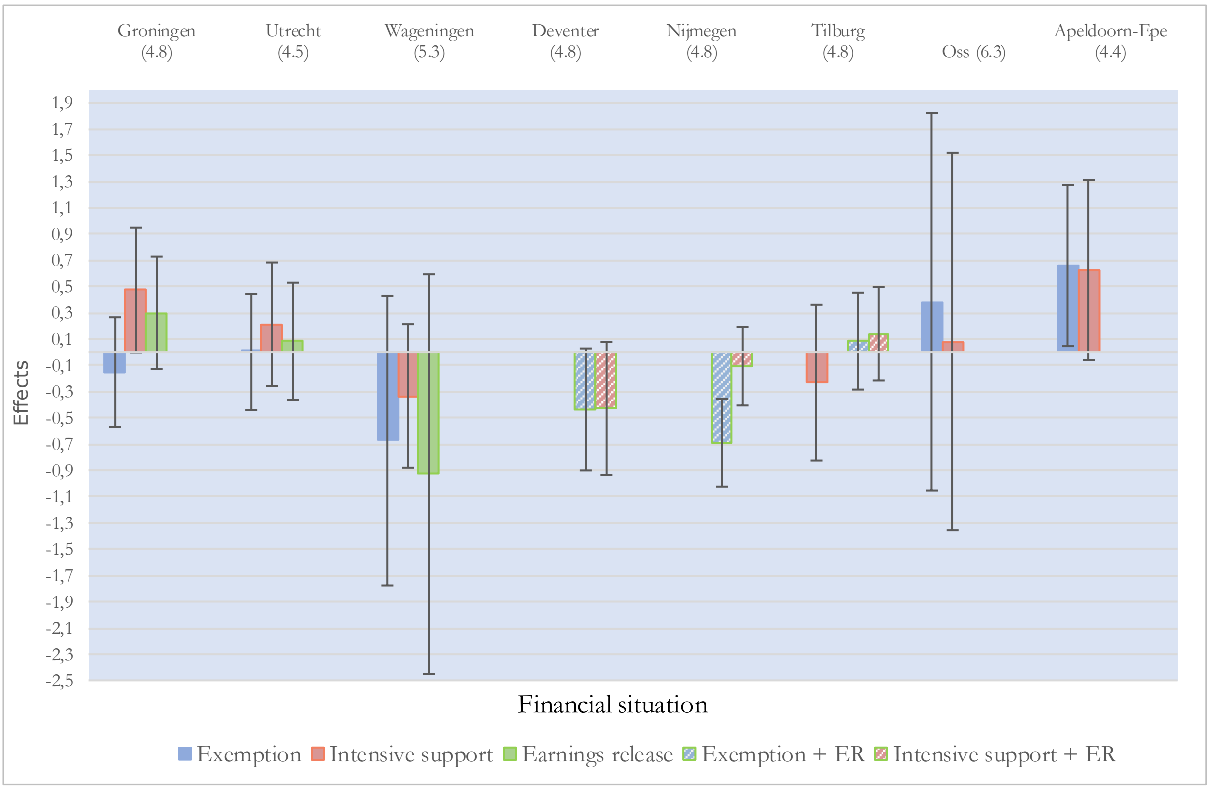
Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. Also, for Groningen the comparison group is the control group.

Figure 14. Effects on trust in the caseworker

In hypothesis two it was contended that providing autonomy and trust evoke positive feelings of reciprocity which converts into extra job search efforts. The findings here for two out of the four cities under scrutiny show that treatments based on this idea of putting more trust in people indeed might lead to increased levels of trust in the caseworker which might in step two lead to increased search efforts. The earlier findings on the effects of the treatments on job search suggests that only when more autonomy and trust is combined with counselling or active support it may trigger active job search notably for people with a large distance to the labor market .

Financial situation

Finally, we viewed the effects on perceived financial situation indicating financial stress. The average scores are very low and range between 4.4. in Apeldoorn-Epe and 5.3 in Wageningen on a scale from zero to ten. It shows that people experience serious financial stress also because a substantial proportion of the participants have to make debts or dissave to make ends meet (20 to 30%). Except for Apeldoorn-Epe where we found positive significant effects on the financial situation for exemption and nearly significant effects for intensive support (p<0.15), and for Wageningen and Nijmegen, where we found negative significant effects for earnings releases and exemption combined with earnings release respectively, most effects are insignificant. In Groningen positive effects are observed for earnings release and nearly significant effects for intensive support (p<0.11). Also in Utrecht and Oss, positive but insignificant effects were observed for all treatments and in Tilburg for exemption and intensive support combined with earnings release.



Note: Treatment effects compared to the mean scores for the control group and 90% confidence interval. Results for the control group (means) in the heading. For Utrecht the outcomes are based on a slightly different model specification controlling for 24 months of earnings history instead of 6 months. Exemption + Earnings Release in Tilburg includes an extra work bonus for fulltime exit. Also, for Groningen the comparison group is the control group.

Figure 15. Effects on financial situation

In Wageningen and Deventer all effects appeared negative but insignificant. Again, the evidence is mixed across cities and across and within groups. In four of the five cities with intensive support as a single treatment a positive effect was found that was (nearly) significant in two cities. For exemption we found positive effects in four cities, of which one was significant, but also negative effects in four cities. People in the earnings release group are rewarded for working more hours with an extra work bonus which could also improve their financial situation. In two of the three cities with earnings release as a single treatment we find positive but insignificant effects on the financial situation whereas in the third city the effect is significantly negative. It appears, cautiously, that notably intensive support as a single treatment might contribute to supporting people in reducing financial stress caused by a poor financial situation or problematic debts.

8. Conclusions and reflection

Before we discuss the overall results of the eight experiments, we reflect on some methodological issues we came across and from which lessons can be drawn for future experiments.

- *Substantial differences in design across the various cities.* In this report we decided to research the outcomes of each experiment separately, because the design, selection of target groups, content and implementation of the treatments were very different across the cities making an overall analysis complicated[41]. In a next step of the Technequality project we might attempt to assess joint effects of comparable treatments in the various cities even though content and implementation differences cannot be excluded complicating the interpretation. For future experiments a uniform design with a view to randomization, target population, treatments and implementation is warranted.
- *Changes in implementation of the regular treatment.* The RCT design is ideally suited to examine causal effects but some strict methodological requirements need to be met to be able to make sound inferences from the outcomes. One of these is that the standard or 'care-as-usual' treatment does not change during the experiment. But that is difficult to safeguard when policies change and the implementation practices also change. The RCT experiments are field experiments which were set up in a local policy context where the municipalities were the owners of the experiments, not the researchers. Also, the chosen and carefully designed alternative treatments need to be maintained during the experiment and may not be changed even though a caseworker or the management might be convinced an alternative treatment would work better for this particular person or group.
- *Selection and experiment effects (John Henry).* For investigating selection effects, (placebo) regressions were executed showing the ITT-employment effects for the various groups already 24 months before the start up to 24 months after the start. For Nijmegen we found that due to selection the control group already behaved differently in the 24 months period before the start of the experiment which is presumably caused by accidental randomization differences. In some cities

[41] The experiments differ first because there were official experiments and informal experiments. The informal experiments were not permitted to change the earnings release rules so they had two treatments only. For the official experiments there are cities with single treatments, such as Groningen, Utrecht and Wageningen, and others with combination treatments such as Deventer, Tilburg and Nijmegen. Two cities used pre-randomization (first randomize and then invite) and all others post-randomization. The combination treatments varied because the components vary. Tilburg had an extra work bonus on top of the extended earnings release. Single treatments also varied. Exemption might mean 'leaving people alone' or ' learning self-management' and the content of extra support depends on the number of contacts, the tools available for support, the caseload and the local policy context that impacts the standard treatment.

such as Groningen and Tilburg no evidence was found for selection but it emerges that the control group already at or shortly after the start behaves very differently from the other groups. This suggests that John Henry effects are in play. The treatment effects that were found can therefore causally not be solely attributed to the treatments, but may also be related to experiment effects. Further scrutiny is needed to be able to correct for these experiment effects. Comparison with the randomized comparison group of non-participants in e.g., Groningen is one of the options but also the use of other statistical models (like PSM-propensity score matching or fixed-effects panel regression models) to control for selectivity in the non-randomized comparison group of non-participants (e.g., in Tilburg and Oss) might be an option. Eventually, also comparison with a randomized group of recipients in other cities receiving the regular treatment might be an option.

- *Limitations due to the design of the RCT field experiments.* The official experiments were designed according to an evaluation format that is developed by ZonMw (the Netherlands Organization for Health Research and Development) and the 'Temporary Regulation Experiments Participation Act' of 1 April 2017. The regulation imposes a number of limitations on the design (a low ceiling for earnings disregard, the inclusion of a special treatment with stricter monitoring and control, combination of exemption and extra support is not allowed, and a two-year instead of a three-year period as asked by the researchers). The local context demanded combination treatments instead of single treatments, particular population groups were excluded (such as youngsters) and the size was limited to reduce costs. Apart from large local differences in design and implementation, these requirements also resulted in small N in each treatment (see next point) and, furthermore, in small interventions (such as the low level of earnings disregard) because of which large effects could not be expected.
- *Lack of power to analyze heterogenous effects.* The small N in each treatment complicates the analysis of heterogenous effects (interactions). It was not possible to disaggregate the analyses to subgroups such as young and older people, people with short and long spell durations, low skilled versus high skilled people or people with bad and good health. The results presented here refer to average effects for each treatment group, but the various treatments might 'work for someone but not for the other'. It also means that the individual effects might cancel each other out within a particular group. Further analyses where similar treatments across cities are joined might lead to more power and effects estimated with higher levels of reliability[42].
- *Lack of power and low significance.* Design differences resulted in small N of each treatment in each city leading to limitations in power. This might explain why we in the end found few, small and insignificant effects. Additionally, in quite a few cases the effects were just insignificant at the 90% level but significant at 85% level. The alternative explanation that 'random significance' might be an issue since the number of significant effects compared to the total number of effects to be

[42] For these analyses the approvement of the municipalities is required also because the outcomes might change.

estimated is rather small, is presumably less relevant because the tests are carefully and separately designed, implemented and prepared in the various cities. Most research instruments were also already validated in previous surveys. Finally, the experiments followed a methodological format for RCT research evaluated by a team of researchers with a high expertise in this type of research of the national science foundation that has judged the tests as academically sound.

- *Model specification and sensitivity analyses.* Several alternative model specifications were estimated but in all models we included spell duration and education level to correct for selectivity but these variables might take-up part of the treatment effect when the treatment works better for people with longer or shorter spell durations or for the low versus high skilled. Interaction effects between treatment, skill level and spell duration might provide new insights in why particular treatments works better or worse for specific people or subgroups. Alternative model specifications with more controls for design and implementation (number of contacts, caseworker, applied tools) and alternative estimating techniques might also help to explain why the effects differ between treatments and cities.

Substantive results

This report presents the first results of the study into the outcomes of these eight experiments. Viewing exit into fulltime and parttime employment, most effects are insignificant except in a few cities such as in Utrecht for exit into parttime work (notably earnings release) and Apeldoorn-Epe for exit into parttime and fulltime work (intensive support)[43]. In some cities, such as Wageningen and Oss, positive but just insignificant effects ($p<0.15$) were found for the exemption group and extra support group for exit into fulltime work. That some positive effects are just not or nearly significant is likely the result of small numbers (lack of power) notably in the smaller cities (Wageningen, Nijmegen and Oss). In cities where the outcomes were negative such as in Nijmegen (exemption and extra support combined with earnings release), Groningen (exemption and intensive support) and Tilburg (intensive support), the CPB-report (De Boer et al. 2020) concludes that these effects could not be assessed with certainty. This is either due to selection effects as in Nijmegen caused by accidental randomization differences or the existence of "John Henry" effects (the control group behaves already different at or just after the start of the experiment) in Groningen and Tilburg. This means that due to these experiment effects, the effects can causally not be attributed to treatment effects and are therefore not certain.

Modest effects, but a story to tell

The conclusion can be that the employment effects are apparently modest but they nevertheless have a story to tell. No evidence was found that the implicit assumption of current policies that enforcement and tight monitoring and control of compliance behavior is the best way to get people back into work is

[43] In these calculations exit into self-employment was not included. Future research using updated register data will allow to include exit into self-employment because of which the outcomes might change.

confirmed, but also no (clear) evidence that the alternative treatments achieve better employment outcomes than the regular treatment. The employment outcomes are more or less similar to the outcomes of current workfare practices, while for exit into fulltime and parttime work they are in some municipalities (Utrecht, Oss, Apeldoorn-Epe) better. These latter positive effects emerge notably when tailored counselling support is provided or when recipients have more opportunities to earn extra money through working parttime. This conclusion was also the main conclusion drawn by the CPB in their evaluation report (De Boer et al. 2020).

The *hypotheses* we have formulated on the positive employment effects of exemption, intensive support and earnings release are not confirmed in most cities while being confirmed in a few cities. But in three cities the effects could only be estimated with uncertainty and if we leave out these cities the picture is slightly more positive. Notably for exemption and intensive support positive and significant effects are found for fulltime work and for earnings release for parttime work. Exemption on the other hand seem to harm exit into parttime work possibly because recipients prefer fulltime jobs and therefore confine their jobs search to these jobs. Tailored support seem to improve exit into fulltime as well as parttime employment notably for people with longer stays in welfare and poor employment chances. The theoretical expectations we had beforehand that exit rates into employment would substantially increase, on average with 10 percent points, notably for the intensive support group, have not come through. The effects are on average smaller in size and insignificant.

Job search, health and wellbeing, trust and financial situation

Viewing the results on job search, health and wellbeing, trust and financial situation, the results are more or less similar to the ones for employment. Again, the outcomes differ widely across the various cities and treatments and only a few outcomes appear significant. With a view to our hypotheses some were confirmed others falsified. Hypotheses one and two on the positive effects of exemption on job search, is not confirmed; it seems that having little contact with the people appears to reduce job search efforts instead of improving it. Earnings release combined with regular support seem to harm job search whereas intensive support tends to improve active job search. We concluded that counselling support is needed for effective job search. In day-to-day practices people with longer residence in welfare have little contact with their caseworker and are more or less left alone. That seems not the best way to reintegrate people in employment. In Utrecht positive and significant effects were also found for the exemption and intensive support groups on self-efficacy in finding work. The evidence for intensive support in the other cities is mixed, positive but insignificant in four cities and negative insignificant in three cities. For earnings release positive but insignificant effects on self-efficacy are found in five cities also when it is combined with intensive support. The findings on self-efficacy suggest, cautiously, that intensive counselling support combined with earnings release might improve people's perception of job search self-efficacy.

For *health* we found no significant positive effects in any of the cities and treatments. The obvious explanation is that the treatments are small and not aimed at supporting people with serious illnesses or mental health issues. For *wellbeing* we find overall positive but insignificant effects for exemption and intensive support, where the effects for the latter are shown to be larger than for the former treatment. But, in some cities we also found negative significant effects which might be attributed to John Henry effects in some cities or a selection effect caused by survey non-response of people who found fulltime work. These findings contradict hypothesis 4 on the positive effects of exemption and support on health and wellbeing but do not come as a surprise because from the literature it is known that only serious life events like the death of a child or a health shock tend to drop these rather stable levels of subjective wellbeing and health over a longer time (Goodin et al. 1999; Headey et al. 2010).

Social trust can be conceived as a yardstick for people's social capital and social integration in society (Coleman 1988). Positive and sometimes significant results were found for social trust and notably for trust in the caseworker (institutional trust). Strongest positive effects on social trust are observed in Groningen for intensive support and in Oss and Deventer for exemption. In Groningen and Wageningen positive effects on social trust were also observed with the earnings release group. In most of the four cities for which information was available, positive effects were observed for trust in the caseworker. In two cities Tilburg (intensive support with earnings release) and Oss (exemption and extra support) they were also significant. In these two cities but also in Wageningen and Apeldoorn-Epe the experiment was labeled a 'trust' experiment. The building up of a trust-relationship with the participant was in these cities considered an important feature of the experiment with expected positive spin-offs for more effective support.

Finally, with respect to the *financial situation* of the participants, the findings show very mixed evidence across the cities with positive and negative effects which are however in most cases insignificant. In Groningen, Oss and Apeldoorn-Epe positive effects on the financial situation are observed for all treatments, in Apeldoorn-Epe they were also significant. Overall, it emerges that notably intensive support or active counselling might contribute to support people in resolving financial strain caused by a poor financial situation or problematic debts.

Theory or implementation failure

As with the employment outcomes, the evidence on our 'soft' outcome measures show varying effects which are overall rather modest. Whether these modest findings on these measures point to a 'theory failure', an 'implementation failure' or both, needs further scrutiny. One issue discussed in the report is that the experiment is designed and set-up according to the rules implied in the regulations of law because of which the room for experimenting with alternative regimes by the municipalities was constrained. But, also within these local bureaucracies, the implementation of the experiment was constrained by the availability of resources and existing implementation policies and practices. That also means that the theoretically designed treatments might not be implemented according to the theory. Some clues, can be derived from the qualitative part of the research in which interviews were held with the caseworkers and project leaders and from which some lessons can be drawn. In the first few months after the start the caseworker in cities such as Tilburg, Oss, Wageningen and Apeldoorn, expressed serious concerns about how to implement the treatments as they were designed in theory (cf. list of local reports). They had no experience with the alternative treatments and shared a long-lasting expertise with the standard workfare approach. They expressed the need for training and InterVision to get to know how they should operate and deal with a view to the content and implementation of the treatments. It might be that the treatments require a new methodology and new expertise they did not have and which they had to build up during the experiment. At the same time the way of management and support of the experiment within the local context appeared to have been rather different because of which the conditions for succesful implementation are different. It might also be that the theory is too abstract or not sufficiently developed to provide clear ideas on how the 'black box' of support and coaching practices should look like (cf. Blonk, 2018). These concerns about theory and implementation might explain the modest results we found in most cities.

Qualitative research outcomes

In May 2020 the researchers published a joint report in which they summarized the findings from the local reports which were partly also based on the results of the qualitative research in the various cities[44]. From these findings they concluded that approaching welfare recipients with more trust, more autonomy and positive attention seem to result in a more relaxed and open relationship with positive spin-offs for both, caseworkers and welfare recipients (Sanders et al. 2020). Notwithstanding the modest effects, most stakeholders involved in the experiments have expressed their enthusiasm and positive evaluation of it and for what it means to participants and caseworkers[45]. The positive spin-offs are in their view associated, foremost, with a better relationship between caseworker and recipient. In several experiments, caseworkers have indicated that they go to work with more pleasure and that they are more satisfied with their work. They also value the improved relationship with the participants that is likely to increase their motivation and initiative. In their role as professionals and their positive experiences with the experiment, they plea for investments in professional support and development and in more room for experimenting with alternative ways of treatment.

[44] In the local reports of Tilburg, Wageningen, Oss and Apeldoorn-Epe, the findings of the qualitative research part were discussed. This part of the research was called the process evaluation and based on focus group interviews with the caseworkers and project leaders and three surveys held among caseworkers for each participant about the support process, at the start and after one and two years.

[45] Supportive evidence for this assertion is that in some cities (Tilburg, Oss) we found significant effects on the outcome measure "trust in the caseworker".

Final remarks

Experimenting with different policy regimes in RCT-field experiments is a challenging and very valuable way to test the effects of alternative policies but is also demanding because some strict methodological requirements need to be fulfilled, implying that the researchers need to have a large say in the design of it. This report focuses on summarizing the first results of the experiments with a view to the employment and health and wellbeing outcomes. A number of interesting substantive issues need further scrutiny such as research into the heterogeneity effects (interaction effects with health, skill-level and spell duration), research into selection and experiment effects and how to correct for these and research into the longer-term effects on sustainable employment and on health and wellbeing. But, also research related to the consequences of technical change for the employment and financial situation of people on social assistance are mentioned in the introduction but are not yet discussed in depth or examined in this report. The latter topic will be the subject of our work in the next stage of the project.

References

Abbring, J. , G. J. Van den Berg & J. Van Ours (2005). The Effect of Unemployment Insurance Sanctions on the Transition Rate from Unemployment to Employment, *The Economic Journal*, Vol. 115 (505), 602–630.

Adamson, J., S. Cockayne, S. Puffer, D. J. Torgerson (2006). Review of randomized trials using the post-randomized consent (Zelen's) design, *Contemporary Clinical Trials*, Volume 27,4:305-319.

Angrist J., and J. Pischke (2009). Mostly Harmless Econometrics: An Empiricists's Companion, Princeton University Press, Princeton.

Athey, S and G. Imbens, 2017, The econometrics of randomized experiments, *Handbook of Economic Field Experiments*, Vol. 1, 73-140.

Atkinson, A. B. (1996) The case for a participation income, *Political Quarterly*, 67, 1, 67–70.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives, 29*(3), 3-30.

Broersma, L., A.J.E. Edzes & J. van Dijk (2013). Have Dutch municipalities become more efficient in managing the costs of social assistance dependency. *Journal of Regional Science*, 53, 2, 274-291.

Bolhaar, J., N. Ketel and D. van Vuuren (2019). Job search periods for welfare applicants: Evidence from a randomized experiment, *American Economic Journal: Applied Economics*, Vol. 11(1): 92-215.

Betkó, J., N. Spierings, M. Gesthuizen & P. Scheepers (2019), The Who and the Why? Selection Bias in an Unconditional Basic Income Inspired Social Assistance Experiment, In: Delsen, L. (Ed.), *Empirical Research on an Unconditional Basic Income in Europe*, Springer Publisher, Series: Contributions to Economics.

Blonk, Roland (2018). We zijn nog maar net begonnen, Openbare rede prof. dr. R. Blonk, Tilburg University (6 April 2018).

Bohnet, I., Frey, B. S. and Huck, S. (2001). More order with less law: on contract enforcement, trust and crowding, *The American Political Science Review*, 95, 1,131–44.

Bond, G. R., Drake, R. E., & Luciano, A. (2015). Employment and educational outcomes in early intervention programs for early psychosis: A systematic review, *Epidemiology and Psychiatric Sciences,* Vol. 24 (5): 446–457.

Card, D., Kluve J., and A. Weber (2010). Active labour market policy evaluations: a meta-analysis, *The Economic Journal,* 120 (November), F452–F477.

Card, D., Kluve, J., & Weber, A. (2015). What works? A meta-analysis of recent active labor market program evaluations (No. w21431). National Bureau of Economic Research.

CBS (2017), De Geluk meter, CBS Statline, Voorburg/Heerlen.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology,* 94, S95–S120.

De Boer, H-W, Bolhaar, J., Jongen, E., & A. Zulkarnain (2020). Evaluatie experimenten Participatiewet: Effecten op de uitstroom naar werk. CPB: Den Haag.

Deaton, A. & N. Cartwright (2018), Understanding and misunderstanding randomized controlled trials, *Social Science & Medicine,* Vol. 210:2-21.

Deci, E. L. and Ryan, R. M. (1985). Intrinsic Motivation and Self-Determination in Human Behavior, New York: Plenum.

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: Dimensions underlying positive and negative affect, *Journal of Personality and Social Psychology,* 48, 1253-1265.

Duflo, E., R. Glennerster and M. Kremer (2008). Using randomization in development economics research: A toolkit, in: Schultz, T. and J. Strauss (eds), *Handbook of Development Economics,* Vol. 4, Ch. 61, North Holland, Amsterdam.

Ellis, R. A., & Taylor, M. S. (1983). Role of self-esteem within the job search process, *Journal of Applied Psychology*, 68, 632–640.

Elo, A.-L., Leppänen, A., & Jahkola, A. (2003). Validity of a single-item measure of stress symptoms. *Scandinavian Journal of Work, Environment & Health*, 29 (6), 444–451.

Fehr, E. and Schmidt, K. M. (2003). Theories of fairness and reciprocity: evidence and economic applications, Advances in Economics and Econometrics, Econometric Society, Eighth World Congress, 1, 208–57.

Fishbein, M., & Ajzèn, I. (2010). Predicting and changing behavior: The reasoned action approach. New York: Psychology Press.

Fouarge, D. J. A. G. (2017), Veranderingen in werk en vaardigheden [Changes in Work and Skills], (inaugural address), Maastricht University.

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory, *Journal of Economic Surveys*, 15, 5, 589-611.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change,* Vol. 140:254-280.

Gerber, Alan S., and Donald P. Green, (2012). Field Experiments: Design, Analysis, and Interpretation.
New York: W.W. Norton.

Gibson, Maria, Wemdy Hearty & Peter Craig (2020). The public health effects of interventions similar to basic income: a scoping review, *Lancet Public Health*; Vol. 5: 165–76.

Goodin, R. E., Headey, B., Muffels, R. and Dirven, H.-J. (1999). The Real Worlds of Welfare Capitalism, Cambridge: Cambridge University Press.

Glaeser, E., David L., Jose S., & Christine S. (2000). Measuring trust, *Quarterly Journal of Economics*, 115(3), 811–846.

Groot, L., Muffels, R. J., & Verlaat, T. (2019). Welfare states' social investment strategies and the emergence of Dutch experiments on a minimum income guarantee, *Social Policy & Society,* 18:2, 277–287, Cambridge University Press.

Headey, B., Muffels, R.J.A., & Wagner, G. (2010). Long-running German panel survey shows that personal and economic choices, not just genes, matter for happiness. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 107(42), 17922-17926

Helliwell, J. F., & Huang, H. (2014). New measures of the costs of unemployment: Evidence from the subjective well-being of 3.3 million Americans, *Economic Inquiry*, 52(4), 1485-1502.

Hemerijck, A. (ed.) (2017). The Uses of Social Investment, Oxford: Oxford University Press.

Hullegie, P, and J. van Ours (2014). Seek and ye shall find: How search requirements affect job finding rates of older workers, *De Economist*, Vol. 162, 377-395.

Klink van der, J.J, Bultmann, U,. Burdorf, A., Schaufeli, W.B., Zijlstra, F.R., Abma, F.I, Brouwer, S,. van der Wilt, G.J. (2016). Sustainable employability--definition, conceptualization, and implications: A perspective based on the capability approach, *Scandinavian Journal of Work Environ Health*, Vol. 42:71-9.

Knack, S. and P. Keefer (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation, *Quarterly Journal of Economics*, Vol. 112, No. 4: 1251-1288.

Knoef, M. and J. van Ours (2016). How to stimulate single mothers on welfare to find a job: evidence from a policy experiment, *Journal of Population Economics*, Vol. 29(4): 1025-1061.

Koning, P. (2009). The effectiveness of public employment service workers in the Netherlands, *Empirical Economics*, Vol. 37, 393-409.

Kunzmann, U., Little, T., Smith, J. (2000). Is age-related stability of subjective well-being a paradox? Cross-sectional and longitudinal evidence from the Berlin Aging Study. *Psychology and aging*, 15(3), 511.

Lyubomirsky, S., & Lepper, H. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 4*, 137–155.

Mani, A., Mullainathan, S., Shafir, E. and Zhao, J. (2013). Poverty Impedes Cognitive Function, *Science*, 341, 976–80.

Marshall, T., Goldberg, R. W., Braude, L., Dougherty, R. H., Daniels, A. S., Sushmita Shoma Ghose, E. D., et al. (2014). Supported employment: Assessing the evidence. *Psychiatric Services*, 65(1), 16–27.

Ministerie van Sociale Zaken en Werkgelegenheid (2017). Tijdelijke regeling experimenten Participatiewet. Staatsblad van het Koninkrijk der Nederlanden, 69, 1-18.

Morel, N., Palier, B. and Palme, J. (2012). Towards a Social Investment Welfare State? Ideas, Policies and Challenges, Bristol: The Policy Press.

Muffels, R. & E. Gielens, (2020). Job Search, Employment Capabilities and Well-being of People on Welfare in the Dutch 'Participation Income' Experiments, In: Delsen, L. (Ed.), *Empirical Research on an Unconditional Basic Income in Europe*, Springer Publisher, Series: Contributions to Economics, pp. 109-138.

Mullainathan, S. & Shafir, E. (2013). Scarcity – The True Cost of Not Having Enough, London: Penguin Books.

OECD (2020), "What is happening to middle-skill workers?", in *OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis*, OECD Publishing, Paris, https://doi.org/10.1787/c9d28c24-en

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudemonic well-being, *Annual Review of Psychology*, 52, 141–166.

Saks, A. M. , Zikic, J., Koen, J. (2015). Job search self-efficacy: Reconceptualizing the construct and its measurement, *Journal of Vocational Behavior*, 86:104-114, ISSN 0001-8791.

Sanders, M. W. J. L., Betkó J. G., Edzes, A., Gramberg, P. J., Groot, L. F. M., Muffels, R. J. A., Rosenkranz, S., Rijnks, R. H., Spierings, C. H. B. M., Venhorst, V. A., & Verlaat, T. L. L. (2020). De bijstand kan beter. *Me Judice*, *2020*(28 mei 2020).

SCP (2019). Eindevaluatie van de Participatiewet, SCP-publicatie 2019-17, Sociaal en Cultureel Planbureau, Den Haag.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction, *American Psychologist*, 55, 5–14.

Sen, A.K. (1999). Development as freedom, New York: Knopf.

Sen, A.K. (2004). Capabilities, lists, and public reason: continuing the conversation, *Feminist Economics*, 10 (3), 77–80.

Thaler, R. H. and Sunstein, C. R. (2008). Nudge: Improving Decisions about Health,Wealth, and Happiness, New Haven, Conn.: Yale University Press.

Uslaner, E. M. (2002). The moral foundations of trust, Cambridge University Press.

van der Klaauw, B., & van Ours, J. C. (2013). Carrot and Stick: How Re-Employment Bonuses and Benefit Sanctions Affect Exit Rates from Welfare, *Journal of Applied Econometrics*, 28 (2), 275–296.

van Parijs, P. (2004). Basic income, a simple and powerful idea for the twenty-first century, P*olitics and Society*, Vol. 32, 1, 7–39.

Veenhoven, R. (1984). *Conditions of happiness*. Dordrecht: Reidel (now Springer).

Verlaat, T., S. Rosenkranz, L. Groot, M. Sanders (2020). Requirement vs. Autonomy: What works in social assistance? (unpublished draft), University of Utrecht.

Ware, J.E. Jr & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care,* 30(6):473-83.

Widerquist, K. (2018). The Devil's in the Caveats: A Critical Analysis of Basic Income Experiments for Researchers, Policymakers, and Citizens, Palgrave Macmillan.

List of local research reports

Betkó, J., N. Spierings, Gesthuizen, M., P. Scheepers (2020). Rapportage experiment Participatiewet gemeente Nijmegen, Nijmegen.

Edzes, A., Rijnks, R., Kloosterman, K., & Venhorst, V. (2020). Bijstand op Maat: Beleidsrapport. (URSI-onderzoeksrapport;No.366)), Groningen: Rijksuniversiteit Groningen.

Gramberg, P. & J. de Swart (2020). Wat werkt op weg naar werk? Eindrapport Experiment Participatiewet gemeente Deventer. Enschede: Saxion Hogeschool.

Muffels, R., K. Blom-Stam & S. van Wanrooij (2020a). Vertrouwensexperiment Tilburg: Werkt het en waarom wel of niet?, Maart 2020, Tilburg University/Tranzo-ReflecT, p.1-117.

Muffels, R., K. Blom-Stam & S. van Wanrooij (2020b). Vertrouwensexperiment Wageningen: Werkt het en waarom wel of niet?, Voorlopig eindverslag, Tilburg University/Tranzo-ReflecT, p.1-113.

Muffels, R., K. Blom-Stam & S. van Wanrooij (2020c). Vertrouwensexperiment Oss: Werkt het en waarom wel of niet?, February 2020, Tilburg University/Tranzo-ReflecT, p.1-110.

Muffels, R., K. Blom-Stam & S. van Wanrooij (2020d). Vertrouwensexperiment Apeldoorn-Epe: Werkt het en waarom wel of niet?, April 2020, Tilburg University/Tranzo-ReflecT, p.1-105.

Verlaat, T., de Kruijk, M., Rosenkranz, S., Groot, L., & Sanders, M. (2020). Onderzoek Weten wat werkt: samen werken aan een betere bijstand, Eindrapport. Utrecht: Universiteit Utrecht.

Annex 1. Design and methodology of each experiment

Table A1: Methodology: randomization, implementation, sample and other issues by city

Official experiments:
Deventer
• *Randomization:* Took place before registration but after selection of a randomly drawn reference group. Due to randomization, a pure comparison with the random reference group is possible. Generally, there were no big differences between experiment groups with regard to background characteristics
• *Implementation:* Included in the effect is the effect of supervision by a fixed team of dedicated caseworkers, notably for the Intensive Support group.
• *Sample:* Particular groups, such as the youngsters, were excluded from the experiment. Small differences were found between the experiment groups and the target population because participants have a slightly more favorable labor market position. Around a quarter of the target group took part in the study.
• *Other:* Due to different inflow moments, participants who entered from the start have more frequently completed a questionnaire than participants entering later.
Tilburg
• *Randomization* : Randomization took place after registration but no big differences in characteristics were found between the experiment groups. Substantial differences were however found with the target population (longer welfare durations, more females).
• *Implementation* : The control group was supervised by consultants from the Work and Income Department while the treatment groups were supervised by externally recruited coaches. The exemption intervention (with work bonus) was 'learning of self-management' of the participants by trained coaches and therefore not ‘leaving alone’. The supervision was therefore not less intensive than in the two Intensive Support groups. The control group, on the other hand, might have been supervised somewhat more intensively than usual. It is concluded that experiment (John Henry) effects might have affected the comparison because the control group behaves differently immediately after the start with a view to the higher outflow to work than usual. Effects compared to the control group can therefore not be attributed only to the interventions.
• *Sample:* More than 35% of the SA beneficiaries were excluded from participation in the experiment. Youngsters below 27 years of age but also beneficiaries with shorter stays (<1 year in SA) who were put under the supervision of a contracted private company (Sagenn/Diamantgroup) were excluded.
• *Other* : In the Tilburg final report, local BUS data and slightly different models were used, because of which the results differ somewhat. Propensity score matching has been applied in the local report to correct for selection in the reference group (all non-participants) to permit comparison of outflow to work between the intervention groups and this group. This was required because the outflow to work figures of the control group were presumably biased.
Nijmegen
• *Randomization* : Took place after registration. Experiment groups appeared comparable with regard to background characteristics. Differences on ‘part-time work’ were however already found a few months before the start in the outcomes between the intervention groups and the control group, because of which the control group has a better start position. At survey baseline, that is at start, differences were also found in, among other things, (mental) health, addiction, and personal problems, because of which the control gro up has from start on a better labor market position. This points to experiment (John Henry) effects.
• *Implementation* : Included in the measured effects is the effect of supervision by a fixed team of dedicated caseworkers, notably for the Intensive Support group.
• *Sample* : Not everyone was permitted to participate in the experiment and randomization took place after registration. As a result, the group of participants has a more favorable labor market position than the target group. Research shows that there are differences between those who registered and those who did not, where those who signed up had better employment chances (John Henry effects).
Utrecht
• *Randomization* : Took place after registration. Generally, no big differences were found between the intervention groups with regard to background characteristics. A few months before the start, small differences in ‘part-time work’ work outcomes were found between the intervention groups and the control group, where the control group has a less favorable labor market position.

• *Implementation*: Included in the treatment effect is the effect of supervision by a fixed team of dedicated caseworkers, notably for the Intensive Support group.
• *Sample*: Some groups were excluded from participation in advance such as people with short-term stays in welfare (<6 months) while randomization took place after registration. As a result, the group of participants appears to have a more favorable labor market position than the target population.
• *Other*: Results from the surveys for Intensive Support should be interpreted with caution because of a slightly lower response rate.
Groningen
• *Randomization*: Took place before registration but after selection of a randomly drawn reference group. Due to randomization, a pure comparison with the random reference group is possible. Generally, there were no big differences found between experimental groups with regard to background characteristics.
• *Implementation*: Likely there are behavioral effects in the control group (John Henry effects) because of which a pure comparison with this group is not viable and comparison with the randomized reference group is preferred. For the survey outcomes, a comparison with the control group is appropriate because the random reference group did not participate and therefore did not fill in the survey questionnaires.
• *Sample*: Randomization of the entire target group results in a high degree of representativity for the target population. Beneficiaries with short-term stays (<½ years) were excluded from the target population.
Wageningen
• *Randomization*: Took place after registration. Experiment groups appeared comparable with regard to background characteristics. In a few months before the start, small differences in employment outcomes were found between the Intensive Support group and the control group, in which the control group is more likely to start working in small jobs and in jobs with earnings exceeding 50% of the minimum wage.
• *Implementation*: At the start, a new working method was developed for all interventions including the control group according to which supervision in this group became more intensive than usual. Therefore, potential experiment effects in the control group (John Henry) are found, partly due to the new working method, because of which the outflow to work in the control group might be overestimated.
• *Sample:* Not everyone was permitted to participate in the experiment and randomization took place after registration. The participants signed a contract with the municipality on their rights and duties during the experiment. Only 3% (11 persons) withdrew from the experiment. About half of the social assistance population participated in the survey.
• *Other*: In the local report for Wageningen, the local BUS data were used and slightly different models than in the joint report based on the CBS data, so the results differ slightly. Propensity score matching has been used in the local report to correct for selection in the reference group, consisting of all non-participants, to permit comparison of outflow to work between the intervention groups and this group.
Apeldoorn-Epe and Oss
• *Randomization*: Took place after registration. Experiment groups appeared comparable both in Oss and Apeldoorn-Epe with regard to background characteristics. Differences were however found between the experiment groups and the target population.
• *Implementation*: The caseworkers of the control group were part of the team of caseworkers of the entire experiment in both cities. In Oss they worked in tandem so that each caseworker could take over from the other in the case of sickness or absence. The caseworkers in both cities got the training of the training Centre (the so-called 'skills method') for supporting the participants on self-management. The local report mentions the possibility that the caseworkers of the 'care-as-usual' group behaved differently from the caseworkers not being part of the experiment, pointing to an experiment effect. The participants might also have behaved differently because of participating in the experiment. For that reason, potential experiment effects (John Henry) were reported in both cities.
• *Sample:* Not everyone was permitted to participate in the experiment. In Oss 13% of the potential participants withdrew from the experiment but 4% already at the start. In Apeldoorn-Epe about 20% of the registered participants withdrew from the experiment after the start and for various reasons. Differences in characteristics were found between the experiment groups and the reference group of all non-participants.
• *Other*: In the local report for Oss and Apeldoorn-Epe, the local BUS data were used and slightly different models than in this joint report based on the CBS data, so the results might differ slightly. Propensity score matching has been used in the local report to correct for selection in the reference group of non-participants (due to selection into the experiment) and to permit comparison of the work outcomes with this group.

Source: Derived from Sanders et al. 2020 with own additions.

Annex 2. Operationalisation of survey outcome measures

Table A.2.1. Definition of survey outcome measures

Subjective health (SH)	• Subjective health question. Likert scale: 1=very bad to 5=excellent. Scores are normalized on a scale from 0-10.
Mental health (MH)	• Mental health scale based on 5 items and 5-points Likert scale: 1=never to 5=always. Scores are normalized on a scale from 0-10. • Items: nervous; sad-down; gloomy-depressed; calm-peaceful; happy
Health (SMH)	• The health measure used is the average of the 0-10 scores on mental health and subjective health.
Capabilities (CAP)	• Based on two questions and 7 items: item is considered important; available in own situation (Likert scale 1=never to 5=always). • Items: to learn and do new things; to set own targets; to have good contacts with others; to have a decent income; to contribute to the life of others. • Capability index: weighted sum of items, weighted with level of importance ranging from 1 to 5 and normalised on 0-10 scale
Social trust (STR)	• ESS survey question on how much trust people put in others on 0-10 scale
Institutional trust (ITR) and trust in the caseworker (TRCW)	• Survey question, partly derived from ESS, on how much trust people put in the government, municipality, SA department, caseworker, on a four-points Likert scale, 1=not at all to 4=full trust and normalised into 0-10 scale. The indicator "trust in the caseworker" uses only the last item.
Job search intensity (JSI)	• Based on the weekly number of hours spent on job search in the last four weeks, the number of applications in the last 4 weeks and the number of invitations for an interview in the last 4 weeks. The scores are normalized on a scale from 0 to 10.
Self-efficacy (SE)	• Based on 4 statements and 5-points Likert scale 1=completely disagree to 5 completely agree • Items: find work when i put effort; confident to find work in future; can make good impression when apply; job fits well to my education/skills
Social network (SPART)	• Frequency of monthly contacts with family, friends, neighbours ranging from 0 to 4 times a month
Financial situation (FINSIT)	• Scale ranging from: 1. have to make debts, 2. dissave, 3. just make ends meet, 4. saving a bit of money, 5. can save money. • Financial stress (INCPOV): poor income when people have debts or dissave (score 4 or 5)
Deprivation poverty (DEPPOV)	• Based on two questions on necessity of item and whether one can afford it on 5 out of 14 selected items. • Item list derived from EU-SILC deprivation questions • 5 items: once a day eating fruit/vegetables; once a day a meal with meat, poultry or fish; replace worn-out furniture; replace worn clothes; repair damaged equipment

Annex 3. ITT and LATE-regression results on employment outcomes

Table A3.1. Estimation of ITT-effects for more than 8 hours, more than 27 hours a week and for more than 70% of minimum wage

	Parttime+Fulltime		**Fulltime**		**Fulltime income**	
	>8 hours		**>27 hours**		**>70% Min. Wage**	
	COEFF	**SE**	**COEFF**	**SE**	**COEFF**	**SE**
Groningen (N=8190)						
Mean reference group	0.151	0.359	0.055	0.228	0.082	0.276
Exemption	-0.031**	0.026	-0.004	0.020	-0.015**	0.023
Intensive support	-0.019**	0.026	-0.001	0.020	-0.010	0.023
Earnings release	-0.007	0.026	0.009	0.020	-0.005	0.023
Utrecht(N=752)						
Mean control group	0.122	0.329	0.059	0.024	0.069	0.254
Exemption	0.024	0.034	0.013	0.026	0.039	0.030
Intensive support	**0.063****	0.034	0.040	0.027	0.040	0.029
Earnings release	**0.076****	0.036	0.020	0.025	0.024	0.028
Wageningen (N=408)						
Mean control group	0.304	0.463	0.104	0.306	0.163	0.371
Exemption	-0.006	0.063	0.044	0.049	-0.008	0.053
Intensive support	0.004	0.064	0.006	0.048	0.028	0.053
Earnings release	-0.012	0.063	-0.013	0.044	-0.042	0.050
Deventer(N=695)						
Mean control group	0.209	0.408	0.092	0.290	0.141	0.341
Exemption + ER	-0.027	0.037	0.010	0.031	-0.033	0.035
Intensive support + ER	0.003	0.040	0.045	0.330	0.019	0.038
Intensive support + ER/App	0.025	0.039	0.021	0.032	-0.009	0.036
Nijmegen (N=289)						
Mean control group	0.443	0.499	0.165	0.373	0.237	0.428
Exemption + ER	**-0.186****	0.062	**-0.081***	0.046	**-0.129****	0.055
Intensive support + ER	**-0.144****	0.067	-0.053	0.051	**-0.106***	0.057
Tilburg (N=683)						
Mean control group	0.203	0.403	0.084	0.304	0.129	0.336
Exemption + ER	0.026	0.034	0.026	0.029	0.002	0.035
Intensive support	**-0.017****	0.036	-0.001	0.026	-0.018	0.029
Intensive support + ER	-0.003	0.036	-0.001	0.026	0.000	0.032
Oss (N=340)						
Mean control group	0.186	0.391	0.186	0.391	0.195	0.391
Exemption	0.091~	0.055	0.072	0.051	0.063	0.060
Intensive support	0.009	0.044	-0.009	0.044	0.001	0.058
Apeldoorn-Epe (N=550)						
Mean control group	0.120	0.326	0.230	0.422	0.142	0.350
Exemption	0.034	0.044	0.010	0.036	0.030	0.039
Intensive support	**0.101****	0.047	0.042	0.035	0.059	0.040

Note: ***p<0.01;**p<0.05;*p<0.10;~p<0.15. ITT-effects at 16-24 months after the start of the model in Eq. (1). Updated CBS-microdata for Tilburg and Wageningen (Jan. 2015-June2020). Due to experiment effects in the control group the ITT-effects for Groningen, Tilburg and Nijmegen cannot be attributed causally to the treatments.
Source: LOEP calculations based on CBS microdata.

Table A3.2. Estimation of LATE-effects for more than 8 hours, more than 27 hours a week and more than 70% of minimum wage

Outcome measure	>8 hours		>27 hours		>70% min. wage	
Cities	**COEFF**	SE	**COEFF**	SE	**COEFF**	SE
Groningen (N=8190)						
Mean control group						
Exemption						
Intensive support						
Earnings release						
Utrecht (N=752)						
Mean control group	0.122	0.329	0.059	0.235	0.069	0.254
Exemption	0.025	0.035	0.014	0.027	0.041	0.03
Intensive support	**0.068***	0.036	~0.042	0.029	0.043	0.03
Earnings release	**0.083****	0.038	0.021	0.027	0.027	0.03
Wageningen (N=375)						
Mean control group	0.299	0.46	0.069	0.255	0.161	0.37
Exemption	0.002	0.079	0.062	0.054	0.003	0.067
Intensive support	-0.042	0.075	0.071	0.053	0.006	0.064
Earnings release	-0.054	0.069	0	0.045	-0.076	0.056
Deventer (N=695)						
Mean control group	0.209	0.408	0.092	0.29	0.141	0.349
Exemption	-0.083	0.115	-0.177	0.117	-0.103	0.109
Intensive support	0.014	0.178	-0.019	0.177	0.09	0.167
Int. Support w./ app	0.129	0.203	-0.05	0.204	-0.045	0.184
Nijmegen (N=289)						
Mean control group	0-.443	0.499	0.165	0.373	0.237	0.428
Exemption + ER	**-0.206*****	0.066	**-0.09****	0.057	**-0.14****	0.058
Intensive support + ER	**-0.197****	0.079	**-0.079***	0.067	**-0.144****	0.067
Tilburg (N=468)						
Mean control group	0.262	0.442	0.107	0.31	0.164	0.372
Exemption + ER	-0.02	0.067	-0.015	0.047	-0.04	0.054
Intensive support	-0.086	0.072	-0.031	0.052	-0.068	0.059
Intensive support + ER	-0.024	0.069	-0.029	0.048	-0.013	0.058
Oss (N=237)						
Mean control group	0.302	0.462	0.163	0.371	0.163	0.371
Exemption	0.038	0.077	0.093	0.065	0.093	0.065
Intensive support	-0.021	0.069	**-0.089***	0.045	~-0.075	0.047
Apeldoorn (N=402)						
Mean control group	0.216	0.413	0.144	0.352	0.151	0.359
Exemption	0.029	0.051	-0.01	0.043	0.019	0.046
Intensive support	**0.176*****	0.056	0.031	0.045	~0.076	0.049

Note: ***p<0.01;**p<0.05;*p<0.10;~p<0.15. LATE-effects at 16-24 months after the start of the model in Eq. (3). Due to experiment effects in the control group the LATE-effects for Groningen, Tilburg and Nijmegen cannot be attributed causally to the treatments. For Groningen the LATE-effects are not given because the reference group and not the control group has been used as the comparison group.
Source: LOEP calculations based on CBS microdata.

Annex 4. ITT-regression results on survey outcomes

Table A.4.1: Estimation results of ITT-effects on various outcome measures for three cities with single treatments, Groningen, Utrecht and Wageningen.

Single treatments (survey 2 results)													
Cities	Groningen (N=757)					Utrecht (N=588)				Wageningen (N=219)			
Treatments	CG	T1-A	T2-B	T3-C	T9-B+E	CG	T1-A	T2-B	T3-C	CG	T1-A	T2-B	T3-C
	CG MN	EX	IS	ER	CS	CG MN	EX	IS	ER	CG MN	EX	IS	ER
Subjective health	5.141	0.083	-0.078	0.140	0.150	4.360	0.032	0.060	-0.244~	5.059	-0.418	0.223	-0.163
SE	1.561	0.137	0.157	0.146	0.152	1.986	0.180	0.180	0.163	1.689	0.337	0.374	0.327
p-value		0.545	0.617	0.338	0.324		0.849	0.739	0.136		0.214	0.551	0.615
Subjective well-being	6.543	**-0.238***	0.119	0.079	0.073	6.120	0.067	0.097	-0.008	6.814	-0.244	-0.313	-0.491~
SE	1.531	0.127	0.166	0.134	0.141	2.083	0.205	0.217	0.207	1.609	0.335	0.375	0.323
p-value		0.062	0.474	0.554	0.605		0.745	0.645	0.969		0.467	0.404	0.131
Institutional trust	5.681	-0.054	-0.028	0.077	0.033	5.518	0.141	0.044	-0.020	6.706	0.072	0.115	-0.374
SE	2.451	0.179	0.217	0.189	0.184	2.562	0.247	0.249	0.250	1.701	0.377	0.358	0.328
p-value		0.763	0.898	0.683	0.858		0.570	0.859	0.937		0.849	0.748	0.255
Social trust	5.445	0.294~	**0.65*****	**0.429***	0.28~	5.370	0.087	0.160	0.016	6.700	-0.233	0.06	-0.213
SE	2.261	0.186	0.209	0.205	0.191	2.475	0.261	0.262	0.262	1.877	0.385	0.429	0.349
p-value		0.115	0.002	0.037	0.144		0.738	0.540	0.950		0.546	0.889	0.543
Trust in caseworker										3.020	0.091	0.123	0.077
SE										0.735	0.154	0.164	0.135
p-value											0.557	0.454	0.569
Self-efficacy	4.956	0.123	0.253	0.271~	0.180	4.963	**0.489***	**0.573***	0.328	5.039	-0.435	-0.17	0.046
SE	2.186	0.164	0.217	0.174	0.152	2.311	0.245	0.298	0.270	2.584	0.528	0.521	0.473
p-value		0.402	0.245	0.119	0.237		0.047	0.055	0.225		0.411	0.745	0.473
Job search intensity	0.236	**-0.09***	0.000	-0.014	-0.005	0.210	0.017	0.007	-0.033	0.189	-0.059	0.161	-0.118
SE	0.631	0.050	0.086	0.060	0.054	0.541	0.580	0.054	0.057	0.590	0.103	0.212	0.094
p-value		0.071	0.996	0.817	0.228		0.776	0.894	0.559		0.564	0.449	0.24
Financial situation	4.782	-0.152	0.474~	0.301	0.082	4.459	0.002	0.213	0.083	5.284	-0.673	-0.333	-0.928~
SE	2.370	0.255	0.290	0.261	0.266	2.322	0.270	0.287	0.273	2.598	0.572	0.667	0.607
p-value		0.551	0.102	0.249	0.759		0.994	0.459	0.760		0.242	0.618	0.125

Note: *p<0.1; **p<0.05; ***p<0.001; ~p<0.15. The outcome measure 'trust in caseworker' is not available for Groningen. OLS-estimates of ITT-effects on various outcome measures using second wave data and based on the empirical model in Eq. (2) for survey data,; control group mean, coefficients, standard errors and p-values. Significant estimates in bold. CG=control group; EX=Exemption; IS=intensive support; ER=earnings release.

Source: LOEP calculations based on CBS microdata.

Table A.4.2: Estimation results of ITT-effects on various outcome measures for three cities with combination treatments, Deventer, Tilburg and Nijmegen.

Combination treatments (survey 2 results)											
Cities	Deventer (N=233)				Tilburg (N=400)				Nijmegen (N=238)		
Treatments	CG	T4-A+C	T5-B+C	T7=B+C	CG	T4-A+C	T2-B	T5-B+C	CG	T4-A+C	T5-B+C
	CG MN	EX w/ER	IS w/ ER	IS w/ Ap	CG MN	EX w/EI	IS	IS w/ EF	CG MN	EX w/EI	IS w/ ER
Subjective health	4.708	0.169	-0.283	0.342	4.946	-0.195	-0.184	-0.307~	5.148	0.130	0.019
SE	1.939	0.251	0.307	0.308	1.857	0.213	0.210	0.191	1.743	0.200	0.236
p-value		0.503	0.358	0.267		0.360	0.382	0.108		0.516	0.936
Subjective well-being	6.293	0.122	0.040	0.049	6.331	0.149	0.301	0.212	6.723	0.002	0.153
SE	1.815	0.283	0.309	0.307	1.849	0.225	0.251	0.217	1.447	0.204	0.182
p-value		0.666	0.896	0.874		0.509	0.232	0.329		0.991	0.403
Institutional trust	6.557	0.156	-0.074	0.150	4.790	-0.178	-0.088~	-0.03~	6.402	**-0.48***	-0.172
SE	2.457	0.322	*0.376*	0.340	2.382	0.329	0.353	0.299	2.288	*0.283*	0.303
p-value		0.628	0.844	0.659		0.283	0.131	0.132		0.092	0.571
Social trust	5.319	0.439	-0.169	-0.321	4.790	0.178	0.088	0.03	6.675	-0.246	-0.136
SE	2.097	0.310	0.350	0.353	1.875	0.277	0.349	0.288	1.958	0.242	0.244
p-value		0.158	0.629	0.364		0.521	0.802	0.917		0.547	0.578
Trust in caseworker					2.753	0.323	**0.323****	**0.31****			
SE					0.741	0.124	0.146	0.122			
p-value						0.010	0.028	0.011			
Self-efficacy	5.395	-0.278	0.371	0.183	5.238	0.069	-0.113	0.442	6.149	0.346	-0.127
SE	2.340	0.327	0.426	0.347	2.442	0.332	0.329	0.316	1.836	0.249	*0.249*
p-value		0.396	0.385	0.599		0.835	0.733	0.163		0.166	0.612
Job search intensity	0.115	0.003	-0.070	0.003	0.192	-0.066	**-0.147***	-0.068	0.144	0.021	0.043
SE	0.338	0.051	0.045	0.036	0.635	0.092	0.087	0.096	0.328	0.057	0.075
p-value		0.956	0.122	0.926		0.470	0.091	0.482		0.717	0.567
Financial situation	4.815	-0.436	-0.429	-0.708~	4.815	0.085	-0.231	0.141	4.969	**-0.689***	-0.106
SE	2.270	0.413	0.537	0.452	2.150	0.349	0.362	0.325	2.936	0.412	0.452
p-value		0.252	0.425	0.120		0.808	0.524	0.605		0.096	0.814

Note: *p<0.1; **p<0.05; ***p<0.001; ~p<0.15. The outcome measure 'trust in caseworker' is not available for Deventer. OLS-estimates of ITT-effects on various outcome measures using second wave data and based on the empirical model in Eq. (2) for survey data,; control group mean, coefficients, standard errors and p-values. Significant estimates in bold. CG=control group; EX=Exemption; IS=intensive support; ER=earnings release.

Source: LOEP calculations based on CBS microdata.

Table A.4.3: Estimation results of ITT-effects on various outcome measures for two informal experiments with single treatments, Oss and Apeldoorn-Epe.

Informal experiments (survey 2 results)						
Cities	**Oss (N=189)**			**Apeldoorn-Epe (N=267)**		
Treatments	**CG**	**T1-A**	**T2-B**	**CG**	**T1-A**	**T2-B**
	CG	**EX**	**IS**	**CG**	**EX**	**IS**
Subjective health	4.858	-0.059	**-0.526***	4.690	0.254	0.122
SE	1.977	0.289	0.312	-1.592	0.224	0.215
p-value		0.838	0.094		0.257	0.572
Subjective well-being	6.651	-0.322	**-0.981*****	6.507	0.006	-0.064
SE	1.318	0.251	0.311	-1.561	0.242	0.239
p-value		0.201	0.002		0.979	0.791
Institutional trust	5.882	**0.549***	0.089	5.130	**0.598***	-0.354
SE	2.094	0.294	0.326	-2.520	0.361	0.357
p-value		0.063	0.785		0.099	0.322
Social trust	5.033	0.308	-0.263	4.428	0.147	-0.143
SE	1.877	0.309	0.339	-1.985	0.305	0.296
p-value		0.322	0.439		0.629	0.631
Trust in caseworker	2.784	**0.478*****	**0.334****	2.754	0.057	0.111
SE	0.702	0.134	0.133	-0.847	0.133	0.131
p-value		0.000	0.013		0.667	0.397
Self-efficacy	4.896	0.571	0.077	4.905	-0.178	**-0.791***
SE	2.278	0.5	0.556	-2.367	0.268	0.306
p-value		0.256	0.890		0.509	0.010
Job search intensity	0.073	0.123	**0.112***	0.238	-0.082	-0.161*
SE	0.278	0.064	0.063	-0.577	0.074	0.087
p-value		0.056	0.077		0.272	0.064
Financial situation	6.324	0.358	0.082	4.386	**0.659***	0.626~
SE	4.347	0.877	0.877	1.722	0.374	0.418
p-value		0.684	0.925		0.080	0.137

Note: *p<0.1; **p<0.05; ***p<0.001; ~p<0.15. CG=control group; EX=Exemption; IS=intensive support; ER=earnings release; OLS-estimates of ITT-effects on various outcome measures using second wave data and based on the empirical model in Eq. (2) for survey data,; control group mean, coefficients, standard errors and p-values. Significant estimates in bold.

Source: LOEP calculations based on CBS microdata.