TI 2018-035/V
Tinbergen Institute Discussion Paper

Graded Return-to-Work as a Stepping Stone to Full Work Resumption

Lieke Kools[1]
Pierre Koning[2]

[1] Leiden University, Netspar
[2] VU Amsterdam, Tinbergen Institute, The Netherlands; IZA

Graded Return-to-Work as a Stepping Stone to Full Work Resumption*

Lieke Kools † Pierre Koning ‡

April 2018

Abstract

There is increasing evidence that graded return-to-work is an effective tool for the rehabilitation of sick-listed workers. Still, little is known on the optimal timing and level of grading in return-to-work trajectories, as well as the allocation of trajectories across worker types. To fill this gap, we analyze whether the effectiveness of graded return-to-work depends on the starting moment of the trajectory and the initial level of graded work resumption. We use administrative data from a Dutch private workplace reintegration provider. In order to correct for the selection bias inherent to the evaluation of activation strategies, we exploit the discretionary room of the case managers in setting up treatment plans. In correspondence with previous literature we find that graded return-to-work reduces sick spells with eighteen weeks within the first two years after reporting sick. However, the probability of work resumption after two years remains unchanged. Work resumption can be achieved faster when graded return-to-work is started earlier or at a higher rate of initial work resumption. These findings however do not hold for individuals who have problems related to mental health.

JEL-codes: I18; C26

Keywords: Activation; Long-term sickness absence; Graded return-to-work

*The authors gratefully acknowledge the financial support from Instituut Gak for this research project. They also thank Bénédicte Rouland, Rob Euwals and Bertjan Teunissen for detailed comments and suggestions on earlier versions of the paper, as well as seminar participants that gave feedback at the EALE conference in St. Gallen in 2017 and the research seminars at CPB Netherlands Bureau for Economic Policy Analysis, the University of Antwerp and RWI in Essen.

†Universiteit Leiden and Netspar. Corresponding author. P.O. Box 9500, 2300 RA Leiden, The Netherlands. Email: l.kools@law.leidenuniv.nl.

‡Universiteit Leiden, VU University Amsterdam, Tinbergen Institute and IZA. E-mail: p.w.c.koning@law.leidenuniv.nl.

1 Introduction

In the past decades many Western countries have seen a rise in uptake of disability benefits (OECD, 2010). In an effort to curb this trend, there has been an increased focus on what disabled individuals can do at work, rather than what they cannot. For example, in England sick notes have been replaced by a statement of fitness for work in 2010 (Wainwright et al., 2011), in Sweden general practitioners are recommended to subscribe part-time sick leave rather than full time sick leave (Kausto et al., 2008) and in Denmark sick-listed employees are since 2004 required to work partially after eight weeks of sick leave unless a physician has stated this is impossible (Hernæs, 2017). In a similar vein, part-time sick leave is often used as a workplace based intervention aimed at speeding up the rehabilitation process of sick-listed employees. In these interventions usually the amount of hours worked gradually increases over time, up to the moment that full work resumption is achieved. The idea is that graded work prevents the loss of working skills and may even speed up the recovery from certain injuries. For instance, Andren and Svensson (2012) argue that particularly individuals with musculo-skeletal problems benefit from graded work activities. Likewise, *Individual Placement and Support* (IPS) interventions for sick workers with mental impairments are built upon the idea that work activities may contribute to the recovery process.[1]

Research shows almost unanimously positive effects of graded work on work rehabilitation[2], whereas interventions like vocational rehabilitation and regular paramedical care rather seem to lengthen sick spells (Markussen and Røed, 2014; Rehwald et al., 2016). This however does not mean that graded return-to-work is beneficial for all individuals (Andren and Svensson, 2012; Andren, 2014; Høgelund et al., 2012). Starting graded work trajectories too soon or for too many hours may induce stress or strain on the body, hampering the recovery process. In light of these considerations, it is important to understand

[1] Corrigan and McCracken (2005) argue that psychiatric problems can be addressed only for some workers in real-life settings, so as to identify the cause of them.

[2] See e.g. Bernacki et al. (2000); Bethge (2016); Hernæs (2017); Høgelund et al. (2010); Kausto et al. (2014); Markussen et al. (2012); Rehwald et al. (2016); Viikari-Juntura et al. (2012). The general finding that graded work increases work resumption is confirmed in peer reviewed papers on the effects of part-time sick leave, active sick leave, phased return to work, and graded return to work. Related to this literature, evidence on graded work exposure or light duties also points at positive results, see e.g. Krause et al. (1998).

what separates an effective graded return-to-work trajectory from an ineffective one.

In this paper we analyze how the specifics of the set-up of a graded return-to-work trajectory determine its effectiveness. More specifically, we analyze if work resumption rates change when the trajectory is started later or at a higher initial rate of work resumption. For this we make use of registered data from a private workplace reintegration provider, which performs case management for mostly small and medium sized firms. This provider offered reintegration services for about 12,000 long-term sick listed workers, of which 62% participated in graded work trajectories between the years 2011 and 2014. We observe detailed worker information on the timing and the degree of grading that is used, as well as information on impairment types, employer, and other individual characteristics. We enrich these data with information on the case managers that were assigned to them by the reintegration provider.

In order to correct for the selection bias inherent to the evaluation of activation strategies, we follow an instrumental variables approach for which we exploit the discretionary room of the case managers in setting up treatment plans. We use the tendency of a case manager to focus on early/intense graded work (graded work propensity) as an instrument to actually receiving such a strategy. In doing so, we follow a strand of literature applying this technique in the context of activation strategies for sick-listed employees, such as Dean et al. (2015), Markussen and Røed (2014) and Rehwald et al. (2016).[3] As case managers may learn on the job or change their preference for graded work, we allow graded work propensities to vary across years. Our key assumption is that the assignment of (new) sick-listed workers to case managers is exogenous. We argue that this assumption is plausible, as the assignment is driven by the direct availability of case managers. Moreover, all the individual information on new sick-listed workers that is available to the case managers at the moment of intake is observed in our data. This means that any selection on observables can be controlled for. Reversely, we also can test for the importance of such selection effects by estimating model specifications without individual controls.

[3]For the Dutch case, where sick-listed employees have to follow a return-to-work plan established in the beginning of the sick-spell, we prefer this approach over the use of proportional hazard models, as used by for example Høgelund et al. (2010) for the case of Denmark, which relies on the non-anticipation assumption. Other methods used in the context of graded return-to-work are propensity score matching (Bethge, 2016) and randomized control trials (Viikari-Juntura et al., 2012).

Our analysis also extends on earlier studies in this field of research by using alternative propensity measures that proxy the specifics of graded work trajectories. In line with earlier work, we will first define case managers' propensity measures as the likelihood of initiating a trajectory for sick-listed workers that haven't started one yet. With the information of workers that have effectively started a trajectory, we next construct propensity measures of case managers that proxy the timing of graded work during the sickness spell as well as the graded work percentage that is applied. This then enables us to evaluate the effects of differences in the timing and the degree of grading of the interventions on (full) work resumption for those individuals that have started a graded work program. We thus gain insight in the optimal timing of graded work and the importance of gradually increasing the degree of grading.

We also shed new light on the determinants of graded work propensities and the implications of this for the interpretation of our findings. Even though the case managers' tendencies to use graded work interventions can be considered as exogenous, we cannot be sure that they are uncorrelated with other case manager characteristics affecting the likelihood of work resumption. For instance, high graded work propensities may be a marker of high quality case managers that also show higher work resumption rates without the use of graded work interventions. If so, the effectiveness of graded work interventions will be overestimated. We therefore estimate model versions with various proxies for case manager quality as additional controls. Among others, these proxies include the current and past work resumption rates of (other) sick-listed workers that were assigned and work resumption rates of workers that are out of our sample. When controlling for these proxies, we are able to assess the extent to which graded work effects are truly driven by the allocation to trajectories, rather than other case manager activities that are correlated with graded work.

In line with earlier literature, we find overall positive effects of graded return-to-work. Graded return-to-work speeds up the recovery process. At the same time, graded work does not necessarily help rehabilitate individuals who would otherwise have not rehabilitated. We find an increase in the number of weeks worked during the first two years after sick-listing of 18 weeks due to graded return-to-work, but no significant effects on

the probability to return to work within two years. Moreover, we find that starting the graded return-to-work trajectory earlier and at a higher rate of work resumption speeds up the recovery process. Starting one week earlier raises the number of weeks worked in the first two years with two weeks. Starting a graded return-to-work trajectory at a work resumption rate which is 10 percentage point higher increases the probability to return to work within two years with 2.5 percentage point. Work resumption rates are more strongly affected by the moment that graded return-to-work is started than by the moments within the trajectory at which the level of work resumption is increased.

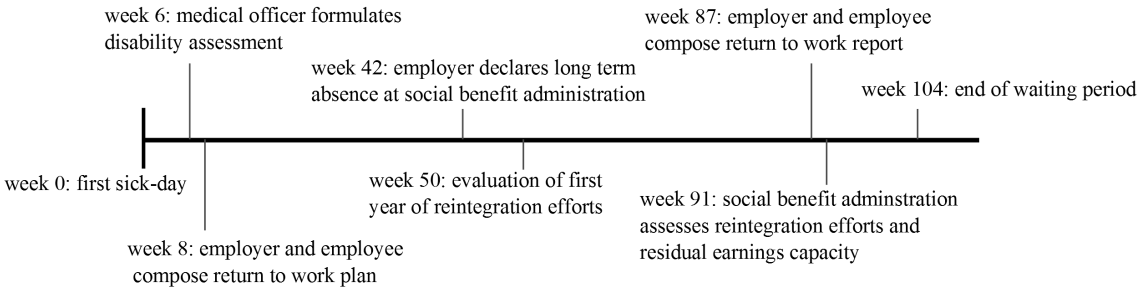
The positive effects of graded return-to-work are especially strong for individuals who have general medical conditions. For them the positive effects persist at the end of the waiting period. For individuals with problems related to mental health we find no significant effects of graded return-to-work. Moreover, and contrary to the overall findings, for these individuals starting the graded return-to-work trajectory one week earlier decreases the probability to return to work within two years with 3 percentage point.

In the following section, we explain the system of sick leave and disability insurance in the Netherlands. Then, in Section 3 we provide descriptive statistics on the sick-listed individuals in the data set, the graded return-to-work trajectories, and the case managers. In Section 4, we explain our empirical strategy and underlying assumptions. We present the results of the analysis in Section 5, followed by concluding remarks in Section 6.

2 Institutional setting

The Dutch disability system used to be notorious for its large and increasing number of beneficiaries; at its peak those receiving benefits amounted up to 12 percent of the insured individuals (Koning and Lindeboom, 2015). Since the beginning of the 21st century disability insurance award rates have been steadily declining, due to a number of reforms to the system. Among these reforms was the introduction of the Gatekeeper Protocol, obliging employers and employees to engage in activities aimed at reintegrating sick-listed workers into the workforce. As a consequence of the Gatekeeper Protocol, disability insurance inflow was estimated to reduce by about 40 percent (van Sonsbeek and Gradus,

Figure 1: Time line of the gatekeeper protocol.



2013). This positive effect can partly be attributed to improved screening, making it more difficult to use DI as an alternative exit root for Unemployment Insurance (de Jong et al., 2011). Moreover, increased employer responsibilities have played a crucial role in curbing the rise in DI beneficiaries, both as a stimulus to actively prevent sickness and as a way to accommodate activation strategies for sick-listed workers (Koning and Lindeboom, 2015).

As a result of the reforms the Netherlands has a rather unique, largely privately organized sickness and disability system (Koning, 2017). This section describes those elements of the system that are relevant for understanding the context within graded return-to-work is used.

2.1 Gatekeeper Protocol

In the Netherlands all workers are insured against income losses due to injuries, irrespective of having incurred the injury at the workplace or not. Individuals can apply for DI benefits after a two year waiting period, during which the employer is obliged to continue payments of at least 70% of the employees regular salary.[4] In practice, most Collective Labor Agreements stipulate full wage payments in the first year and 70% in the second year. During these two years, sick-listed workers can start with graded work or adapted work. As long as the waiting period proceeds and the worker has not fully recovered, wage payments are continued.

During the waiting period the employer and the employee are obliged to undertake efforts towards re-integration of the sick-listed employee. The Gatekeeper Protocol (in

[4]For comparison, in Scandinavian countries employers are responsible for two to three weeks of continued wage payments, after which the Social Insurance Administration (Sweden/Norway) or municipalities (Denmark) take over the burden (Andren, 2014; Markussen and Røed, 2014; Rehwald et al., 2016).

Dutch: *Wet verbetering Poortwachter*) gives directions as to what these efforts should entail. Figure 1 shows a time line of the concrete steps that need to be taken under the Gatekeeper Protocol. In the sixth week a disability assessment should be conducted by a medical officer (company doctor). This assessment is used as input for a reintegration plan, due in week 8. This plan is composed by the employer and employee, and should stipulate the reintegration aim[5] and the planned steps towards reaching this aim. A case manager should be appointed to keep track of the reintegration process and the return-to-work plan may be reevaluated at set dates.

After 42 weeks of sick-listing, the employer has to declare the sick-listed employee to the Social Benefit Administration (responsible for Disability Insurance) and after a year the reintegration efforts undertaken so far have to be evaluated. In the 87th week the employer and employee have to compose a return to work report, including all the reintegration efforts taken. This report will be assessed by the Social Benefit Administration in the 91th week, when also the residual earnings capacity of the individual is established. Finally, at the end of the waiting period the individual can apply for (wage-related) DI benefits granted that (1) both employer and employee can show they have taken adequate reintegration measures and (2) the individual has a residual earnings capacity of less than 65% of his/her pre-disability earnings. In case the employer has not shown sufficient collaboration, the waiting period can be extended with one more year at maximum.

2.2 Private insurance of continued wage payments and case management

Employers can insure themselves against the risk of the continued wage payments during the waiting period via private insurers. Approximately 76% of Dutch employers has such an insurance (de Jong et al., 2014). The employees of the uninsured and insured firms are similar in terms of age and gender, however insured firms are usually smaller than the uninsured firms. 78% of firms with 2-10 employees has an insurance for continued wage payments, whereas only 27% of firms with more than 100 employees has such an insurance. For small employers the risk of continued wage payments is similar to large

[5]Preferably, the reintegration aim should be (partial/adapted) employment with the current employer 'first track' reintegration). Only if this is out of reach, one can aim at fitting employment with another employer ('second track' reintegration).

firms, the relative burden however is higher. Insurers can offer the possibility to not only insure wages, but also insure all the costs that come with the obligations of the Gatekeeper Protocol. At least 67% of the insured firms have such a 'broad' insurance (de Jong et al., 2014). One such obligation is to assign a case manager that serves as a link between all the parties involved and keeps track of the progress of the sick-listed employee.[6]

During the waiting period, the sick-listed employee is allowed to work partially. The employee can either do therapeutic work wherein he or she is considered as an extra pair of hands, or do graded work. In the latter case, the employee engages in productive work and the employer pays for those productive hours worked and the insurer only pays the hours foregone. For example, if an employee engages for 20% in graded work, he gets paid 100% of his pre-sickness wage of which 80% is covered by the insurer and 20% by the employer. As the case managers are hired by the insurer, they have a direct financial incentive to actively keep track of the individuals' residual earnings capacity and to try to get the individual to participate in paid work for as much as deemed possible. With full insurance and full sick pay coverage, direct financial incentives are obviously less strong for employers and employees, but they do have an interest in work resumption anyway. For employers, sickness absence may be costly for other reasons than wage continuation, non cooperation may lead to an extension of the waiting period, and potential DI benefit costs after the waiting period are experience rated. Moreover, non-cooperation with reintegration plans inhibits the risk of getting fired or loosing eligibility to DI benefits for sick-listed employees.[7]

The data used in this paper come from a private workplace reintegration provider that is the sole provider of case management for two large insurers, together holding a market share of about 30% of the insurances for continued wage payments (Dutch Association of

[6]There are many variations possible when it comes to these insurances. There is freedom of choice in the percentage of wages insured (77% of firms chooses to insure 100% of the wages paid in the first year and 70% of the wages paid in the second year of sick leave), firms can opt-in for a deductible (77% of firms chooses to keep two weeks to two months of sick leave on their own account), and firms can choose for a stop-loss insurance (only chosen by 5% of firms of which most are firms with more than 100 employees). Of the firms surveyed by de Jong et al. (2014) 9% answered that their insurance only covers continued wage payments, 67% answered their insurance covers wage payments and the costs for gatekeeper obligations, 4% has some other type of package, and 19% does not know what their insurance covers.

[7]The evidence also confirms that private workplace reintegration providers usually increase reintegration activities in the waiting period (Everhardt and de Jong, 2011). This suggests that the provision of insurance does not (fully) remove the incentive to achieve work resumption.

Insurers, 2016). The workplace reintegration provider offers different types of products, from the registration of sickness absence to case management for individuals at risk of long term absence. In the current study, we focus on the individuals assigned to case management. Employers who take out the 'broad' insurance package with either of the two insurers are automatically directed to our workplace reintegration provider for case management. Those who are only insured against continued wage payments can opt to work with a case manager from within their own company, hire an external case manager, or hire the services of the case manager of our workplace reintegration provider.

In a typical case management trajectory a sick-listed employee can be directed to our workplace reintegration provider when a disability assessment is made by the company doctor. When there is an indication for imminent long-term absenteeism at that time and the contract with the provider includes case management, the employee gets assigned to a case manager who establishes a more detailed diagnosis and writes the return to work plan. The assignment of sick-listed employees to case managers is based on caseload, i.e. the case manager that has time takes on the sick-listed employee. Stated differently, case managers are not specialized in specific health problems, sectors, or regions.[8]

The case managers working at our workplace reintegration provider are not doctors. Usually, case managers have a background in law, HR, or (para)medical care. They purely serve as a manager of the reintegration process: consulting with the occupational physician, keeping in regular contact with the employer and sick-listed employee, identifying the steps to be taken by the employer and employee, putting together the return to work plan, and administrating the process. Based on cost-benefit analysis case managers can decide to buy rehabilitation products from external parties, such as paramedical care, job training, and coaching. They do not provide this care themselves.

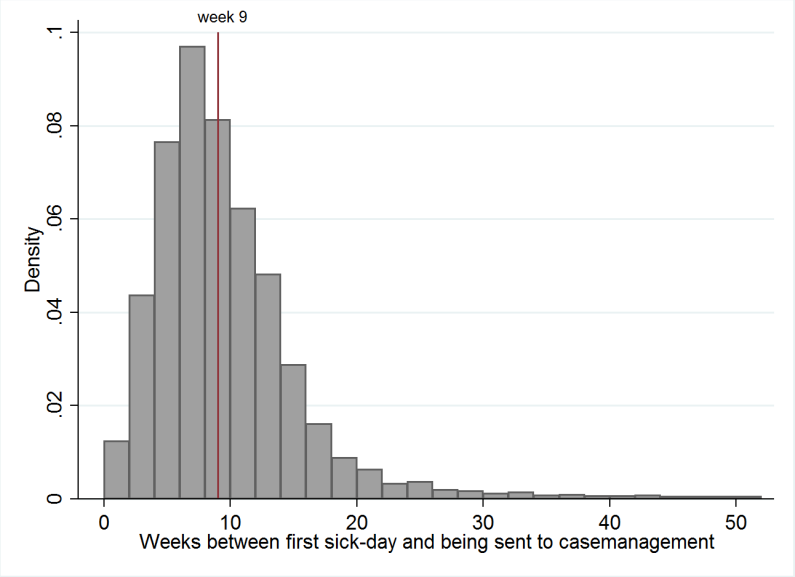
Table 1: Descriptive statistics sick-listed employees.

	all	no graded rtw	graded rtw	p-value[a]
number of sick-listed employees	11,741	4,504 (38.4%)	7,237 (61.6%)	
% female	47.3%	49.6%	45.9%	0.000
age at start of case management	42.4	41.9	42.8	0.000
weeks until start of of case management	9.2	9.3	9.1	0.207
gross pre-sickness wage (euro/day)	255.86	235.12	268.76	0.458
firm size				
- 1 employee	15.2%	17.0%	14.1%	0.000
- 2 to 9 employees	36.3%	37.5%	35.5%	0.031
- 10 to 49 employees	35.8%	32.8%	37.7%	0.000
- 50 or more employees	2.6%	1.9%	3.1%	0.000
- number of employees unknown	10.1%	10.9%	9.5%	0.020
type of condition				
- general medical - mild	7.7%	10.9%	5.7%	0.000
- general medical - medium	13.5%	11.7%	14.7%	0.000
- general medical - severe	11.5%	10.5%	12.1%	0.007
- physical - mild	7.1%	6.9%	7.3%	0.395
- physical - severe	3.6%	3.3%	3.8%	0.127
- neck, shoulder, arm complaints	6.9%	5.6%	7.7%	0.000
- hip, ankle, knee complaints	6.3%	4.7%	7.4%	0.000
- back complaints	7.3%	6.2%	8.1%	0.000
- psychiatric	1.8%	1.9%	1.7%	0.442
- psychological - mild	11.4%	10.4%	12.0%	0.007
- psychological - severe	2.8%	2.6%	2.9%	0.303
- psychosocial - mild	10.7%	10.1%	11.0%	0.106
- psychosocial - severe	1.8%	1.4%	2.0%	0.004
- social problems	2.1%	2.1%	2.0%	0.751
- conflict	4.0%	8.6%	1.1%	0.000
- other[b]	1.5%	3.2%	0.4%	0.000
time allocated to claimant (min/week)	17.0	23.1	13.2	0.000
weeks until closing of file	42.1	36.0	45.9	0.000
returns to work within one year	59.6%	53.6%	63.3%	0.000
returns to work within two years	76.7%	59.3%	87.6%	0.000

[a] Two-sided t-test on difference between sample with graded work and no graded work, with unequal variances.

[b] Other contains conditions such as flu and complaints due to pregnancy.

Figure 2: Histogram of application moments.



3 Data

3.1 Characteristics of sick-listed employees

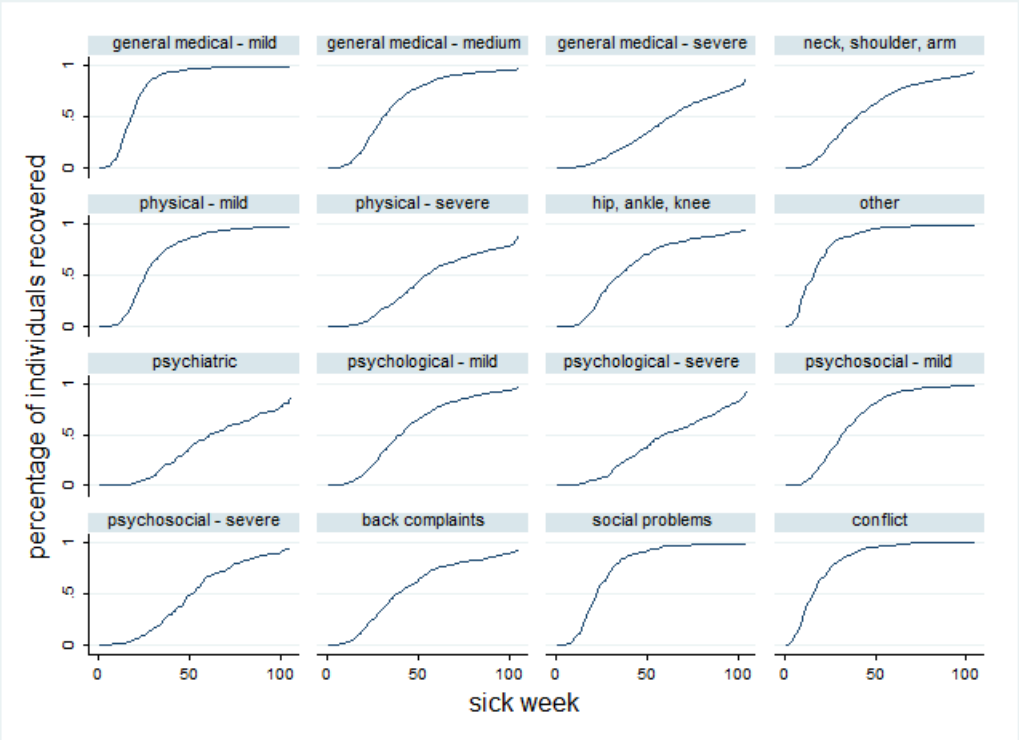
We have access to all files on sick-listed employees that were assigned to case management at our private workplace reintegration provider between the years 2011 and 2014. We exclude those individuals that hold specific insurance contracts, which include extra services before case management and/or earlier entry into case management (when there is not yet a risk of long term sickness). These contracts are predominantly held by self-employed.[9] The client files include characteristics like gender, gross (pre-sickness) wage earnings, and age. Moreover, they include the exact dates of the first sick day, of the entry day at the workplace reintegration provider, and of (partial) recovery. These files are merged to a file containing the interventions applied to each sick-listed employee and a file containing information on the assigned case manager. The data covers 11,741 sick-listed employees that are assigned to 68 case managers.

Table 1 shows the characteristics of the sick-listed employees in our sample. Almost

[8]The workplace reintegration provider has only one office, located in the centre of the country. Contact with the sick-listed employee is mostly maintained via phone and email.

[9]Table 10 of Appendix A shows the selection of our data in more detail. As becomes apparent from the table, we also exclude observations that were assigned to caseworkers with less than 20 clients in a particular year.

Figure 3: Recovry patterns by type of diagnosis



half of the individuals is female and they are on average 42 years old. The time between sick-listing and the sick leave file arriving at the provider is on average nine weeks, whereas individuals are legally obliged to start their return to work activities by the eight week. Figure 2 shows that roughly half of the individuals do enter case management before the eighth week of sickness absence. However, it also shows that there is quite some spread in the moment at which the individuals start case management. As the elapsed duration until intake is likely to affect both the likelihood of graded work and work resumption, we should take this into account in our empirical analysis. We have no information on possible reintegration efforts by the employer and employee between the moment of sick-listing and the moment of entry at the workplace reintegration provider.

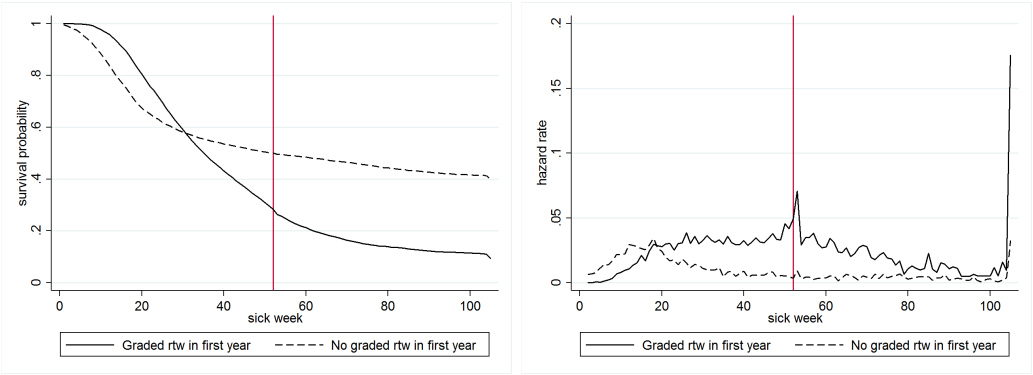
Individuals earn on average €255.86 a day and mostly work in small to average sized firms. 32.7% of the individuals has a general medical condition, 10.7% has physical problems, 20.5% has musculoskeletal problems, 30.6% has psychiatric, psychological, or social problems, 4.0% has a conflict at work, and 1.5% has some kind of other condition, such as flue or complaints due to pregnancy. When it comes to general medical conditions one

must think of individuals who are recovering from surgery or suffer from chronic illness. The average individual has 17 minutes per week allocated to him by the case manager. Individuals exit the trajectory on average after 42.1 weeks, with 59.6% of individuals returning to work within a year, and 76.7% of individuals returning to work within two years. Figure 3 shows the percentage of individuals that recovered in each sick week, stratified with respect to type of diagnosis. It should be stressed here that we only consider the individuals that were directed to the workplace reintegration provider after some period of sickness. As a result of this type of selection, recovery and work resumption rates remain close to zero in these first weeks. In line with expectations, we observe that individuals with less severe problems on average recover faster than those with more severe problems. The different type of musculo-skeletal problems (neck/shoulder/arm, hip/ankle/knee, and back) show similar recovery patterns.

Table 1 also shows the characteristics of the sick-listed employees for those who did and those who did not participate in a graded return-to-work arrangement. We define an individual to be in graded return-to-work when his wage value, e.g. the degree of pre-sickness productive work time resumption, exceeds 0%. Roughly 60% of the individuals in our data set participate in graded return-to-work at some point during their sick spell. The two groups are comparable in terms of age, gender, and moment of application; the differences in means are statistically significant in some cases, but not substantial. The graded individuals do not earn significantly more than the non-graded individuals. The compositions of the groups are slightly different when it comes to the diagnoses. For example, people who have a conflict at work rarely enter a graded return-to-work trajectory. Presumably, cooperation of the employer and possibly work place adaption is more troublesome in situations where there is a conflict.

Those in graded return-to-work have on average less time devoted to them by their case manager than those who are not in graded return-to-work. Despite the longer average sickness duration, those participating in graded return-to-work have a higher probability of returning to work in the longer run. This is also reflected in Figure 4 showing survival probabilities and hazard rates for individuals who started a graded return-to-work in the first year of their sick leave and for individuals who did not start a graded return-to-work

Figure 4: Survival and hazard rates for individuals with and without graded return-to-work in first year of absence



in the first year, respectively. Individuals participating in graded return-to-work have a lower probability to recover in the first weeks of illness, but start to perform better than those not participating in graded return-to-work from about the 25th week onward leading to substantially lower probabilities of non-recovery in the 70th week. From that point on the lines run roughly parallel to each other. The hazard rate spikes after the first year of sick-leave and at the end of the second year. These spikes correspond to the two annual evaluation moments in the Gatekeeper Protocol.

3.2 Characteristics of case managers

Table 2 shows case manager characteristics of our sample. We have information on 68 case managers, who are predominantly female (70.6%). They have on average about 68 sick-listed employees assigned to them per year. There is quite some spread however, with case managers treating up to 123 individuals a year at maximum. We dropped those case manager-years in which a case manager treated less than 25 individuals in a particular year.[10]

In principle individuals are assigned to case managers based on caseload. That is, new clients are directed to those who have time. However, there seems to be some clustering at

[10] In the appendix, we present the results of robustness analyses that take different cutoffs (see Tables 18, 19, 20 and 21 in Appendix B). When setting the cutoff too low, the average behavior of case managers with only a few clients is more likely to be a poor representation of grading practices. This will weaken the explanatory power of the instrument. When setting the cutoff too high, however, many observations need to be dropped, thus decreasing the efficiency of the estimations. As we will show, both the point estimates as the standard errors turn out to be hardly affected by the choice of cutoff.

Table 2: Descriptive statistics of the 68 case managers

	mean	sd	min	max
a. characteristics of case manager				
female	70.6%			
age on 1 nov 2014	39.1	10.1	25	65
number of clients per year	68.4	23.1	25	123
b. characteristics of the clients of case managers				
fraction of clients female	48.5%	14.8%	20.9%	76.6%
average age at start of case management	42.4	1.7	37.6	46.1
weeks until start of case management	9.1	1.1	60.4	11.1
average gross pre-sickness wage (euro/day)	253.08	242.16	76.45	1317.26
median gross pre-sickness wage (euro/day)	108.12	5.04	84.36	110.00
fraction of clients from firm size categories				
- 1 employee	15.1%	5.7%	2.6%	30.6%
- 2 to 9 employees	36.5%	5.8%	24.0%	51.9%
- 10 to 49 employees	35.4%	7.8%	13.3%	56.0%
- 50 or more employees	2.8%	3.5%	0.0%	23.1%
- number of employees unknown	10.2%	3.4%	3.6%	18.3%
fraction of clients with condition type				
- general medical - mild	8.3%	6.8%	0.0%	28.8%
- general medical - medium	13.3%	6.0%	3.7%	41.0%
- general medical - severe	10.9%	4.6%	0.0%	25.3%
- physical - mild	7.4%	7.0%	0.0%	39.3%
- physical - severe	3.7%	2.8%	0.0%	17.6%
- neck, shoulder, arm complaints	6.7%	3.9%	0.0%	19.0%
- hip, ankle, knee complaints	6.4%	3.8%	0.0%	16.4%
- back complaints	7.2%	3.2%	0.0%	17.5%
- psychiatric	1.7%	1.4%	0.0%	6.3%
- psychological - mild	11.6%	7.8%	0.0%	40.7%
- psychological - severe	2.8%	3.0%	0.0%	19.3%
- psychosocial - mild	10.4%	7.2%	0.0%	33.1%
- psychosocial - severe	1.7%	1.8%	0.0%	8.5%
- social problems	2.3%	3.6%	0.0%	21.4%
- conflict	4.1%	2.5%	0.0%	11.9%
- other[a]	1.5%	2.0%	0.0%	8.0%
c. activities and results of case managers				
fraction of clients in graded work	60.2%	8.2%	33.9%	77.4%
average time allocated to client (min/week)	17.0	3.1	10.6	28.4
average weeks until closing of file	41.0	6.2	21.2	57.2
fraction of clients returned to work within one year	60.8%	10.3%	23.3%	92.0%
fraction of clients returned to work within two years	76.9%	8.5%	40.7%	94.1%

[a] 'Other' contains conditions such as flu and complaints due to pregnancy.

certain case managers based on gender and type of diseases. More specifically, the spread of the case manager averages is relatively high for these variables. This could hint at some form of specialization, in the sense that case managers select those individuals that they know best how to deal with. However, when it comes to the diagnoses of the clients, the variation is more likely to be a result of the reporting behaviour of the case managers than reflecting selection. This is because the diagnoses are established by the case managers after the clients are assigned to them. The results from the sensitivity checks reported in Section 5.4 will show that our results are unlikely to be driven by potential specialization of case managers.

Case managers differ substantially in their use of graded return-to-work, with some case managers only having 33.6% of their clients in graded return-to-work and others having up to 82.6% of their clients participating in graded return-to-work. Also the average time allocated to the individuals vary greatly among case managers.

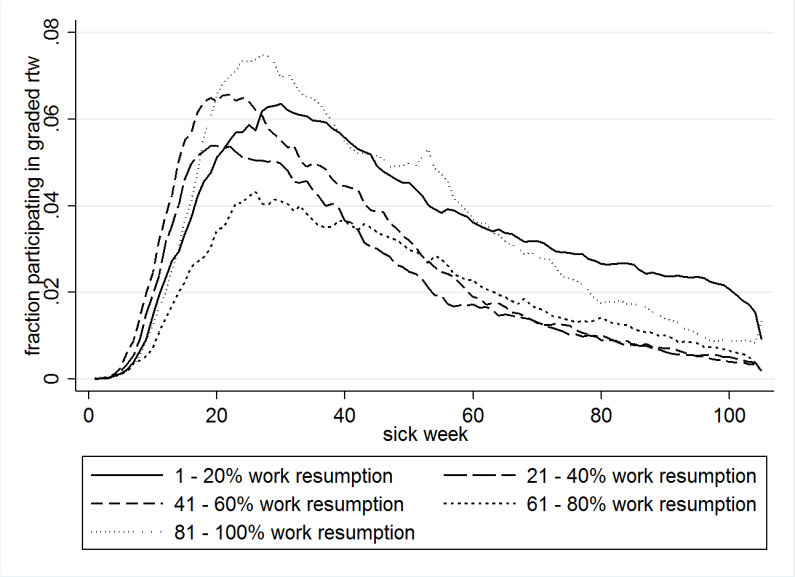
3.3 Setup of graded return-to-work trajectories

Within the group of clients that started a graded work trajectory, relevant outcomes measures are the moment and the degree at which grading is started. The variable 'wage value', which we use to construct our graded return-to-work index, may contain any integer value ranging from 0 to 100 and can be updated up to 24 times at maximum in a two-year-trajectory. Case managers are encouraged to fill in the variable succinctly, as any degree of work resumption implies lower costs for the workplace reintegration provider. The extent to which we can use this detailed information depends on the variation in the graded return-to-work trajectories. In this section we explore the different trajectories in detail.

Figure 5 shows the percentage of individuals participating in graded return-to-work in a certain week, where we define five categories of graded work: 1-20%, 21-40%, 41-60%, 61-80%, and 81-100% of the pre-sickness wage value, respectively.[11] The figure shows that in the first weeks of sickness individuals usually work modest amounts of time (21-60%

[11] When calculating this percentage, we include individuals from the first sick day up until the end of the 105th sick week (so also after recovery). As a result, the numerator remains unchanged.

Figure 5: Percentage of individuals having participated in graded return-to-work per week.



graded work). Towards the 20th week, individuals participate more often in high degrees of graded work resumption (81-100%) or very low degrees (<20%). In the later weeks (when most have recovered), those who are still in graded return-to-work mostly work modest amounts of time, i.e. $< 20\%$ graded work resumption.

Table 3 shows the variation in grading practice of the different case managers. On average case managers wait 20.85 weeks before starting the graded return-to-work and do so at a degree of 36.01%. The fastest case manager waits on average 12.56 weeks and the slowest 25.92. The case manager that starts grading at the lowest degree does so at 28.26% on average and the one that starts the highest does so at 55.15% on average. There are some case managers that never start a graded return-to-work arrangement after 32 weeks, while others start almost a third of the trajectories that late. Also, some case managers never start a graded return-to-work arrangement at 1-20% of pre-sickness wage value, whereas others start almost half the arrangements at this level. We thus conclude there is quite some variation in the grading practice of the different case managers.

Table 3: Variation in grading practices across case managers.

	mean	sd	min	max
average weeks waited until start of graded rtw	20.85	2.83	12.56	25.92
average degree of grading at start of graded rtw	36.01%	4.24%	28.26%	55.15%
fraction of graded rtw that started:				
1 - 8 weeks	13.90%	5.82%	3.85%	31.34%
9 -16 weeks	35.14%	6.39%	22.95%	55.56%
16 - 24 weeks	22.42%	6.07%	8.96%	36.84%
24 - 32 weeks	11.97%	3.84%	3.70%	23.08%
after 32 weeks	16.56%	6.51%	0.00%	28.32%
fraction of graded rtw started at a grade between:				
1 - 20% of pre-sickness wage value	26.4%	8.5%	0.0%	47.4%
21 - 40% of pre-sickness wage value	34.6%	7.3%	7.1%	60.0%
41 - 60% of pre-sickness wage value	31.3%	8.7%	17.9%	78.6%
61 - 80% of pre-sickness wage value	4.0%	2.9%	0.0%	15.6%
81 - 100% of pre-sickness wage value	3.7%	2.9%	0.0%	14.3%

4 Estimation strategy

To identify the effectiveness of graded return-to-work at reducing the length of sick spells, we use an instrumental variable (IV) method which was introduced by Duggan (2005). Duggan analyzes how expenses on new drugs affect total medical expenditures by exploiting the variation in psychiatrists' preferences in drugs prescription as an instrument for individual expenses on certain types of new drugs. In a similar fashion, more recent applications exploit variation in strictness of disability examiners and judges in awarding disability benefits (Maestas et al., 2013; French and Song, 2014) and the propensities of employment offices or individual caseworkers to use certain interventions (Dean et al., 2015; Markussen and Røed, 2014; Rehwald et al., 2016; Markussen et al., 2017). Our approach is most similar to Markussen et al. (2012), who exploit variation in physicians' use of graded absence certificates to identify the effect of part-time sick leave on absence duration.

In our case, employees are send to the reintegration provider after some weeks of absence. The provider assigns them to case managers that have substantial discretionary room in choosing specific treatments. Case managers are encouraged to use graded return-to-work whenever possible. However, the actual grading practice may vary among the case managers. First, this is because different case managers may make different assessments

of when an individual is ready to start graded return-to-work and the individuals' ability to work. Second, one cannot simply assign an individual to graded return-to-work in all relevant work environments. The case manager has to negotiate the possibilities of adapted work duties with the employer, who is not always willing to allow for such flexibility (Wainwright et al., 2011).[12] One case manager may be better in this negotiation process than the other, speeding up the process towards graded return-to-work. Hence, whether an individual participates in graded return-to-work and when he starts to do so, may depend on the case manager he is assigned to. This means the case manager's *propensity to grade* can be used to instrument the graded return-to-work variable.

Within the context of the current analysis, the validity of instrumental variables estimation essentially requires four conditions to be met. First, the probability of graded work should be affected by the concerning case managers' propensity to use a graded work for all other individuals that are assigned to him ('relevance'). In light of the time span of four years that is covered, assuming the tendency to use graded work to be constant over time may be too restrictive. Case managers may change their behavior over time, as they may learn from earlier experiences. We therefore construct propensities by case manager for each year in our sample. This also potentially increases the efficiency of our estimates.[13]

Our second condition for IV to work is that sick listed individuals are assigned randomly to case managers. Stated differently, this implies that sick-listed individuals with long and short expected sick durations do not cluster among certain case managers. With the information on sick-listed workers in our data, we can test for randomness by excluding client characteristics in our model. If this yields different coefficient values for graded work, this suggests there is clustering on worker types. In a similar vein, we can re-run the analyses while excluding case managers who have abnormal client group compositions. The results of both of these analyses are reported in Section 5.4.1. Obviously, testing for clustering on unobservable characteristics is more complex, but it should be stressed that

[12] When performing a decomposition analysis of the observed variation in graded-work applications across case managers and employers, we see indeed that the individual's employer is more important than the individual's case manager. As long as individual's are randomly assigned to case managers, however, this does not burden our analysis. At most, it decreases the efficiency of our method.

[13] At the same time, the sample size per case manager per time unit should be sufficient.

case managers did not receive more information than the registered data we have. This renders it plausible there was no selection on unobservables.

Third, we rely upon the assumption that graded work effects are not correlated with the general ability of case managers in getting individuals back to work (i.e., the 'exclusion restriction'). For instance, if high quality case managers have a strong tendency to use graded work, the IV model will overestimate the effect of graded work. We therefore will conduct various sensitivity tests that use proxies for the overall quality of case managers. In particular, such proxies include both current and lagged work resumption rates for clients that were assigned to case managers or work resumption rates for individuals that were on graded work already at the moment of intake. The results of these analyses are reported in Section 5.4.2

Finally, individuals who would not be treated by a high propensity case manager, should also not be treated by a low propensity case manager. This monotonicity assumption implies that the graded work propensities should impact all individuals equally in our sample. For instance, this assumption may be violated if one case manager is more inclined to use graded work for individuals with mental issues, but less inclined to use graded work for individuals with musculo-skeletal problems. With this in mind, we will conduct tests on the equality of graded work propensity impacts on the actual use of graded work – i.e, the first stage estimates. The results of these checks are reported in Section 5.4.3.

4.1 Specification of the effect of graded work

When specifying the IV model that estimates the effect of graded-work on the incidence of work resumption and the number of sickness weeks, we closely follow Markussen and Røed (2014) and Rehwald et al. (2016). In these analyses, the aim is to estimate the effect of the provision of graded work (G). As we will show later on, we extend their analysis in two ways. First, we will develop propensities for the weeks waited until the start of graded work (W). For the individuals with graded work, this enables us to estimate the impact of the timing of graded work on full work resumption. Second, we will focus on the impact of the level of graded work at the start of a graded work trajectory (S). For ease of exposition, we will consider a single time period for which we construct case

manager propensities. As argued above, we can extend this by allowing for case manager propensities for each year in our sample.

To start with, we structure the cross sectional data on the sick-listed individuals to a panel where every period t corresponds to one week. We include all individual-weeks in the first year of the sick-spell up to and including the week in which graded work started or, in case of the absence of a graded work treatment, until the sick spell ended (i.e., individual went back to work or entered the DI scheme). Then, we run an OLS regression on a dummy indicating whether the individual is or is not starting to participate in graded work that week. In this regression we control for time constant individual characteristics x_i for individual i (e.g. age, age squared, sex, sick type, log gross (pre-sickness) wage, log gross (pre-sickness) wage squared, firm size, year of application, type of insurance contract, sick duration until application at the re-integration office), together with period dummies ($date_{it}$), and dummies for all possible outcomes of elapsed sick weeks (d_{it}):

$$\begin{aligned} graded_{ijt} &= \mathbf{x}_i' \, \boldsymbol{\theta}^g + \delta_1^g \; d_{it} + \delta_2^g \; date_{it} + u_{ijt}^g, \qquad (1) \\ i &= 1, \dots n \quad \text{(individuals)}, \\ j &= 1, \dots J \quad \text{(case managers)}, \\ t &= 1, \dots T \quad \text{(periods)}, \end{aligned}$$

where u_{ijt} is i.i.d. and clustered at the level of case manager-year combinations. The parameters θ^g, δ_1^g and δ_2^g describe the effects of individual characteristics, the elapsed sick weeks and period dummies, respectively.

Using the estimated individual errors $\hat{u}_{ijt}^g$, we next construct the case manager propensities to treat ψ_i^g. We sum the errors over the periods for every individual i, i.e.

$$\hat{u}_{ij}^g = \sum_{t=1}^{T_i} \hat{u}_{ijt}^g, \qquad (2)$$

where T_i is the last period individual i is at risk of making a transition into treatment. Following Markussen and Røed (2014) and Rehwald et al. (2016), $\hat{u}_{ij}^g$ can be interpreted as the difference between the duration until treatment of individual i and the average

duration until treatment for individuals with the same pre-treatment characteristics as individual i. We next take the average of all $\hat{u}_{ij}^g$ per case manager, while leaving out $\hat{u}_{ij}^g$ for the sick-listed employee concerned, i.e.

$$\psi_i^g = \frac{1}{n_j - 1} \sum_{k \in N_j^{-i}} \hat{u}_{kj}^g, \tag{3}$$

where N_j is the set of individuals corresponding to case manager j. For ease of interpretation, we rescale these ψ_i^g from 0 to 1, with 0 indicating the lowest propensity to use graded work and 1 indicating the highest propensity to use graded work.

In order to estimate the effect of graded return-to-work on the probability to return to work (y_i), we collapse the data to one observation per individual. This observation may either be the probability of work resumption or the number of weeks that have been worked over a certain time window. We estimate the effect of having participated in graded return-to-work on the return-to-work probability, using the propensity to grade (ψ_i^g) as an instrumental variable. We control for the same individual characteristics as in the propensity regressions. This yields the following IV model:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^g + \gamma^g \, \widehat{G}_i + \epsilon_i^g, \tag{4}$$

$$G_i = \mathbf{x}_i' \, \boldsymbol{\pi}^g + \alpha^g \, \psi_i^g + \eta_i^g. \tag{5}$$

4.2 Specification of the effect of timing and initial degree of graded work

Following the IV estimation procedures as in equations (1) to (5), the variation in graded-work propensities of case managers that we exploit essentially stems from two sources. First, case managers show differences in the likelihood of starting graded work interventions. Second, there is variation in the timing of treatments across case managers for those individuals that start graded work. To estimate the isolated impact of the duration until graded return-to-work on absence duration, we select only those individuals that enter graded return-to-work at some point during their sickness absence, next recalculate the propensities as explained in equations (1), (2), and (3) and denote these as ψ_i^w. Next, for this sub-sample of individuals, we define the variable W_i as the number of weeks until the

start of graded return-to-work for individual i and estimate the effect of this variable on the absence durations using ψ_i^w as an instrument:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^w + \gamma^w \, \widehat{W}_i + \epsilon_i^w, \tag{6}$$

$$W_i = \mathbf{x}_i' \, \boldsymbol{\pi}^w + \alpha^w \, \psi_i^w + \eta_i^w. \tag{7}$$

As with any IV model, it is important to stress at this point that our parameter of interest in the above equation, γ^w, should be interpreted as a local average treatment effect (LATE). This parameter denotes the effect of waiting one week extra before starting the trajectory on absence duration for those individuals. This result does not necessarily extrapolate to all individuals or to the whole support of the weeks waited variable, W_i.

Our data also allow us to focus on the tendency of case managers to start graded work at a high or low degree. For this purpose, we calculate a propensity based only on the percentage of pre-sickness hours worked during the first week of graded return-to-work, i.e. the starting level denoted by S_{ij}, for the selected sample of individuals with graded work. We estimate a regression corresponding to equation (1),

$$S_{ij} = \mathbf{x}_i' \, \boldsymbol{\theta}^s + \delta_1^s \, d_i + \delta_2^s \, date_i + u_{ij}^s, \tag{8}$$

$$i = 1, \ldots n \quad \text{(individuals)},$$

$$j = 1, \ldots J \quad \text{(case managers)}, \tag{9}$$

where u_{ij}^s is i.i.d. and clustered with respect to case manager-year combinations. Based on the outcomes of this regression, we calculate similar propensities as in equation (3) for individual i with case manager j. We denote these as ψ_i^s. We instrument the initial degree of grading with the average initial degree of grading for all other sick listed workers that were assigned to this case manager. This enables us to conduct an IV regression as above using the degree of graded work resumption rate at the start of graded return-to-work as the intervention:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^s + \gamma^s \, \widehat{S}_i + \epsilon_i^s, \tag{10}$$

$$S_i = \mathbf{x}_i' \, \boldsymbol{\pi}^s + \alpha^s \, \psi_i^s + \eta_i^s. \tag{11}$$

5 Results

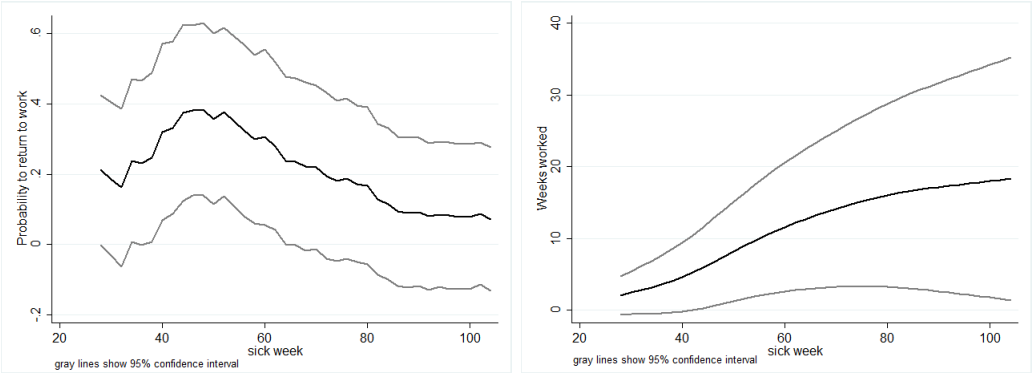
5.1 Overall effects of graded return-to-work

Table 4 shows the effects of graded return-to-work trajectories on (1) a dummy variable indicating whether the sick-listed employee returned to work within 1 year; (2) a dummy variable indicating whether the sick-listed employee returned to work within 2 years; (3) the number of weeks worked in the first year; (4) the number of weeks worked in the first two years. Panel (a) shows the OLS results, panel (b) shows the IV results and panel (c) shows the reduced form or 'Intention-to-treat' estimates for the case manager propensity measure. The results for the regressions underlying the propensities and the estimated coefficients for the control variables of the regressions are shown in Tables 13 and 14 of Appendix B.

Columns 1 to 4 of Table 4 present the baseline results, where we consider an individual as treated if he enters a graded return-to-work trajectory within the first year of sick leave.[14] Based on the OLS results, one would conclude that graded return-to-work trajectories have substantial and positive rehabilitation effects. The IV estimates however show only moderate and statistically insignificant effects, suggesting positive selection into the treatment. This is best illustrated by the outcomes at the end of the second year. The OLS estimates indicate a 30 percentage point increase in return to work probabilities for individuals on a graded return-to-work trajectory, whereas the IV estimates show only a 7.5 percentage point (insignificant) increase. Similarly, the reduced form estimates indicate that individuals assigned to a case manager with the highest propensity to use

[14] 5.3% of untreated individuals do start a graded return-to-work trajectory in the second year of sick leave. Since these trajectories start later in time than outcome variables (1) and (3), we consider these individuals as untreated. When we do consider them as treated and estimate the effects at the end of the two year waiting period, outcome variables (2) and (4), we find slightly smaller effects: return to work probabilities increase by 0.0488 (0.112), the number of weeks worked increases by 1.728 (8.680).

Figure 6: Cumulative effects of graded work per sick weeks



graded return-to-work are only 2 percentage point (insignificant) more likely to rehabilitate within two years than those assigned to the case manager with the lowest propensity to use graded work.

Columns 5 to 8 of Table 4 show the results when only considering graded return-to-work trajectories which started in the first 26 weeks of sick leave as a treatment (individuals who started a graded return-to-work trajectory after the 26 weeks are considered untreated). Compared to the earlier results with 52 weeks as a maximum, there are noticeable differences in the effects. The probability to return to work increases with 30.8 percentage point compared to 12.7 percentage point and the number of weeks worked increases with 8.9 weeks compared to 1.2. One explanation for this difference in outcomes is that graded return-to-work trajectories are more effective when started earlier, which is the hypothesis we will further explore in Section 5.2. Another explanation is that there is a lock-in for graded return-to-work trajectories that occurs in the first weeks of grading. If so, we would expect differences in effectiveness of graded work to fade out over time. This is confirmed when comparing the long-term effects that are shown in column 2 and 6 of Table 4.

To illustrate the evolution of the effects in more detail, Figure 6 shows the effects of graded return-to-work trajectories that started in the first half year on the return to work probability as well as the number of weeks worked. The effect on the return-to-work probabilities is increasing up to week 46, after which the effect declines. It appears that graded return-to-work speeds up the recovery process, with the return-to-work probabilities being almost equal after two years. In line with this, the steep increase in weeks worked between

Table 4: Overall effects of graded return-to-work on full work resumption

Intervention:	**Graded rtw started in week 1-52**				**Graded rtw started in week 1-26**			
	Returned to work		**Weeks worked in**		**Returned to work**		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104	1 year	2 years	week 1-52	week 1-104
(a) OLS estimates								
Graded rtw	0.184***	0.300***	0.251	14.98***	0.280***	0.225***	4.865***	17.78***
	(0.010)	(0.009)	(0.287)	(0.719)	(0.009)	(0.008)	(0.264)	(0.636)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.198	0.181	0.296	0.244	0.239	0.131	0.319	0.262
(b) IV estimates								
Graded rtw	0.127	0.075	1.173	6.642	0.380***	0.070	8.901**	18.30**
	(0.122)	(0.109)	(3.581)	(8.531)	(0.125)	(0.104)	(3.759)	(8.624)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.195	0.117	0.296	0.231	0.230	0.101	0.303	0.262
stage 1: Ψ[graded rtw]	0.270***				0.268***			
	(0.0268)				(0.0267)			
(c) Reduced form estimates of propensity								
Ψ[graded rtw]	0.034	0.020	0.317	1.793	0.102***	0.019	2.386**	4.907**
	(0.033)	(0.030)	(0.970)	(2.333)	(0.035)	(0.028)	(1.014)	(2.372)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.167	0.068	0.296	0.201	0.168	0.068	0.297	0.202

Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types, and firm size.
Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant.
Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

week 40 and 60 does not persist, such that the line flattens out towards the end of the second year.

The effect of graded return-to-work spells started in the first half year on weeks worked in the first year is comparable to the effect found in Markussen et al. (2012) with data from sick-listed workers in Norway. They find that part-time sick leave decreases the absence spells with eight to ten weeks. Rehwald et al. (2016) find substantially bigger results, amounting to a 30 week increase in weeks in regular employment in the first year.[15] Contrary to our results, both Markussen et al. (2012) and Rehwald et al. (2016) find positive long run effects. The first shows that employment two years after sick listing increases with 16 to 21 percentage point, the latter finds a increase of 27 weeks worked during the second year and an increase of 26 weeks in the third year. When comparing these outcomes with ours, one should bear in mind that employers in the Netherlands are committed to facilitate the return-to-work for the sick-listed workers for at least two years. Accordingly, we may expect that individuals in the control group – i.e., those without graded return-to-work – are likely to receive other services. This in turn may explain why the long-term impacts we find are smaller and insignificant. Still, our evidence also suggests that graded return-to-work may speed up the recovery process, particularly when starting early.

5.2 Effects of the timing and initial level of graded work

We argued earlier that both the timing and the initial level of graded work may determine the effectiveness of graded work trajectories. To investigate the importance of these two parameters, we select the sample of individuals who entered a graded return-to-work trajectory in the first year of their sick leave. Using a similar setup as for our benchmark model, we first estimate the effect of starting a graded return-to-work trajectory one week later on the outcome measure. The results are reported in Panel (a) of Table 5. The first stage results show that being assigned to the case manager with the highest propensity of graded work leads to a four week reduction in waiting time until graded return-to-work,

[15] Markussen et al. (2012) only consider grading decisions made within the first eight weeks of sick leave. In the field experiment of Rehwald et al. (2016), graded return-to-work should be started within four weeks after a meeting which is held in the first eight weeks.

Table 5: Effect of starting graded return-to-work one week later or at a higher starting level: IV estimates.

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
(a) Duration until start of graded return to work trajectory				
Sick weeks until grading start	-0.044***	-0.001	-1.497***	-2.177***
	(0.010)	(0.005)	(0.257)	(0.489)
n	5,906	5,906	5,906	5,906
R^2	-0.028	0.033	0.105	0.111
stage 1: ψ[weeks waited]	-3.935***			
	(0.791)			
(b) Level of work resumption at start (linear specification)				
Starting level (0-100)	0.006***	0.003**	0.135***	0.318***
	(0.002)	(0.001)	(0.052)	(0.111)
n	5,913	5,913	5,913	5,913
R^2	0.142	0.018	0.314	0.199
stage 1: Ψ[degree grading]	25.37***			
	(0.564)			

Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size.
Propensities are calculated on the sample of graded individuals. Claimants are excluded when their assigned case manager graded fewer than 25 claimants in the same year as the claimant.
Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

as compared to the case manager with the lowest propensity. The second stage results indicate that waiting one week extra before starting graded return-to-work, decreases the probability to rehabilitate within one year with 4.4 percentage point, whereas the probability to return to work within two years is not affected. This again suggests that graded return-to-work speeds up the recovery process, rather than increasing the long-term probability of recovery. Starting graded return-to-work one week later results in a trajectory that lasts half a week longer, so that the number of weeks worked in the first year decreases by 1.5. When taking a time horizon of two years, the number of weeks worked even decreases by 2.2.

In panel (b) of Table 5 we consider the effect of the level of work resumption at the start of the graded return-to-work trajectory on work resumption. In this setup, the instrument orders case managers in terms of their preference to start a return-to-work trajectory at high levels of graded work resumption. As the table shows, being assigned to a case manager that tends to start trajectories at high rates rather than to one that tends to start at low rates, increases the starting level of work resumption by 25 percentage point. From the second stage estimates we infer that starting at a 10 percentage point higher level of grading results in a 6 percentage point higher chance of recovering in the first year.[16] The return to work probability for the first two years is increased by 2.5 percentage point. The number of weeks worked in the first year increases by 1.4, whereas individuals work 3.2 additional weeks in the first two years. This suggests that a higher initial level of grading also improves long-term recovery rates. It may thus be that giving sick-listed individuals an easy start by re-introducing them to work for a very limited amount of hours, may actually harm them. It may be that the potential positive effects of graded work cannot be established if the individual cannot properly participate in work processes and is not viewed as a full-fledged employee.

[16]Note that more than 90 % of the trajectories have an initial degree of graded work that is less than 60%. Accordingly, variation in the degrees we study typically reflects differences between one, two or three days of working at the start of graded work.

5.3 Effects for different types of medical conditions

Table 6 shows IV estimates for samples of specific medical conditions that are registered by the reintegration provider. Panel (a) shows the baseline estimates for all graded return-to-work trajectories and panel (b) those for all graded return-to-work trajectories started in the first 26 weeks. The first stage estimation results are similar in size across medical conditions, suggesting that the extent to which case managers can affect the use of graded work is equal across groups in the first 26 weeks of absence. The second stage estimates however vary across medical conditions. While graded return-to-work increases first year return-to-work probabilities substantially for general medical as well as musculo-skeletal problems, it seems to have little effect on workers with mental problems. This corresponds with the findings of Høgelund et al. (2010) and Andren (2014) who both find no effects of graded return-to-work for individuals with mental disorders, but positive effects for individuals with other disorders. Also Hernæs (2017) finds larger effects for individuals with musculo-skeletal problems, than for individuals with psychological problems. After two years of sickness, the effect for individuals with musculo-skeletal problems tends to zero, whereas the effect for individuals with general medical problems remains high. This suggests that graded work can be meaningful for individuals with chronic illnesses or individuals that recover from medical treatments.

Panel (c) and (d) of Table 6 show the effects of the timing and initial level of graded work on work resumption for different medical conditions, respectively. Starting the trajectory one week later decreases the probability to return to work for individuals with general medical problems within one year with 4 percentage point, but the effect becomes small and insignificant after two years of absence. For musculo-skeletal conditions we do not find any significant effects on the probability to return to work. There also is no evidence of effects of starting later in the short run for individuals with mental conditions. In the long run, it even seems harmful to start graded work early for this group. In particular, the effect of starting the graded return-to-work trajectory one week later amounts to an increase in the probability to return to work within two years that is equal to about 3 percentage point. Finally, starting at a higher initial level of graded work resumption

Table 6: IV estimation results on work resumption for different medical conditions.

	General medical		**Musculo-skeletal**		**Mental**	
	Returned to work		**Returned to work**		**Returned to work**	
	1 year	2 years	1 year	2 years	1 year	2 years
(a) Overall effect: trajectories started week 1-52						
Graded rtw	0.572*	0.563**	0.477	-0.203	-0.0234	-0.108
	(0.327)	(0.244)	(0.540)	(0.413)	(0.352)	(0.373)
stage 1: Ψ[graded rtw]	0.191***		0.155		0.170**	
	(0.072)		(0.095)		(0.074)	
(b) Overall effect: trajectories started week 1-26						
Graded rtw	0.789***	0.468**	0.539*	-0.061	0.051	-0.259
	(0.238)	(0.205)	(0.323)	(0.261)	(0.261)	(0.296)
stage 1: Ψ[graded rtw]	0.281***		0.229***		0.266***	
	(0.066)		(0.076)		(0.079)	
(c) Duration untill start trajectory						
Sick weeks until grading start	-0.040**	-0.0003	-0.016	0.016	-0.010	0.030*
	(0.017)	(0.008)	(0.011)	(0.010)	(0.018)	(0.017)
stage 1: Ψ[weeks waited]	-5.025***		-5.179***		-3.559**	
	(1.704)		(1.738)		(1.402)	
(d) Initial degree of grading						
Starting level (0-100)	0.005*	0.002	0.010***	0.003	0.007***	0.001
	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)
stage 1: Ψ[degree graded]	30.59***		23.31***		26.71***	
	(0.859)		(1.141)		(0.841)	

The group 'general medical' consists of individuals with the conditions general medical - mild/medium/severe. The group musculo-skeletal consists of individual with the conditions neck, shoulder, arm, hip, ankle, knee or back complaints. The group mental consists of individuals with the conditions psychiatric, psychological - mild/severe, psychosocial - mild/severe or social problems. Individuals with physical mild/severe conditions are not considered because of the small sample size. Also individuals labels as 'other' or having a conflict are excluded.
Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size.
Claimants are excluded when their assigned case manager treated fewer than 10 claimants of the same type in the same year as the claimant.
Panel (a) and (b) is based on 3,971 observations with general medical conditions, 1,947 with musculo-skeletal conditions, and 3,380 with conditions related to mental health.
Panel (c) and (d) is based on 1,667 observations with general medical conditions, 982 with musculo-skeletal conditions, and 1,807 with conditions related to mental health.
Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

results in higher probabilities to return to work after one year and no significant effect after two years for all three types of medical conditions. The results for weeks worked correspond to the results for the return-to-work probability and can be found in Table 17 of Appendix B.

5.4 Sensitivity tests

5.4.1 Endogeneity and specialization

We stated earlier that new clients were assigned to case managers based only on their caseload. As a result, there would be no specialization of case managers that could result in a positive correlation between the propensity to grade and the likelihood to return to work for reasons other than graded return-to-work itself. Based on observed characteristics of clients that are assigned to the same case managers, we can test for this assumption and biases that may stem from specialization. At the same time, we should bear in mind that for analyses based on samples of graded individuals only, there may be non-random selection – even if there is no specialization to start with. This may occur if for instance case managers that only grade few individuals do this because they only grade the most easy cases, which are also easy to grade early and at a high starting degree. In that case, there will be a positive correlation between the weeks waited propensity (or the degree grading propensity) and the likelihood to return to work, for other reasons than starting the trajectory early (or starting at a high degree).

To ensure that potential non-random distribution of clients over case managers does not affect our results, we run a set of sensitivity analyses which are reported in Table 7. As the most prominent effects on graded work were found in the first year of absence, we focus on the return-to-work dummy within one year as the variable that is to be explained. First, we re-run the regressions while excluding specific sets of covariates. If these covariates are correlated with both the probability to recover as well as the propensity to grade, the baseline analysis is subject to omitted variable bias. We exclude sick types in column (2), sick weeks until application in column (3), and all covariates expect the time dummies in column (4). Overall, the results are similar to the baseline. As for the overall effect

Table 7: Sensitivity tests for specialization effects for return-to-work within one year

Dependent: **rtw within 1 year**	(1) Baseline	(2) Exclude covariates	(3) Exclude covariates	(4) Exclude covariates	(5) Exclude abnormal groups	(6) Exclude abnormal groups	(7) Include
		Sick type	Weeks until application	All expect year dummies	>3 sd from mean	>2 sd from mean	graded work propensity
(a) Overall effect: trajectories started in week 1-52							
Graded rtw	0.127	0.130	0.099	-0.027	0.253**	0.263	
	(0.122)	(0.137)	(0.124)	(0.132)	(0.118)	(0.279)	
stage 1: Ψ[graded rtw]	0.270***	0.254***	0.266***	0.271***	0.285***	0.233***	
	(0.027)	(0.024)	(0.026)	(0.023)	(0.032)	(0.050)	
(b) Overall effect: trajectories started in week 1-26							
Graded rtw	0.380***	0.477***	0.370***	0.391***	0.331***	0.340*	
	(0.125)	(0.140)	(0.121)	(0.138)	(0.093)	(0.192)	
stage 1: Ψ[graded rtw]	0.268***	0.239***	0.274***	0.255***	0.301***	0.322***	
	(0.027)	(0.024)	(0.028)	(0.025)	(0.027)	(0.061)	
(c) Duration until start trajectory							
Weeks waited	-0.044***	-0.040***	-0.050***	-0.043***	-0.037***	-0.026***	-0.049***
	(0.010)	(0.007)	(0.013)	(0.007)	(0.009)	(0.009)	(0.013)
Ψ[graded rtw]							-0.060
							(0.053)
stage 1: Ψ[weeks waited]	-3.935***	-5.345***	-3.351***	-5.079***	-4.230***	-5.450***	-3.550***
	(0.791)	(0.769)	(0.837)	(0.776)	(0.781)	(1.012)	(0.853)
(d) Initial degree of grading							
Degree grading (0-100)	0.006***	0.005**	0.006***	0.004*	0.007***	0.005	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)
Ψ[graded rtw]							0.107**
							(0.045)
stage 1: Ψ[degree grading]	25.37***	24.57***	25.54***	25.07***	25.66***	26.13***	25.46***
	(0.564)	(0.576)	(0.560)	(0.596)	(0.670)	(1.479)	(0.537)

Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant.
Panel (a) and (b) are based upon 11,741 observations of which 8,464 remain in column (5) and 3,807 in column (6).
Panel (c) and (d) are based upon 5,913 observations of which 4,492 remain in column (5) and 2,105 in column (6).
Clustered (case manager - year) standard errors between parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

estimates for trajectories that started in weeks 1 to 52, we find the most substantial difference from the baseline occurs when excluding all variables, with a decrease in the point estimate from 0.127 (0.122) to -0.0270 (0.132); see panel (a). Coefficient estimates are hardly affected when we concentrate on trajectories starting in the first 26 weeks of absence; see panel (b). It is only when we exclude the sick type dummies that the point estimate increases from 0.380 to 0.477. To put these findings in perspective, it is important to bear in mind that sick type is not known at the start of the case management trajectory, but determined by the case manager after the client is assigned to him. Differences in reporting a condition for example as general medical or physical, may be influenced by case manager beliefs. These beliefs may in turn be correlated with the propensity to grade. A similar explanation may also hold for the change in the effect estimate of the initial degree of grading that occurs when we exclude sick type dummies – see panel (d). In all cases, the sizes of the difference in point estimates do not lead to concerns about the validity of our approach.

As a second sensitivity check, we exclude case managers with abnormal client group compositions. We define a group composition by the group averages of the characteristics of the clients per case manager-year combination. If a group average of one of the characteristics is more than three (column 4) or two (column 5) standard deviations away from the mean of the group averages the group composition is defined as abnormal. In effect, it means that if a case managers has an extremely high or low number of clients of the same sick type or gender or extremely high or low average ages, sick duration until application at the office, or wage levels among his clients, the clients belonging to this case manager in the respective year are removed from the sample. Excluding these observations results in slightly smaller point estimates than the baseline, but not statistically significantly different.

Finally, we have conducted similar sensitivity tests that apply to the sample with individuals with graded work only – i.e., the analyses on timing and initial degree of graded work. In particular, we add the propensity used in panel (a) to the regressions in panel (c) and (d), to check if the weeks waited and probability to recover are correlated

with the overall propensity to grade.[17] With results that are virtually identical, the picture that emerges is that sample selection effects are negligible.

5.4.2 Case manager quality

Graded return-to-work is only one of the pieces in the case manager's toolbox. He may also make use of other interventions such as paramedical care, job training, or coaching. Or he may assert control by contacting the employee at the right moments and giving valuable advice. A case manager thus can be effective in many different ways. For the validity of our approach, we assume that the case managers' propensity to grade is not correlated with overall case manager quality – i.e., the exclusion restriction. This assumption may not hold when, for example, high quality case managers are also better at motivating the employer and employee in setting up graded return-to-work arrangements, so that there exists a positive correlation between the propensity to grade and the likelihood to return to work for reasons other than graded return-to-work itself. Or, on the contrary, it could be that lower quality case managers tend to overestimate the ability of individuals to participate in graded return-to-work and enter individuals in graded return-to-work too early, leading to a negative correlation between propensities to grade and the likelihood to return to work.

A straightforward measure of case manager quality is his success: does he or she manage to get individuals back to work quickly? Similar to the propensity to grade, we therefore define a 'propensity to cure' that measures the ability of the case manager to get individuals, other that the individual concerned, back to work quickly. Column (2) of Table 8 shows the results of the IV regressions when we control for case manager quality using this propensity to cure. Again, we take the return-to-work within one year as the relevant outcome measure. Being appointed to the highest quality case manager rather than the lowest quality case manager increases the likelihood to return to work with 36 percentage point. At the same time, the effect of graded return-to-work itself decreases to -0.027 when all trajectories are considered and to 0.191 when only the trajectories in the

[17] The correlation between the propensity to grade and the weeks waited propensity equals 0.3433. The correlation between the propensity to grade and the degree grading propensity equals -0.240.

first half year are considered. Following these results, one could conclude that half of the effect of the graded return-to-work trajectories started in the first half year could actually be ascribed to general case manager quality and grading itself is less effective on its own.

That being said, using case manager quality in the way described may be problematic for the same reason as not controlling for quality at all. In particular, the overall quality that we measure may partly be due to the appropriate usage of graded return-to-work, such that the propensity may absorb a too large part of the effect.[18] Under the assumption that case manager quality is relatively constant[19], but the usage of graded work may vary, we could partly resolve this problem by using a lagged quality measure. Using a lagged quality measure as a control variable, we find point estimates of the effect of graded work – as shown in panels (a) and (b) – that are virtually equal to those in the baseline model. Moreover, the maximum effect of case manager quality on work resumption is about 20 percentage point.[20]

Lastly, the original data set also included individuals that already participated in graded return-to-work before entering case management. These individuals were excluded from the sample for the baseline regressions, as the case manager had no influence on their graded return-to-work status. As an auxiliary source of information, we calculated the case managers' propensity to cure on the sample of individuals graded at the start, for each case manager-year combination with at least ten observation in the sample graded at start. As such, we have a proxy of case manager quality apart apart from the ability to appropriately use graded return-to-work.[21] Using this measure of quality we find only a slight decrease in the point estimate of graded work. Based on these estimates, it thus seems that the quality of the case manager does not drive the effects we find. Being appointed to the highest quality case manager rather than the lowest quality case manager does increase the likelihood to return to work with about 14 percentage point.

[18] The correlation between the propensity to grade (within the first half year) and the propensity to cure is 0.200.

[19] In principle some may have more natural ability at the job than others. At the same time, there may be room to increase job performance by learning from past cases or through training.

[20] When we assume case manager quality is fully constant over time, we could also control for it using case manager fixed effects. In that case the effect of graded return-to-work trajectories started in the first half year on work resumption within one year is estimated at 0.268.

[21] The case manager can still exert influence on the level of grading throughout the rest of the trajectory.

Table 8: Sensitivity tests for the importance of case manager quality

Dependent: **rtw within 1 year**	(1) Baseline	(2)	(3)	(4)
		Quality: Propensity to cure		
		Regular	Lagged	Graded at start
(a) Overall effect: trajectories started in week 1-52				
Graded rtw	0.127	-0.027	0.266*	0.080
	(0.122)	(0.088)	(0.143)	(0.133)
Ψ[cure]		0.356***	0.202***	0.143***
		(0.044)	(0.071)	(0.038)
stage 1: Ψ[graded rtw]	0.270***	0.277***	0.267***	0.268***
	(0.027)	(0.027)	(0.032)	(0.027)
stage 1: Ψ[cure]		-0.059*	-0.091**	-0.008
		(0.030)	(0.037)	(0.023)
(b) Overall effect: trajectories started in week 1-26				
Graded rtw	0.380***	0.191**	0.396***	0.323**
	(0.125)	(0.085)	(0.123)	(0.133)
Ψ[cure]		0.342***	0.188***	0.137***
		(0.043)	(0.060)	(0.035)
stage 1: Ψ[graded rtw]	0.268***	0.268***	0.284***	0.257***
	(0.027)	(0.027)	(0.030)	(0.031)
stage 1: Ψ[cure]		-0.0004	-0.036	0.004
		(0.028)	(0.034)	(0.021)
(c) Duration until start of trajectory				
Weeks waited	-0.044***	-0.041***	-0.048***	-0.050***
	(0.010)	(0.009)	(0.010)	(0.011)
Ψ[cure]		0.153**	0.074	0.063
		(0.061)	(0.083)	(0.053)
stage 1: Ψ[weeks waited]	-3.935***	-3.970***	-4.125***	-3.490***
	(0.791)	(0.791)	(0.833)	(0.848)
stage 1: Ψ[cure]		0.398	0.165	-0.324
		(0.894)	(1.094)	(0.741)
(d) Initial degree of grading				
Degree grading (0-100)	0.006***	0.005***	0.005**	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)
Ψ[cure]		0.157**	0.068	0.065
		(0.066)	(0.080)	(0.046)
stage 1: Ψ[degree grading]	25.37***	25.43***	25.51***	25.25***
	(0.564)	(0.574)	(0.757)	(0.608)
stage 1: Ψ[cure]		-0.712	0.560	1.146**
		(0.709)	(0.810)	(0.505)

Claimants are excluded when their assigned case managers treated fewer than 25 claimants that year.
Panel (a) and (b) are based upon 11,741 observations of which 8,319 remain in column (3) and 10,244 in column (4).
Panel (c) and (d) are based upon 5,913 observations of which 4,408 remain in column (3) and 5,591 in column (4).
Baseline results under the sample of individuals included in column (3): (a) 0.180 (0.131), (b) 0.355*** (0.122), (c) -0.0483*** (0.00973), (d) 0.00510** (0.00236).
Baseline results under the sample of individuals included in column (4): (a) 0.107 (0.138), (b) 0.360*** (0.139), (c) -0.0515*** (0.0114), (d) 0.00596*** (0.00209).
Clustered (case manager - year) standard errors between parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

We have also conducted the above-mentioned sensitivity tests on the regressions for the effects of the timing and the initial degree of grading. As panel (c) and panel (d) of Table 8 show, including proxies for case manager quality does not change these results considerably. This suggests that high quality case managers may be inclined to use graded work more often, but not at an earlier stage or at a higher starting level than low quality case managers.

5.4.3 Monotonicity

For the interpretation of our results as local average treatment effects, we need instrumental monotonicity to hold. That is, an individual who would not be treated when assigned to a high propensity case manager, should also not be treated when assigned to a low propensity case manager. This would be violated if certain case managers are more likely to grade individuals with psychological problems, whereas other case managers are more likely to grade individuals with musculo-skeletal complaints. The correlation between the case managers propensity to grade and the individuals grading status should thus be roughly equal for each subgroup of individuals. Panel a of table 9 shows the first stage correlation coefficients for each subgroup of diagnosis. For most diagnoses these are comparable, but not for all. When clustering sick types to the subgroups that we have used earlier in section 5.3, however, we get first-stage results that are very similar (panel b). It thus appears that the differences in first-stage estimates for the detailed subcategories largely stem from small group size.

6 Conclusion

In this paper we investigate the conditions under which graded return-to-work arrangements are most effective at rehabilitating sick-listed employees. We use administrative data from a Dutch private rehabilitation provider and exploit the differences in grading practices between case managers to identify the effect of graded return-to-work. Our analysis relies on the fact that the assignment of new sick-listed clients to case managers is based on caseload. Based on this assumption, we effectively compare the full work re-

Table 9: First stage results for different subcategories

subgroup	overall: started in week 1-52	started in week 1-26	N	duration until start trajectory	initial degree of grading	N
a. detailed subcategories						
general medical - mild	0.174*	0.175	907	-1.040	24.17***	348
	(0.104)	(0.109)		(1.393)	(6.003)	
general medical - medium	0.343***	0.334***	1,588	-6.012**	24.65***	895
	(0.067)	(0.067)		(2.458)	(3.580)	
general medical - severe	0.198***	0.202***	1,350	-8.520***	32.40***	632
	(0.075)	(0.075)		(2.387)	(4.770)	
physical - mild	0.177*	0.120	838	-0.757	25.16***	455
	(0.093)	(0.092)		(2.647)	(7.668)	
physical - severe	0.488***	0.495***	427	1.166	29.52***	216
	(0.099)	(0.101)		(2.103)	(5.026)	
neck, shoulder, arm complaints	0.179	0.228*	810	-0.311	32.03***	463
	(0.146)	(0.127)		(4.670)	(10.29)	
hip, ankle, knee complaints	0.318***	0.391***	743	-7.416***	20.91***	447
	(0.106)	(0.105)		(2.158)	(5.683)	
back complaints	0.081	0.036	860	5.988	50.97	477
	(0.232)	(0.222)		(9.330)	(35.04)	
psychiatric	0.043	0.082	210	-10.67	11.64	86
	(0.196)	(0.195)		(8.060)	(9.802)	
psychological - mild	0.140*	0.172**	1,338	-1.664	24.19***	687
	(0.077)	(0.085)		(1.924)	(4.475)	
psychological - severe	0.042	-0.116	328	-10.58	-2.960	146
	(0.173)	(0.162)		(6.954)	(12.57)	
psychosocial - mild	0.370***	0.288***	1,254	-2.800	20.93***	706
	(0.084)	(0.098)		(2.144)	(3.369)	
psychosocial - severe	0.445***	-0.001	209	1.262	33.69***	127
	(0.156)	(0.169)		(7.123)	(8.998)	
social problems	0.423***	0.481***	244	-7.968***	29.91***	137
	(0.096)	(0.101)		(2.301)	(4.790)	
conflict	0.228	0.440**	464	-1.023	14.54*	67
	(0.197)	(0.187)		(3.003)	(8.334)	
other[a]	0.269**	0.247**	171	-0.255	38.80**	24
	(0.110)	(0.106)		(5.230)	(15.41)	
F-test on equality of coefficients:						
F(15, 181) / F(15,130)	1.98	1.94		1.70	1.53	
p-value	0.0190	0.0220		0.0578	0.1049	
b. rough subcategories						
general medical	0.261***	0.176	3,845	-5.752***	27.18***	1,875
	(0.049)	(0.108)		(1.344)	(2.166)	
musculo-skeletal	0.298***	0.333***	2,413	-5.367***	25.85***	1,387
	(0.060)	(0.067)		(1.377)	(2.621)	
mental	0.246***	0.207***	3,373	-2.527*	21.40***	1,803
	(0.048)	(0.075)		(1.430)	(2.623)	
F-test on equality of coefficients:						
F(2,181) / F(2, 130)	0.21	1.15		1.68	1.00	
p	0.8094	0.3186		0.1909	0.3701	

[a] 'Other' contains conditions such as flue and complaints due to pregnancy.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

sumption rates of case managers with a high propensity to grade to those with a low propensity to grade. We extend this method by also constructing propensities for the speed of starting graded work and the initial level of graded work.

Generally, we find positive effects of graded return-to-work on the number of weeks worked by sick-listed workers. When initiated in the first half year of sick leave, graded return-to-work increases the probability to return to work within one year by 38 percentage point. After two years of absence, however, we do not find statistically significant effects on the probability to return to work. Graded return-to-work increases the number of weeks worked in the first two years after sick-listing with 18 weeks. Overall, these results suggest that graded return-to-work speeds up the recovery process, rather than having a permanent impact on work resumption.

Our evidence suggests that the timing of graded work and the initial level of graded work are crucial determinants of the success of trajectories. Broadly speaking, graded work trajectories should start early and at an initial level that should be sufficiently substantial. Even though starting graded work one week earlier does not affect the return to work rate after two years, it does raise the number of weeks worked in the first two years after sick-listing with two weeks. In addition, starting a graded return-to-work trajectory at a work resumption rate which is 10 percentage point higher increases the probability to return to work within two years with 2.5 percentage point.

The positive effects of graded return-to-work we find are especially strong for individuals who have general medical conditions, such as chronic illnesses. For these the positive effects persist at the end of the waiting period. For individuals with problems related to mental health, however, we find no significant effects of graded return-to-work. For these individuals, speeding up the start of graded work even causes work resumption rates to decrease.

References

Andren, D., 2014. Does part-time sick leave help individuals with mental disorders recover lost work capacity? Journal of Occupational Rehabilitation 24, 344–360.

Andren, D., Svensson, M., 2012. Part-time sick leave as a treatment method for individuals with musculoskeletal disorders. Journal of Occupational Rehabilitation 22, 418–426.

Bernacki, E. J., Guidera, J. A., Schaefer, J. A., Tsai, S., 2000. A facilitated early return to work program at a large urban medical center. Journal of Occupational and Environmental Medicine 42 (12), 1172–1177.

Bethge, M., 2016. Effects of graded return-to-work: a propensity-score-matched analysis. Scandinavian Journal of Work Environment and Health 42 (4), 273–279.

Corrigan, P. W., McCracken, S. G., 2005. Place first, then train: an alternative to the medical model of psychiatric rehabilitation. Social Work 50 (1), 31–39.

de Jong, P., Gielen, M., Haanstra-Veldhuis, V., 2014. Verzekeringsgraad kleine werkgevers - eindrapportage, research commissioned by the ministry of social affairs.

de Jong, P., Lindeboom, M., van der Klaauw, B., 2011. Screening Disability Insurance applications. Journal of the European Economics Association 9 (1), 106–129.

Dean, D., Pepper, J., Schmidt, R., Stern, S., 2015. The effects of vocational rehabilitation for people with cognitive impairments. International Economic Review 56, 399–426.

Duggan, M., 2005. Do new prescription drugs pay for themselves? The case of second-generation antipsychotics. Journal of Health Economics 24, 1–31.

Dutch Association of Insurers, 2016. Dutch insurance industry in figures 2016.

Everhardt, T. P., de Jong, P. R., 2011. Return to work after long term sickness. The role of employer based interventions. De Economist 159, 361–380.

French, E., Song, J., 2014. The effect of Disability Insurance receipt on labor supply. American Economic Journal: Economic Policy 6, 291–337.

Hernæs, Ø., September 2017. Activation against absenteeism: Evidence from a sickness insurance reform in Norway, IZA Discussion Paper No. 10991.

Høgelund, J., Holm, A., Eplov, L. F., 2012. The effect of part-time sick leave for employees with mental disorders. The Journal of Mental Health Policy and Economics 15, 157–170.

Høgelund, J., Holm, A., McIntosh, J., 2010. Does graded return-to-work improve sick-listed workers' chance of returning to regular working hours? Journal of Health Economics 29, 158–169.

Kausto, J., Miranda, H., Martimo, K.-P., Viikari-Juntura, E., 2008. Partial sick leave - review of its use, effects and feasibility in the Nordic countries. Scandinavian Journal of Work Environment and Health 34 (4), 239–249.

Kausto, J., Viikari-Juntura, E., Virta, L. J., gould, R., Koskinen, A., Solovieva, S., 2014. Effectiveness of new legislation on partial sickness benefit on work participation: A quasi-experiment in Finland. BMJ Open 4, e006685.

Koning, P., 2017. Privatizing sick pay: Does it work? IZA World of Labor 324, 1–9.

Koning, P., Lindeboom, M., 2015. The rise and fall of Disability Insurance enrollment in the Netherlands. Journal of Economic Perspectives 29, 151–172.

Krause, N., Dasinger, L. K., Neuhauser, F., 1998. Modified work and return to work: A review of the literature. Journal of Occupational Rehabilitation 8 (2), 113–139.

Maestas, N., Mullen, K. J., Strand, A., 2013. Does Disability Insurance receipt discourage work? Using examinar assignment to estimate casual effects of SSDI receipt. American Economic Review 103, 1797–1829.

Markussen, S., Mykletun, A., Røed, K., 2012. The case for presenteeism - Evidence from Norway's sickness insurance program. Journal of Public Economics 96, 959–972.

Markussen, S., Røed, K., 2014. The impacts of vocation rehabilitation. Labour Economics 31, 1–13.

Markussen, S., Røed, K., Schreiner, R. C., 2017. Can compulsory dialogues nudge sick-listed workers back to work? The Economic Journal, doi:10.1111/ecoj.12468.

OECD, 2010. Sickness, disability and work: Breaking the barriers. A synthesis of findings across OECD countries. OECD Publishing, Paris.

Rehwald, K., Rosholm, M., Roulande, B., February 2016. Does activating sick-listed workers work? Evidence from a randomized experiment, IZA Discussion Paper No. 9771.

van Sonsbeek, J.-M., Gradus, R. H. J. M., 2013. Estimating the effects of recent disability reforms in the Netherlands. Oxford Economic Papers 65, 832–855.

Viikari-Juntura, E., Kausto, J., Shiri, R., Kaila-Kangas, L., Pekka Takala, E., Karppinen, J., Miranda, H., Luukkonen, R., Pekka Martimo, K., 2012. Return to work after early part-time sick leave due to musculoskeletal disorders: A randomized controlled trial. Scandinavian Journal of Work Environment and Health 38 (2), 134–143.

Wainwright, E., Wainwright, D., Keogh, E., Eccleston, C., 2011. Fit for purpose? Using the fit note with patients with chronic pain: A qualitative study. British Journal of General Practice 61 (593), 794–800.

A Additional data descriptives

Table 10: Data selection steps

Selection step	Number of observations
Total number of clients	35,040
Selection on contract type and insurer [a]	- 14,156
Individual died or left because of problems with insurance	-139
No case management/reported goal different than back to work	- 2,093
Intervention/graded rtw took place before application at Keerpunt	-5,030
Implausible dates	- 121
Individual could not have been observed for two years at October 7, 2016	- 5
year 2009 - 2010 deleted (only few observations)	- 221
Observations left	13,275
Individuals excluded due to missing values or being assigned to Case managers with less than 25 clients that year	- 1,534
Observations used for analysis	11,741

[a] Different contract types follow different processes leading onto application at the workplace reintegration provider. The selected contract types follow similar procedures. The main criterion for selection was that the individuals should not have been in contact with the workplace reintegration provider before the application date.

Table 11: Additional case manager characteristics

average working hours per year:	
- less than 800	33.8%
- 800 - 1000	20.6%
- more than 1000	26.5%
- unknown	19.1%
senior reintegration specialist	7.4%
education:	
- secondary/vocational	10.3%
- bachelor	54.4%
- master/docterate	35.3%
workplace education	
- less than 10 courses	16.2%
- 10 - 19 courses	44.1%
- 20 - 29 courses	26.5%
- 30 - 39 courses	10.3%
- 40 or more courses	1.5%
- unknown	0.0%

Table 12: Reasons for sick-listing

	Freq.	Percent
general medical - mild	1,509	8.75
general medical - medium	2,423	14.05
general medical - severe	1,827	10.59
physical - mild	1,304	7.56
physical - severe	625	3.62
neck, shoulder, arm complaints	1,199	6.95
hip, ankle, knee complaints	1,092	6.33
back complaints	1,324	7.68
psychiatric	291	1.69
psychological - mild	1,895	10.99
psychological - severe	440	2.55
psychosocial - mild	1,787	10.36
psychosocial - severe	298	1.73
social problems	384	2.23
conflict	557	3.23
other	290	1.68
total	17,245	100

B Additional results

The propensities are calculated based on an an OLS regression using all individual-week observations up to and including the first week into graded work or up to the end of the first year of sickness, using a dummy indicating whether one entered graded work as left hand side variable and individual characteristics as right hand side variables. The results are shown in panels (a) of Tables 13, 14, 15, and 16. Females are more likely to be assigned to graded work. The likelihood to participate is hump shaped in age and income. Those who apply later are more likely to participate in graded work. In the year 2012 people were less likely to participate in graded work. The results do not seem to differ substantially for the different categories. The coefficients for the sick week dummies are plotted in Figure 7. Using the errors from these regressions we calculate the year-case manager propensities to treat. The distribution of these propensities, before rescaling, is shown in Figure 8.

Panels (b) and (c) of Tables 13, 14, 15, and 16 show all the coefficient estimates for the control variables of the regressions underlying the baseline IV estimates in Tables 4 and 5.

Table 17 show the results for different medical conditions corresponding to Table 6 for the other to outcome measures. Tables 18, 19, 20, and 21 show the results using different cut-offs for the minimum number of clients per case manager.

Table 13: Effect of graded return-to-work when started in week 1-52, including coefficients on control variables.

(a) Stage 0 - dependent: participates in graded return-to-work

sex	0.000	(0.001)	condition:			contract type:		
age at application	0.001***	(0.000)	general medical - medium	0.010***	(0.001)	B	0.001	(0.002)
age at application^2	0.000***	(0.000)	general medical - severe	0.000	(0.001)	C	0.003	(0.002)
ln(gross wage)	0.005***	(0.001)	neck, shoulder, arm complaints	0.007***	(0.001)	D	0.001	(0.002)
ln(gross wage)^2	0.000***	(0.000)	physical - mild	0.008***	(0.001)	E	0.003**	(0.002)
sick weeks until application	0.001***	(0.0001)	physical - severe	0.004**	(0.002)	F	0.008***	(0.002)
sick weeks until application^2	0.000	(0.000)	hip, ankle, knee complaints	0.012***	(0.002)	G	0.003	(0.002)
application year:			other	-0.011***	(0.002)	H	0.004**	(0.002)
2012	-0.001	(0.002)	psychiatric	-0.002	(0.002)	I	0.004**	(0.002)
2013	0.001	(0.002)	psychological - mild	0.004***	(0.001)	firm size:		
2014	0.002	(0.003)	psychological - severe	-0.002	(0.002)	2-9 employee	0.001*	(0.001)
Constant	-0.016	(0.088)	psychosocial - mild	0.006***	(0.001)	10-49 employees	0.004***	(0.001)
			psychosocial - severe	0.005*	(0.002)	50 or more employees	0.010***	(0.002)
			back complaints	0.007***	(0.001)	unknown	0.001	(0.001)
Observations	290,929		social problems	0.008***	(0.002)			
R-squared	0.011		conflict	-0.012***	(0.002)			

(b) Stage 1 - dependent: participates in graded return-to-work

ψ	0.270***	(0.027)	condition:			contract type:		
sex	0.002	(0.009)	general medical - medium	0.200***	(0.024)	B	0.027	(0.027)
age at application	0.013***	(0.003)	general medical - severe	0.107***	(0.024)	C	0.049*	(0.027)
age at application^2	0.000***	(0.000)	neck, shoulder, arm complaints	0.193***	(0.024)	D	0.035	(0.025)
ln(gross wage)	0.102***	(0.018)	physical - mild	0.166***	(0.025)	E	0.059**	(0.026)
ln(gross wage)^2	-0.006***	(0.002)	physical - severe	0.159***	(0.032)	F	0.115***	(0.035)
sick weeks until application	0.000	(0.002)	hip, ankle, knee complaints	0.236***	(0.026)	G	0.039	(0.040)
sick weeks until application^2	0.000***	(0.000)	other	-0.254***	(0.039)	H	0.070***	(0.026)
application year:			psychiatric	0.046	(0.039)	I	0.066**	(0.031)
2012	-0.048***	(0.012)	psychological - mild	0.148***	(0.022)	firm size:		
2013	-0.063***	(0.012)	psychological - severe	0.078**	(0.031)	2-9 employee	0.021	(0.015)
2014	-0.070***	(0.013)	psychosocial - mild	0.163***	(0.024)	10-49 employees	0.060***	(0.014)
Constant	-0.311***	(0.076)	psychosocial - severe	0.211***	(0.037)	50 or more employees	0.129***	(0.028)
			back complaints	0.181***	(0.023)	unknown	0.007	(0.018)
Observations	11,741		social problems	0.152***	(0.038)			
			conflict	-0.300***	(0.027)			

(c) Stage 2 - dependent: returned to work within 1 year

intervention	0.127	(0.122)	condition			contract type:		
sex	-0.031***	(0.009)	general medical - medium	-0.169***	(0.030)	B	0.047*	(0.025)
age at application	0.0003	(0.003)	general medical - severe	-0.531***	(0.024)	C	0.042	(0.027)
age at application^2	0.000	(0.000)	neck, shoulder, arm complaints	-0.271***	(0.032)	D	-0.018	(0.028)
ln(gross wage)	0.017	(0.020)	physical - mild	-0.100***	(0.028)	E	-0.002	(0.027)
ln(gross wage)^2	-0.003*	(0.002)	physical - severe	-0.446***	(0.035)	F	0.030	(0.037)
sick weeks until application	-0.010***	(0.002)	hip, ankle, knee complaints	-0.197***	(0.036)	G	0.034	(0.035)
sick weeks until application^2	0.000	(0.000)	other	-0.439***	(0.066)	H	-0.023	(0.027)
application year:			psychiatric	-0.478***	(0.037)	I	0.027	(0.031)
2012	0.171***	(0.017)	psychological - mild	-0.319***	(0.029)	firm size:		
2013	0.181***	(0.020)	psychological - severe	-0.510***	(0.035)	2-9 employee	0.012	(0.014)
2014	0.149***	(0.021)	psychosocial - mild	-0.170***	(0.026)	10-49 employees	0.017	(0.016)
Constant	0.713***	(0.076)	psychosocial - severe	-0.416***	(0.045)	50 or more employees	0.054	(0.035)
			back complaints	-0.274***	(0.030)	unknown	0.017	(0.019)
Observations	11,741		social problems	-0.073**	(0.032)			
R-squared	0.195		conflict	-0.108**	(0.047)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 14: Effect of graded return-to-work when started in week 1-26, including coefficients on control variables.

(a) Stage 0 - dependent: participates in graded return-to-work

sex	-0.001	(0.001)	condition:			contract type:		
age at application	0.001***	(0.000)	general medical - medium	0.006***	(0.002)	B	0.003	(0.003)
age at application^2	0.000***	(0.000)	general medical - severe	-0.017***	(0.002)	C	0.005*	(0.003)
ln(gross wage)	0.008***	(0.002)	neck, shoulder, arm complaints	0.001	(0.003)	D	0.002	(0.003)
ln(gross wage)^2	0.000***	(0.000)	physical - mild	0.008***	(0.003)	E	0.006**	(0.003)
sick weeks until application	0.001**	(0.000)	physical - severe	-0.009***	(0.003)	F	0.011***	(0.004)
sick weeks until application^2	0.000	(0.000)	hip, ancle, knee complaints	0.007**	(0.003)	G	0.007*	(0.004)
application year:			other	-0.027***	(0.004)	H	0.006*	(0.003)
2012	-0.003	(0.004)	psychiatric	-0.020***	(0.004)	I	0.008**	(0.003)
2013	0.000	(0.005)	psychological - mild	-0.005**	(0.002)	firm size:		
2014	0.002	(0.007)	psychological - severe	-0.017***	(0.003)	2-9 employee	0.002	(0.001)
Constant	-0.014	(0.106)	psychosocial - mild	0.001	(0.002)	10-49 employees	0.006***	(0.001)
			psychosocial - severe	-0.013***	(0.004)	50 or more employees	0.014***	(0.003)
			back complaints	0.000	(0.003)	unknown	0.003	(0.002)
Observations	147,713		social problems	0.009**	(0.004)			
R-squared	0.007		conflict	-0.028***	(0.003)			

(b) Stage 1 - dependent: participates in graded return-to-work

ψ	0.268***	(0.027)	condition:			contract type:		
sex	-0.009	(0.009)	general medical - medium	0.088***	(0.025)	B	0.038	(0.028)
age at application	0.007**	(0.003)	general medical - severe	-0.114***	(0.024)	C	0.065**	(0.029)
age at application^2	0.000***	(0.000)	neck, shoulder, arm complaints	0.059**	(0.024)	D	0.035	(0.028)
ln(gross wage)	0.078***	(0.017)	physical - mild	0.104***	(0.027)	E	0.065**	(0.030)
ln(gross wage)^2	-0.005***	(0.002)	physical - severe	-0.037	(0.032)	F	0.085**	(0.034)
sick weeks until application	-0.016***	(0.002)	hip, ancle, knee complaints	0.096***	(0.027)	G	0.069*	(0.039)
sick weeks until application^2	0.000	(0.000)	other	-0.269***	(0.037)	H	0.063**	(0.028)
application year:			psychiatric	-0.163***	(0.036)	I	0.069**	(0.032)
2012	-0.038***	(0.010)	psychological - mild	0.004	(0.023)	firm size:		
2013	-0.033***	(0.011)	psychological - severe	-0.113***	(0.030)	2-9 employee	0.015	(0.014)
2014	-0.052***	(0.012)	psychosocial - mild	0.053**	(0.024)	10-49 employees	0.048***	(0.013)
Constant	-0.005	(0.072)	psychosocial - severe	-0.066	(0.041)	50 or more employees	0.124***	(0.029)
			back complaints	0.048**	(0.024)	unknown	0.022	(0.018)
Observations	11,741		social problems	0.103**	(0.040)			
			conflict	-0.347***	(0.027)			

(c) Stage 2 - dependent: returned to work within 1 year

intervention	0.380***	(0.125)	condition			contract type:		
sex	-0.027***	(0.009)	general medical - medium	-0.176***	(0.020)	B	0.036	(0.026)
age at application	-0.001	(0.003)	general medical - severe	-0.473***	(0.024)	C	0.023	(0.028)
age at application^2	0.000	(0.000)	neck, shoulder, arm complaints	-0.267***	(0.022)	D	-0.028	(0.028)
ln(gross wage)	-0.001	(0.020)	physical - mild	-0.117***	(0.024)	E	-0.020	(0.028)
ln(gross wage)^2	-0.002	(0.002)	physical - severe	-0.411***	(0.026)	F	0.011	(0.035)
sick weeks until application	-0.004	(0.003)	hip, ancle, knee complaints	-0.203***	(0.025)	G	0.011	(0.035)
sick weeks until application^2	0.000	(0.000)	other	-0.370***	(0.064)	H	-0.039	(0.027)
application year:			psychiatric	-0.408***	(0.039)	I	0.008	(0.032)
2012	0.179***	(0.017)	psychological - mild	-0.301***	(0.021)	firm size: (CHECK)		
2013	0.186***	(0.018)	psychological - severe	-0.455***	(0.034)	2-10 employee	0.008	(0.014)
2014	0.160***	(0.019)	psychosocial - mild	-0.168***	(0.020)	10-49 employees	0.006	(0.016)
Constant	0.634***	(0.073)	psychosocial - severe	-0.362***	(0.035)	50 or more employees	0.022	(0.035)
			back complaints	-0.268***	(0.021)	unknown	0.010	(0.018)
Observations	11,741		social problems	-0.090***	(0.026)			
R-squared	0.230		conflict	-0.014	(0.053)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 15: Effect of starting moment of graded return-to-work, including coefficients on control variables

(a) Stage 0 - dependent: participates in graded return-to-work

sex	-0.008***	(0.002)	condition:			contract type:		
age at application	-0.001**	(0.001)	general medical - medium	-0.095***	(0.007)	B	0.008	(0.007)
age at application^2	0.000	(0.000)	general medical - severe	-0.149***	(0.007)	C	0.005	(0.007)
ln(gross wage)	0.006	(0.005)	neck, shoulder, arm complaints	-0.120***	(0.007)	D	-0.005	(0.007)
ln(gross wage)^2	-0.001	(0.000)	physical - mild	-0.087***	(0.008)	E	0.003	(0.007)
sick weeks until application	-0.002***	(0.001)	physical - severe	-0.137***	(0.008)	F	0.003	(0.009)
sick weeks until application^2	0.000***	(0.000)	hip, ancle, knee complaints	-0.114***	(0.007)	G	-0.001	(0.009)
application year:			other	-0.087***	(0.020)	H	-0.001	(0.007)
2012	-0.007	(0.008)	psychiatric	-0.144***	(0.010)	I	0.004	(0.008)
2013	0.006	(0.010)	psychological - mild	-0.122***	(0.007)	firm size:		
2014	-0.008	(0.013)	psychological - severe	-0.153***	(0.008)	2-9 employee	-0.004	(0.003)
Constant	0.273	(0.275)	psychosocial - mild	-0.106***	(0.007)	10-49 employees	-0.002	(0.003)
			psychosocial - severe	-0.149***	(0.009)	50 or more employees	0.000	(0.006)
			back complaints	-0.113***	(0.007)	unknown	0.007	(0.004)
Observations	71,670		social problems	-0.055***	(0.011)			
R-squared	0.027		conflict	-0.076***	(0.013)			

(b) Stage 1 - dependent: participates in graded return-to-work

ψ	-3.935***	(0.791)	condition:			contract type:		
sex	0.838***	(0.268)	general medical - medium	4.687***	(0.478)	B	-0.822	(0.897)
age at application	0.123	(0.079)	general medical - severe	12.37***	(0.563)	C	-0.882	(0.998)
age at application^2	-0.001	(0.001)	neck, shoulder, arm complaints	7.064***	(0.582)	D	0.607	(0.933)
ln(gross wage)	-0.705	(0.654)	physical - mild	3.876***	(0.463)	E	-0.234	(0.965)
ln(gross wage)^2	0.062	(0.056)	physical - severe	10.28***	(0.755)	F	0.059	(1.079)
sick weeks until application	0.841***	(0.058)	hip, ancle, knee complaints	6.403***	(0.550)	G	-0.239	(1.170)
sick weeks until application^2	-0.003*	(0.002)	other	4.237**	(1.834)	H	0.156	(0.991)
application year:			psychiatric	10.77***	(1.368)	I	-0.337	(0.962)
2012	-0.073	(0.338)	psychological - mild	7.849***	(0.572)	firm size:		
2013	-1.238***	(0.326)	psychological - severe	13.21***	(1.166)	2-9 employee	0.290	(0.378)
2014	0.127	(0.377)	psychosocial - mild	5.751***	(0.492)	10-49 employees	0.066	(0.383)
Constant	5.933**	(2.566)	psychosocial - severe	12.50***	(1.203)	50 or more employees	-0.068	(0.838)
			back complaints	6.620***	(0.610)	unknown	-0.753	(0.531)
Observations	5,906		social problems	1.747**	(0.687)			
			conflict	2.983***	(0.939)			

(c) Stage 2 - dependent: returned to work within 1 year

intervention	-0.044***	(0.010)	condition			contract type:		
sex	0.005	(0.015)	general medical - medium	0.131**	(0.054)	B	-0.064	(0.049)
age at application	0.003	(0.005)	general medical - severe	0.105	(0.128)	C	-0.071	(0.056)
age at application^2	0.000	(0.000)	neck, shoulder, arm complaints	0.134*	(0.077)	D	-0.050	(0.050)
ln(gross wage)	0.006	(0.051)	physical - mild	0.098*	(0.051)	E	-0.091*	(0.051)
ln(gross wage)^2	-0.003	(0.004)	physical - severe	0.118	(0.116)	F	-0.020	(0.059)
sick weeks until application	0.028***	(0.009)	hip, ancle, knee complaints	0.143*	(0.074)	G	-0.016	(0.059)
sick weeks until application^2	0.000***	(0.000)	other	0.000	(0.127)	H	-0.093*	(0.052)
application year:			psychiatric	0.114	(0.121)	I	-0.084	(0.053)
2012	0.142***	(0.030)	psychological - mild	0.095	(0.083)	firm size:		
2013	0.129***	(0.034)	psychological - severe	0.137	(0.142)	2-10 employee	0.010	(0.020)
2014	0.126***	(0.031)	psychosocial - mild	0.146**	(0.066)	10-49 employees	-0.003	(0.020)
Constant	1.101***	(0.167)	psychosocial - severe	0.204	(0.139)	50 or more employees	0.034	(0.043)
			back complaints	0.122	(0.075)	unknown	-0.010	(0.027)
Observations	5,906		social problems	0.059	(0.044)			
R-squared	-0.028		conflict	0.025	(0.078)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 16: Effect of initial degree of graded return-to-work, including coefficients on control variables.

(a) Stage 0 - dependent: initial degree of graded return-to-work

sex	-1.566***	(0.566)	condition:			contract type:		
age at application	-0.016	(0.171)	general medical - medium	-6.155***	(1.290)	B	-2.310	(1.702)
age at application^2	0.000	(0.002)	general medical - severe	-11.13***	(1.410)	C	-2.014	(1.793)
ln(gross wage)	-1.440	(1.328)	neck, shoulder, arm complaints	-6.723***	(1.456)	D	-3.437**	(1.737)
ln(gross wage)^2	0.019	(0.117)	physical - mild	-4.538***	(1.444)	E	-2.251	(1.715)
sick weeks until application	0.101	(0.152)	physical - severe	-11.81***	(1.777)	F	-0.986	(2.162)
sick weeks until application^2	0.003	(0.005)	hip, ancle, knee complaints	-5.661***	(1.463)	G	-0.400	(2.347)
application year:			other	6.618	(4.229)	H	-2.102	(1.784)
2012	2.477	(1.891)	psychiatric	-12.34***	(2.448)	I	0.262	(1.928)
2013	3.799	(2.481)	psychological - mild	-11.64***	(1.366)	firm size:		
2014	3.855	(3.172)	psychological - severe	-10.35***	(2.029)	2-9 employee	1.162	(0.826)
Constant	85.950***	(12.63)	psychosocial - mild	-9.015***	(1.344)	10-49 employees	1.321	(0.824)
			psychosocial - severe	-10.02***	(2.133)	50 or more employees	1.320	(1.636)
			back complaints	-8.308***	(1.441)	unknown	0.956	(1.111)
Observations	5,913		social problems	-4.814**	(2.024)			
R-squared	0.098		conflict	4.507*	(2.676)			

(b) Stage 1 - dependent: initial degree of graded return-to-work

ψ	25.373***	(0.564)	condition:			contract type:		
sex	-1.499***	(0.512)	general medical - medium	-8.477***	(1.443)	B	-2.953*	(1.727)
age at application	-0.125	(0.167)	general medical - severe	-14.59***	(1.407)	C	-2.777	(1.727)
age at application^2	0.002	(0.001)	neck, shoulder, arm complaints	-9.749***	(1.653)	D	-4.255**	(1.823)
ln(gross wage)	-1.943	(1.759)	physical - mild	-6.011***	(1.435)	E	-3.643**	(1.780)
ln(gross wage)^2	0.066	(0.140)	physical - severe	-14.91***	(1.629)	F	-1.534	(2.162)
sick weeks until application	-0.189	(0.136)	hip, ancle, knee complaints	-8.306***	(1.515)	G	-0.950	(2.370)
sick weeks until application^2	0.009*	(0.005)	other	2.851	(5.991)	H	-3.603**	(1.712)
application year:			psychiatric	-16.02***	(2.167)	I	-0.883	(1.957)
2012	4.470***	(0.258)	psychological - mild	-14.56***	(1.404)	firm size:		
2013	4.994***	(0.309)	psychological - severe	-14.75***	(2.229)	2-9 employee	1.660**	(0.833)
2014	2.090***	(0.334)	psychosocial - mild	-11.20***	(1.239)	10-49 employees	2.013***	(0.750)
Constant	42.303***	(6.564)	psychosocial - severe	-13.31***	(2.395)	50 or more employees	1.501	(1.688)
			back complaints	-10.44***	(1.472)	unknown	1.985*	(1.075)
Observations	5,913		social problems	-6.074***	(1.721)			
			conflict	3.661	(3.570)			

(c) Stage 2 - dependent: returned to work within 1 year

intervention	0.006***	(0.002)	condition:			contract type:		
sex	-0.023**	(0.013)	general medical - medium	-0.033	(0.025)	B	-0.012	(0.032)
age at application	-0.002	(0.004)	general medical - severe	-0.367***	(0.035)	C	-0.024	(0.033)
age at application^2	0.000	(0.000)	neck, shoulder, arm complaints	-0.133***	(0.033)	D	-0.053	(0.037)
ln(gross wage)	0.042	(0.032)	physical - mild	-0.049*	(0.025)	E	-0.061*	(0.034)
ln(gross wage)^2	-0.006**	(0.003)	physical - severe	-0.258***	(0.049)	F	-0.025	(0.044)
sick weeks until application	-0.007**	(0.003)	hip, ancle, knee complaints	-0.101***	(0.033)	G	-0.002	(0.046)
sick weeks until application^2	0.000**	(0.000)	other	-0.225***	(0.085)	H	-0.080**	(0.035)
application year:			psychiatric	-0.282***	(0.064)	I	-0.067*	(0.039)
2012	0.121***	(0.027)	psychological - mild	-0.174***	(0.039)	firm size:		
2013	0.155***	(0.030)	psychological - severe	-0.377***	(0.052)	2-9 employee	-0.010	(0.018)
2014	0.110***	(0.031)	psychosocial - mild	-0.051*	(0.030)	10-49 employees	-0.019	(0.018)
Constant	0.641***	(0.152)	psychosocial - severe	-0.282***	(0.054)	50 or more employees	0.030	(0.041)
			back complaints	-0.115***	(0.030)	unknown	0.016	(0.028)
Observations	5,913		social problems	0.007	(0.035)			
R-squared	0.142		conflict	-0.132**	(0.060)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 7: Duration coefficients

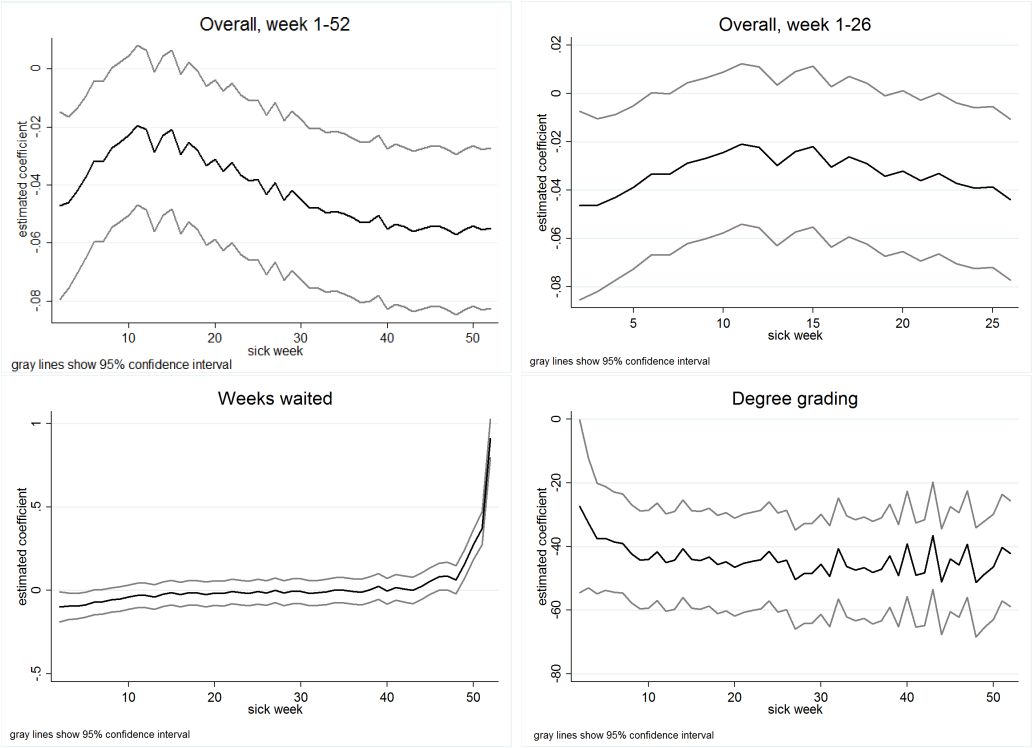


Figure 8: Propensities to treat before scaling.

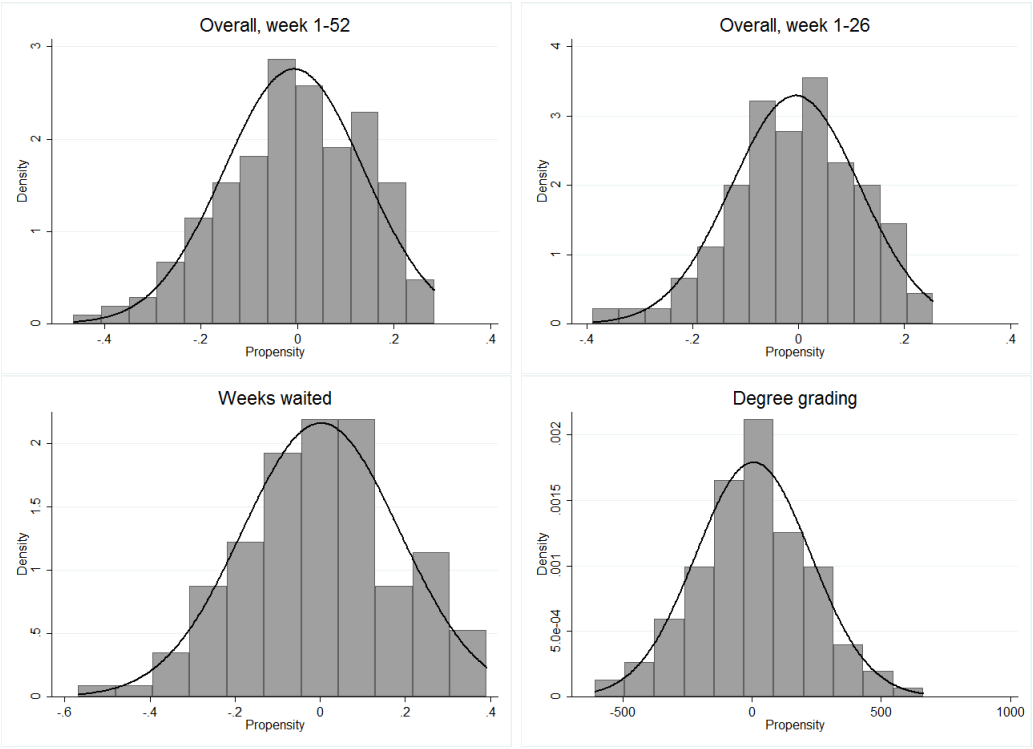


Table 17: IV estimation results for different medical conditions - weeks worked.

	General medical		Musculo-skeletal		Mental	
	Weeks worked in		**Weeks worked in**		**Weeks worked in**	
	week 1-52	week 1-104	week 1-52	week 1-104	week 1-52	week 1-104
(a) Baseline, all trajectories started in week 1-52						
Graded rtw	9.066	30.54	8.680	6.885	7.341	6.824
	(9.855)	(20.87)	(17.21)	(29.57)	(10.85)	(24.74)
stage 1: Ψ[graded rtw]	0.191***		0.155		0.1703**	
	(0.072)		(0.095)		(0.074)	
(b) Baseline, all trajectories started in week 1-26						
Graded rtw	20.06***	42.12***	14.31	18.34	5.354	-0.427
	(7.555)	(15.51)	(11.33)	(21.25)	(7.752)	(19.92)
stage 1: Ψ[graded rtw]	0.281***		0.229***		0.266***	
	(0.066)		(0.076)		(0.079)	
(c) Duration until start of graded return to work trajectory						
Weeks waited	-1.577***	-1.924**	-0.867**	-0.791	-0.532	0.490
	(0.464)	(0.767)	(0.356)	(0.684)	(0.442)	(1.116)
stage 1: Ψ[weeks waited]	-5.025***		-5.179***		-3.559**	
	(1.704)		(1.738)		(1.402)	
(d) Level of work resumption at start						
Degree grading (0-100)	0.156**	0.303**	0.250***	0.546***	0.193***	0.355***
	(0.066)	(0.133)	(0.078)	(0.174)	(0.062)	(0.125)
stage 1: Ψ[degree grading]	30.59***		23.31***		26.71***	
	(0.859)		(1.141)		(0.841)	

The group general medical consists of individuals with the conditions general medical - mild/medium/severe. The group musculo-skeletal consists of individual with the conditions neck, shoulder, arm, hip, ankle, knee or back complaints. The group mental consists of individuals with the conditions psychiatric, psychological - mild/severe, psychosocial - mild/severe or social problems. Individuals with physical mild/severe conditions are not considered because of the small sample size. Also individuals labels as 'other' or having a conflict are excluded.

Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size.

Claimants are excluded when their assigned case manager treated fewer than 10 claimants of the same type in the same year as the claimant.

Panel (a) and (b) is based on 3,971 observations with general medical conditions, 1,947 with musculo-skeletal conditions, and 3,380 with conditions related to mental health.

Panel (c) and (d) is based on 1,667 observations with general medical conditions, 982 with musculo-skeletal conditions, and 1,807 with conditions related to mental health.

Clustered (case manager - year) standard errors between parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 18: Overall results (1-52 weeks) using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
(a) 15 clients or more per caseworker (N=12,534)				
Graded rtw	0.093	0.086	-0.784	4.090
	(0.117)	(0.103)	(3.489)	(8.245)
stage 1: Ψ[graded rtw]	0.385***			
	(0.033)			
(b) 20 clients or more per caseworker (N=12,258)				
Graded rtw	0.129	0.079	0.487	5.821
	(0.115)	(0.109)	(3.375)	(8.189)
stage 1: Ψ[graded rtw]	0.343***			
	(0.032)			
(c) 25 clients or more per caseworker (N=11,741)				
Graded rtw	0.127	0.075	1.173	6.642
	(0.122)	(0.109)	(3.581)	(8.531)
stage 1: Ψ[graded rtw]	0.270***			
	(0.027)			
(d) 30 clients or more per caseworker (N=11,343)				
Graded rtw	0.145	0.041	1.243	5.922
	(0.121)	(0.108)	(3.626)	(8.469)
stage 1: Ψ[graded rtw]	0.268***			
	(0.029)			
(e) 35 clients or more per caseworker (N=10,810)				
Graded rtw	0.188	0.054	2.757	7.682
	(0.124)	(0.110)	(3.683)	(8.734)
stage 1: Ψ[graded rtw]	0.271***			
	(0.030)			

Only graded rtw spells started in the first half year are considered
Contains results of IV regressions
Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 19: Overall results (1-26 weeks) using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
(a) 15 clients or more per caseworker (N=12,534)				
Graded rtw	0.344***	0.0736	7.605**	16.02**
	(0.115)	(0.100)	(3.596)	(8.113)
stage 1: Ψ[graded rtw]	0.386***			
	(0.034)			
(b) 20 clients or more per caseworker (N=12,258)				
Graded rtw	0.348***	0.061	7.331**	15.56*
	(0.111)	(0.103)	(3.355)	(7.945)
stage 1: Ψ[graded rtw]	0.386***			
	(0.035)			
(c) 25 clients or more per caseworker (N=11,741)				
Graded rtw	0.380***	0.070	8.901**	18.30**
	(0.125)	(0.104)	(3.759)	(8.624)
stage 1: Ψ[graded rtw]	0.268***			
	(0.027)			
(d) 30 clients or more per caseworker (N=11,343)				
Graded rtw	0.337***	0.031	7.803**	14.84*
	(0.119)	(0.103)	(3.699)	(8.227)
stage 1: Ψ[graded rtw]	0.268***			
	(0.028)			
(e) 35 clients or more per caseworker (N=10,810)				
Graded rtw	0.335***	0.0272	8.125**	13.99
	(0.123)	(0.109)	(3.745)	(8.520)
stage 1: Ψ[graded rtw]	0.244***			
	(0.026)			

Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 20: Weeks waited results using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
(a) 15 clients or more per caseworker (N=6,672)				
Weeks waited	-0.046***	-0.005	-1.397***	-2.234***
	(0.008)	(0.004)	(0.200)	(0.435)
stage 1: Ψ[weeks waited]	-6.751***			
	(0.995)			
(b) 20 clients or more per caseworker (N=6,436)				
Weeks waited	-0.0428***	-0.003	-1.402***	-2.066***
	(0.008)	(0.004)	(0.218)	(0.417)
stage 1: Ψ[weeks waited]	-5.064***			
	(0.826)			
(c) 25 clients or more per caseworker (N=5,906)				
Weeks waited	-0.044***	-0.001	-1.497***	-2.177***
	(0.010)	(0.005)	(0.257)	(0.489)
stage 1: Ψ[weeks waited]	-3.935***			
	(0.791)			
(d) 30 clients or more per caseworker (N=5,411)				
Weeks waited	-0.042***	0.001	-1.359***	-1.891***
	(0.010)	(0.005)	(0.229)	(0.454)
stage 1: Ψ[weeks waited]	-4.230***			
	(0.825)			
(e) 35 clients or more per caseworker (N=4,658)				
Weeks waited	-0.040***	0.002	-1.337***	-1.854***
	(0.011)	(0.006)	(0.245)	(0.457)
stage 1: Ψ[weeks waited]	-4.290***			
	(0.903)			

Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 21: Degree grading results using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
(a) 15 clients or more per caseworker (N=6,679)				
Degree grading	0.006***	0.002**	0.150***	0.327***
	(0.002)	(0.001)	(0.047)	(0.093)
stage 1: Ψ[degree grading]	28.650***			
	(0.481)			
(b) 20 clients or more per caseworker (N=6,443)				
Degree grading	0.006***	0.003***	0.148***	0.335***
	(0.002)	(0.001)	(0.049)	(0.099)
stage 1: Ψ[degree grading]	25.178***			
	(0.459)			
(c) 25 clients or more per caseworker (N=5,913)				
Degree grading	0.006***	0.003**	0.135***	0.318***
	(0.002)	(0.001)	(0.052)	(0.111)
stage 1: Ψ[degree grading]	25.373***			
	(0.564)			
(d) 30 clients or more per caseworker (N=5,415)				
Degree grading	0.007***	0.003**	0.183***	0.408***
	(0.003)	(0.001)	(0.055)	(0.119)
stage 1: Ψ[degree grading]	23.244***			
	(0.638)			
(e) 35 clients or more per caseworker (N=4,661)				
Degree grading	0.007***	0.002	0.203***	0.387***
	(0.003)	(0.001)	(0.065)	(0.137)
stage 1: Ψ[degree grading]	19.691***			
	(0.670)			

Clustered (case manager - year) standard errors between parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$