Journal of Health Economics 65 (2019) 189–209

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Graded return-to-work as a stepping stone to full work resumption☆

Lieke Kools[a,b,*], Pierre Koning[a,c,d,e]

[a] Universiteit Leiden, P.O. Box 9500, 2300 RA Leiden, The Netherlands
[b] Netspar, The Netherlands
[c] Vrije Universiteit Amsterdam, The Netherlands
[d] Tinbergen Institute, The Netherlands
[e] IZA Germany

ARTICLE INFO

Article history:
Received 9 April 2018
Received in revised form 19 March 2019
Accepted 29 March 2019
Available online 6 April 2019

JEL classification:
I18
C26

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

ABSTRACT

While there is increasing evidence that graded return-to-work is an effective tool for the rehabilitation of sick-listed workers, little is known on the optimal timing and level of grading in return-to-work trajectories. We use administrative data from a Dutch private workplace reintegration provider to fill this gap. 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. We find that graded return-to-work has the potential to speed up the recovery process, but does not necessarily help rehabilitate workers who would otherwise have not rehabilitated. Work resumption can be achieved faster when graded return-to-work is started earlier and may permanently increase when started at a higher rate of work resumption. These findings however do not hold for individuals who have problems related to mental health.

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 For example, in England sick notes have been replaced by a statement of fitness for work in 2010 (Wainwright et al., 2011), in Sweden general practitioners are recommended to subscribe part-time sick leave rather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 Norway sick-listed employees are since 2004 required to work partially after 8 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.

Research shows almost unanimously positive effects of graded work on work rehabilitation,[1] whereas interventions like vocational rehabilitation and regular paramedical care rather seem to lengthen sick spells (Markussen and Røed, 2014; Rehwald et al., 2018). 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,

☆ The authors gratefully acknowledge the financial support from Instituut Gak for this research project. They also thank the referees to this paper, 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.

* Corresponding author.
E-mail addresses: l.kools@law.leidenuniv.nl (L. Kools), p.w.c.koning@law.leidenuniv.nl (P. Koning).

[1] 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. (2018), 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 also points at positive results, see e.g. Krause et al. (1998).

https://doi.org/10.1016/j.jhealeco.2019.03.009
0167-6296/

hampering the recovery process. In light of these considerations, it is important to understand what separates an effective graded return-to-work trajectory from an ineffective one.

In this paper we analyze how the set-up of a graded return-to-work trajectory determines its effectiveness. More specifically, we analyze if work resumption rates change when the trajectory is started at a higher initial rate of work resumption or if it is starter later. 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 workplace 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 graded work (i.e., the '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. (2018).[2] 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. All the individual information on new sick-listed workers that is available to the case managers at the moment of intake is included in our data. This means that any selection on observables can be controlled for.

Our analysis extends on earlier studies in this field of research by also using a propensity measure for the initial degree of grading. 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 have not started one yet. We next construct a propensity score measure for the initial graded work percentage that is applied. This then enables us to evaluate the effects of differences in the degree of grading interventions on work resumption. As both propensities for the frequency and the degree of grading are correlated, we also estimate a model for work resumption with both the grading tendency and the degree of grading as explanatories. As such, we can analyze whether different dimensions of grading are interrelated and isolate their effects on work resumption. In addition, we 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 are plausibly exogenous, we cannot be sure that they are uncorrelated with other case manager characteristics affecting the likelihood of work resumption. We therefore conduct sensitivity tests with proxies for case manager quality as additional controls, including past work resumption rates of other sick-listed workers that were assigned to the case manager and work resumption rates of sick-listed workers that are out of sample.

We find that graded return-to-work has the potential to speed up the recovery process, but does not necessarily help rehabilitate individuals who would otherwise have not rehabilitated. Graded work that is initiated in the first 26 weeks of absence yields an increase of 18 weeks in the number of weeks worked during the first two years after sick-listing, but has no significant effects on the probability to return to work within two years. 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 2 years with four percentage point. This suggests that a sufficiently high degree of grading is crucial to realize work resumption levels that are permanent. Finally, the positive effects of graded return-to-work are especially strong and persistent for individuals who have general medical conditions. For individuals with problems related to mental health we find no significant overall effects of graded return-to-work.

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% of the insured workers (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 that focused on the sickness period that precedes DI claims. 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 this reform, disability insurance inflow was estimated to reduce by about 40% (van Sonsbeek and Gradus, 2013). This positive effect can partly be attributed to improved screening, making it more difficult to use DI as an alternative exit route 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 largely privately organized sickness and disability system (Koning, 2017). This particularly holds for the 2 year waiting period that precedes DI claims. In this period, the employer is obliged to continue payments of at least 70% of the employees regular salary.[3] In practice, most Collective Labor Agreements stipulate full wage payments in the first year and 70% in the second year. 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 Dutch: *Wet verbetering Poortwachter*) gives directions as to what these efforts should entail.[4] Employers can insure themselves against the risk of the continued wage payments during the

[2] 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).

[3] 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., 2018).

[4] For a detailed description of the concrete steps that need to be taken under the Gatekeeper protocol, we refer to de Jong et al. (2011).

waiting period via private insurers or even opt for 'broad insurance' that includes all the costs and activities that come with the obligations of the Gatekeeper protocol. Approximately 76% of Dutch employers has insurance for the risk of continued sick payments and at least 67% has such a broad insurance (de Jong et al., 2014). This predominantly includes smaller employers.

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 an extra pair of hands, or do graded work. In the latter case, the employee engages in productive work, the employer pays for those productive hours worked, and the insurer only pays for 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. For employers, sickness absence may be costly for other reasons than wage continuation, noncooperation may lead to an extension of the waiting period, and potential DI benefit costs after the waiting period are experience rated. Moreover, for sick-listed employees, non-cooperation with reintegration plans inhibits the risk of getting fired or loosing eligibility to DI benefits.[5]

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 (de Jong et al., 2014). 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 is directed to our workplace reintegration provider after 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. 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. This means case managers are not specialized in specific health problems, sectors, or regions.[6] It is important to stress that 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.

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. All individuals are followed for two years after the first sick day (or until they recover if that is within two years). For the last individual in our sample the sick spell starts in the week of October 20, 2014. The last week in our sample is the week of October 17, 2016. We exclude those individuals that hold specific insurance contracts with extra services before case management and/or earlier entry into case management (when there is not yet a risk of long term sickness). These excluded contracts are predominantly held by self-employed. Provided that we use the case managers' grading practice as a means of identification we need sufficient observations per case manager. We therefore exclude those clients that were assigned to caseworkers with less than 25 clients in a particular year.[7] 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 both 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[8] ; 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.[9] 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 Fig. 1 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 in the first year, respectively. By construction, individuals participating in graded return-to-work have a lower probability to recover in the first weeks of illness. Still,

[5] 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.

[6] The workplace reintegration provider has only one office, located in the center of the country. Contact with the sick-listed employee is mostly maintained via phone and email. Also, the workplace reintegration provider does not specialize in a certain sector. When we compare sectoral information from the data source to the sectoral information from Statistics Netherlands, we find that employees from non-profit service sectors, like care and education, are underrepresented in the data. Contrasting to for-profit firms, organizations in these sectors typically have specialized reintegration providers. For the remaining group of private for-profit organizations, sector shares of the employers of sick-listed workers in our sample are in line with those for the Netherlands.

[7] Table A1 of Appendix A shows the selection of our data in more detail.

[8] Fig. A1 in Appendix A 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 take this into account in our empirical analysis.

[9] Information on the education level of individuals was not recorded. If necessary, case managers obtained this information during the intake of new clients instead. In the empirical literature, information on education levels of sick-listed workers is usually derived from public registered data or survey data. Recording this information by private organizations is however not common in the Netherlands. Still, we argue that wage earnings can be a good proxy for the education level of workers.

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 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
– 2–9 employees	36.3%	37.5%	35.5%	0.031
– 10–49 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[b]	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[c]	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 1 year	59.6%	53.6%	63.3%	0.000
Returns to work within 2 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] When it comes to general medical conditions one must think of individuals who are recovering from surgery or suffer from chronic illness.
[c] Other contains conditions such as flu and complaints due to pregnancy.

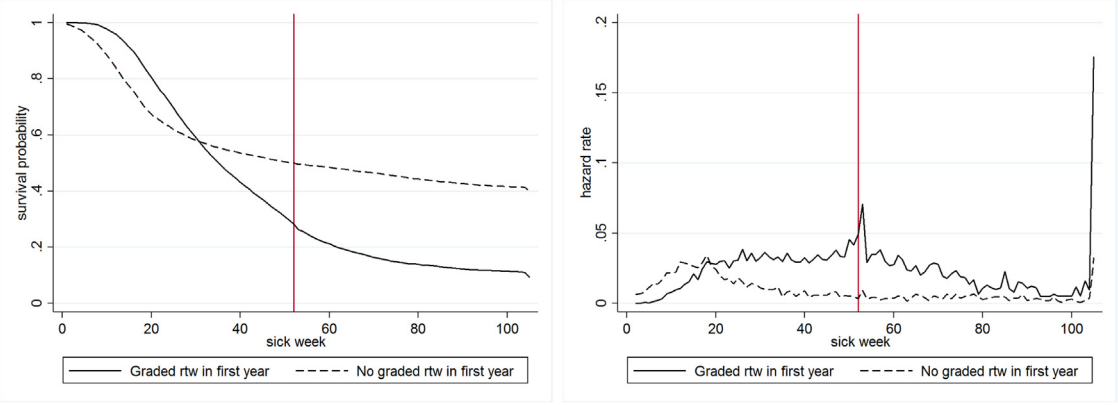

Fig. 1. Survival and hazard rates for individuals with and without graded return-to-work in first year of absence.

the group starts 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. The hazard rate spikes after the first year of sick-leave and at the end of the second year, which mirrors 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. As noted earlier, we excluded those case manager-years in which a case manager treated less than 25 individuals.[10]

[10] In Appendix A, we present the results of robustness analyses that take different cutoffs (see Tables A7–A9). 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. When we look into all the case managers in the original sample, 24% is never dropped, 43% is dropped only once, and for 29% of case managers all observations

Table 2
Descriptive statistics of the 68 case managers.[a]

	Mean	sd	Min	Max
(a) Characteristics of case manager				
Female	70.6%			
Age on 1 November 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[b]	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[c]	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] See Table A2 in Appendix A for descriptives of the working hours and the education level of case managers.
[b] When it comes to general medical conditions one must think of individuals who are recovering from surgery or suffer from chronic illness.
[c] 'Other' contains conditions such as flu and complaints due to pregnancy.

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 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. We will return to this issue in Section 5.4.

Case managers differ substantially in their use of graded return-to-work, with some 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. Apart from using graded return-to-work, case managers can also apply other interventions to their clients. Case managers documented these interventions in a separate file that can be linked to the individual clients. Unfortunately, this 'intervention file' turned out to be incomplete, so that it contained far fewer interventions than the reintegration provider would have expected, especially in the years 2011 and 2012. We therefore chose to abstain from using this file for the analysis. However, assuming that the missing interventions are missing at random, we conclude that clients who do not enter graded return-to-work are not more often enrolled in other interventions. Most of the interventions consist of advisory meetings with workers or psychological and physical care. Vocational training is not included in the options for interventions.

3.3. Setup of graded return-to-work trajectories

Within the group of clients that started a graded work trajectory, relevant outcome 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, contains 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.

are dropped. The case managers that are dropped are predominantly those with lower education who work fewer hours or who are active in the year 2011.

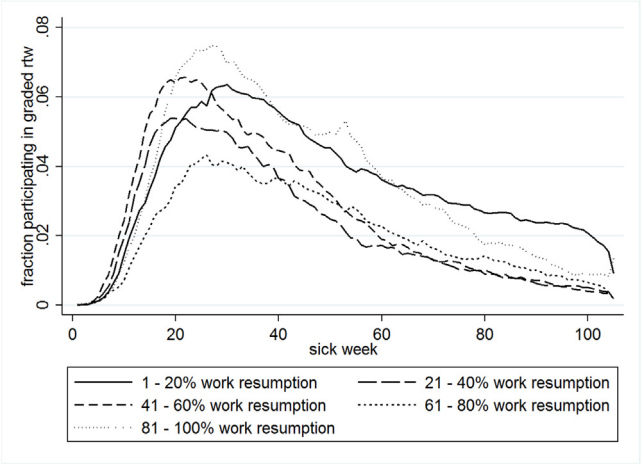

Fig. 2. Percentage of individuals participating in graded return-to-work per week.

Fig. 2 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% 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.

4. Estimation strategy

4.1. IV assumptions

To identify the effectiveness of graded return-to-work at increasing full work resumption, 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 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., 2018; Markussen et al., 2018). 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 sent to the reintegration provider after some weeks of absence. The provider assigns them to a case manager, who has substantial discretionary room in choosing the specific treatment for the client. Case managers are encouraged to use graded return-to-work whenever possible. However, the actual grading practice may vary among the case managers. First, case managers may make different assessments of when an individual is ready to start graded return-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 (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.

In the current analysis, the validity of instrumental variables estimation essentially requires four conditions to be met. First, someone's probability of graded work should be affected by the concerning case managers' propensity to use a graded work for the other clients assigned to him. With a time span of four years that is covered, assuming the tendency to use graded work to be constant over time may be too restrictive. We therefore construct propensities by case manager for each year in our sample. Second, we assume that sick workers are assigned randomly to case managers. This implies that sick-listed individuals with long and short expected sick durations do not cluster among certain case managers. We can test this assumption by excluding client characteristics or by excluding case managers who have abnormal client group compositions.[13] 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'). We will address this issue in sensitivity tests with proxies of overall case manager quality as additional controls. Finally, we assume that propensities affect the probability of graded work equally across types of individuals – the 'monotonicity' assumption.[14]

4.2. Specification of the effect of graded work

When specifying the IV model for 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. (2018). In these analyses, the aim is to estimate the effect

[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.

[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] Obviously, testing for clustering on unobservable characteristics is more complex, but it should be stressed that case managers did not receive more information than the registered data we have. This renders it plausible there was no selection on unobservables.

[14] To analyze the stringency of the monotonicity assumption, we have compared first-stage IV coefficients across diagnosis subgroups (see Table A10 of Appendix B for the results). For most diagnoses these are comparable, but not for all. When clustering sick types to broader subgroups that we will use in Section 5.3, however, we get first-stage results that are very similar – see panel (b). It thus appears that the differences in first-stage estimates for the detailed subcategories largely stem from small group size.

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	26.4%	8.5%	0.0%	47.4%
21–40% of pre-sickness wage	34.6%	7.3%	7.1%	60.0%
41–60% of pre-sickness wage	31.3%	8.7%	17.9%	78.6%
61–80% of pre-sickness wage	4.0%	2.9%	0.0%	15.6%
81–100% of pre-sickness wage	3.7%	2.9%	0.0%	14.3%

of the provision of graded work (G). As we will show later on, we extend their analysis by developing case manager propensities for the initial degree of grading (S). For ease of exposition, we consider a single year for which we construct case manager propensities. This setup can be adapted by allowing for case manager propensities for each year.

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}):

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

where we cluster standard errors across 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.

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

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. (2018), $\hat{u}_{ij}^g$ 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. 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:

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

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

where we cluster standard errors across all combinations of case managers and years in our sample. As with any IV model, it is important to stress that our parameter of interest in the above equation, γ^g, should be interpreted as a local average treatment effect (LATE). This parameter denotes the effect of increases in the likelihood of grading for only those individuals that are susceptive to grading practices of case managers. This result does not necessarily extrapolate to all individuals in our sample.

4.3. Specification of the effect of the degree of grading

Similar to the propensities for the frequency of graded work, we construct case manager propensities that can be used to explain the effect of the degree of grading (as a share of the maximum work hours) on the likelihood of work resumption. 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 those individuals that are not graded, we set this percentage equal to zero. We estimate a regression on this outcome variable that largely corresponds to Eq. (1):

$$\begin{aligned} &S_{ij} = \mathbf{x}_i'\boldsymbol{\theta}^s + \delta_1^s \;\; d_i + \delta_2^s \;\; date_i + u_{ij}^s, \\ &i = 1, \ldots n \,(\text{individuals}) \\ &j = 1, \ldots J \,(\text{case} \quad \text{managers}), \end{aligned} \tag{6}$$

Table 4
Effects of graded return-to-work (rtw) 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 return-to-work	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 return-to-work	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_i^g (propensity graded rtw)	0.270***				0.268***			
	(0.0268)				(0.0267)			
(c) Reduced form estimates of propensity								
ψ_i^g (propensity 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 in parentheses.
*/**/*** indicate p-levels of 10%, 5% and 1%, respectively.

Using the error estimates of the above equation, we calculate propensities as in Eq. (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, together with $\mathbf{x}$ as control variables:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^s + \gamma^s \hat{S}_i + \epsilon_i^s, \tag{7}$$

$$S_i = \mathbf{x}_i'\boldsymbol{\pi}^s + \alpha^s \psi_i^s + \eta_i^s. \tag{8}$$

where we allow for clustering effects across combinations of case manager and years.

As the initial degree of grading is set equal to zero for those individuals that are not graded, it is likely that the proxies for the frequency and the degree of grading are correlated. For a broader interpretation of our results, we therefore also specify a model that allows for both the frequency and the level of grading as endogenous regressors. This corresponds to e.g. Markussen and Røed (2014), who analyze the effectiveness of different treatment strategies on return-to-work rates of sick-listed workers. In our case, this implies the following IV model with two endogenous regressors and two propensities as instruments and with the additional coefficients $\alpha^{g,s}$ and $\alpha^{s,g}$:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^s + \gamma^g \hat{G}_i + \gamma^s \hat{S}_i \epsilon_i^s \tag{9}$$

$$G_i = \mathbf{x}_i'\boldsymbol{\pi}^g + \alpha^g \psi_i^g + \alpha^{g,s} \psi_i^s + \eta_i^g \tag{10}$$

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

where we allow for clustering effects across combinations of case managers and years.

5. Results

5.1. The overall effect 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 2 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 all estimated coefficients of the regressions are shown in Tables A3 and A4 of Appendix B.

Columns 1–4 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.[15] Based on the OLS results, one would conclude that graded return-to-work trajectories have substantial and positive 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 graded return-to-work are only 2 percentage point more likely to rehabilitate within two years than those assigned to the case manager with the lowest propensity to use graded work.

Columns 5–8 show the results when only considering graded return-to-work trajectories which started in the first 26 weeks of sick leave as a treatment, with individuals who started graded work after the 26 weeks considered as untreated. Compared to the earlier results with 52 weeks as a maximum, there are noticeable differ-

[15] 5.3% of untreated individuals 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 0.049 (0.112), the number of weeks worked increases by 1.728 (8.680).

Table 5
Effects of the starting degree of grading on full work resumption.

	Returned to work		Weeks worked in	
	1 year	2 years	Week 1–52	Week 1–104
(a) OLS estimates				
Starting level (0–100)	0.004***	0.005***	0.054***	0.306***
	(0.000)	(0.000)	(0.005)	(0.014)
N	11,741	11,741	11,741	11,741
R^2	0.203	0.135	0.303	0.244
(b) IV estimates				
Starting level (0–100)	0.007***	0.005***	0.132***	0.421***
	(0.002)	(0.001)	(0.049)	(0.113)
N	11,741	11,741	11,741	11,741
R^2	0.187	0.134	0.289	0.238
Stage 1: Ψ_i^s (propensity starting level)	22.50***			
	(0.613)			
(c) Reduced form estimates of propensity				
Ψ_i^s (propensity starting level)	0.154***	0.111***	2.971***	9.463***
	(0.036)	(0.033)	(1.101)	(2.571)
N	11,741	11,741	11,741	11,741
R^2	0.169	0.070	0.297	0.203

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.
*$p<0.1$.
**$p<0.05$
*** indicates significance at 1%.

ences in the effects. The probability to return to work increases with 38.0 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 may be that graded return-to-work trajectories are more effective when started earlier anyway, but there may also be lock-in effects for trajectories that occur in the first weeks of grading. If the latter holds, 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 columns 2 and 6.[16]

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. (2018) find substantially bigger results, amounting to a 30 week increase in weeks in regular employment in the first year.[17] Contrary to our results, both Markussen et al. (2012) and Rehwald et al. (2018) find positive long run effects. The first shows that employment two years after sick listing increases with 16–21 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. As pointed out in Section 3, 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 speeds up the recovery process, particularly when starting early.

5.2. The effect of the starting degree of grading

We next investigate whether graded-work work trajectories should be implemented at a high or low degree of grading. To this end, Table 5 shows the main estimation results of the effect of the initial degree of grading on work resumption and weeks worked for trajectories starting in the first year of sick-listing. Table A5 in Appendix B presents all underlying coefficient estimates.

Table 5 shows that 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 about 22 percentage point. This effect roughly corresponds to an increase of one day per working week. From the second stage estimates we infer that a 10 percentage point higher level of grading results in a 7 percentage point higher chance of recovering in the first year and of 5 percentage point after 2 years.[18] The number of weeks worked in the first year increases by 1.3, whereas individuals work 4.1 additional weeks in the first two years. These findings suggest that a higher initial level of grading improves both short-term and long-term recovery rates. It appears that the potentially 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.

Table 6 presents the estimation results for models that include both endogenous regressors and both grading propensities as instruments. For ease of interpretation, we re-specify the propen-

[16] To illustrate the evolution of the effects in more detail, Fig. A4 in Appendix B shows the effects of graded return-to-work trajectories that started in the first half year on the return to work probability and on 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 2 years. In line with this, the steep increase in weeks worked between weeks 40 and 60 does not persist, such that the line flattens out towards the end of the second year.

[17] 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. (2018), graded return-to-work should be started within four weeks after a meeting which is held in the first eight weeks.

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

Table 6
Effects of the frequency and the starting degree of graded return-to-work on full work resumption.

	Returned to work		Weeks worked in	
	1 year	2 years	Week 1–52	Week 1–104
(a) OLS estimates				
Graded return-to-work	0.081***	0.288***	−3.793***	8.824***
	(0.016)	(0.013)	(0.409)	(1.006)
Starting level (0–100)	0.003***	0.0004	0.112***	0.171***
	(0.000)	(0.000)	(0.008)	(0.019)
N	11,741	11,741	11,741	11,741
R^2	0.206	0.181	0.309	0.250
(b) IV estimates				
Graded return-to-work	−0.171	−0.162	−6.274	−13.24
	(0.202)	(0.188)	(5.800)	(14.14)
Starting level (0-100)	0.010***	0.008***	0.254***	0.677***
	(0.003)	(0.003)	(0.096)	(0.233)
N	11,741	11,741	11,741	11,741
R^2	0.154	0.065	0.282	0.195
Stage 1 estimates: graded return-to-work				
Ψ_i^g (propensity graded rtw)	0.269*** (0.042)			
Ψ_i^{so} (propensity starting level – orthogonal)	0.448*** (0.053)			
Stage 1 estimates: starting level				
Ψ_i^g (propensity graded rtw)	7.825*** (1.848)			
Ψ_i^{so} (propensity starting level – orthogonal)	35.74*** (1.955)			
(c) Reduced form estimates of propensities				
Ψ_i^g (propensity graded rtw)	0.034 (0.036)	0.020 (0.033)	0.299 (0.983)	1.741 (2.498)
Ψ_i^{so} (propensity starting level – orthogonal)	0.286*** (0.052)	0.216*** (0.050)	6.249*** (1.565)	18.26*** (3.848)
N	11,741	11,741	11,741	11,741
R^2	0.172	0.072	0.299	0.206

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.
Ψ_i^g and Ψ_i^s are highly correlated ($\rho = 0.843$). Therefore, we use the normalized value of $\Psi_i^{so} = \Psi_i^s - \beta\Psi_i^g$ as an instrument rather than Ψ_i^s, with β the estimated OLS coefficient of Ψ_i^s on Ψ_i^g.
Clustered (case manager - year) standard errors between parentheses.
**$p<0.05$.
*$p<0.1$.
*** indicates significance at 1%.

sity of the initial degree of grading as orthogonal to the frequency propensity and again normalize the difference to values between zero and one.[19] The first stage coefficients of all propensities in the table are significant. Moreover, both the Kleibergen–Paap F-test and the Sanderson–Windmeijer F-test on weak instruments are large and have P-values equal to zero.[20] So even though the propensities for the frequency and level of grading are correlated, these tests suggest that both the coefficients of the effect the frequency and the level of graded work are identified from the data.

Turning to the second stage estimates in Table 6, our findings are generally in line with those obtained from models with single endogenous regressors. The effects of using of graded work remain insignificant and become smaller for all outcome measures, while the effects of the initial degree of grading remain significant and become more sizable. Based on our estimates, the graded work trajectories should be at least have starting levels between 15% and 25% to improve the likelihood to work resumption or to increase the number of weeks that are worked. The effects on return to work rates also indicate that the effects of graded work persist beyond the absence period of two years.

5.3. Effects for different types of medical conditions

Table 7 presents 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. Panel (c) shows the effects of the initial degree of grading.[21] The first stage estimation results are all similar in size across medical conditions, suggesting that the extent to which case managers can affect the use and degree of graded work is equal across groups. 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. A possible explanation for this is that psychological problems are more often related to the specific work environment, so that graded return-to-work implies returning to the source of the problems rather than

[19] That is, we regress the propensity for the initial degree of grading on the propensity of the frequency of grading and use the (normalized) error terms of this regression as a second propensity. Accordingly, the first stage and reduced form estimates can be interpreted as the isolated effect of the propensities.
[20] The Kleibergen–Paap F-test for two instruments and two endogenous regressors equals 31.2, whereas the F-tests on the excluded instruments by Sanderson and Windmeijer equal 70.0 and 90.2.
[21] The results for weeks worked correspond to the results for the return-to-work probability and can be found in Table A6 of Appendix B.

Table 7
IV estimation results of graded return-to-work and starting degree of grading on full work resumption for different medical conditions.

	General medical		Musculo-skeletal		Mental	
	Returned to work 1 year	2 years	Returned to work 1 year	2 years	Returned to work 1 year	2 years
(a) Overall effect: trajectories started in week 1–52						
Graded return-to-work	0.572*	0.563**	0.477	−0.203	−0.023	−0.108
	(0.327)	(0.244)	(0.540)	(0.413)	(0.352)	(0.373)
Stage 1: Ψ_i^g (propensity graded rtw)	0.191***		0.155		0.170**	
	(0.072)		(0.095)		(0.074)	
(b) Overall effect: trajectories started in week 1–26						
Graded return-to-work	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_i^g (propensity graded rtw)	0.281***		0.229***		0.266***	
	(0.066)		(0.076)		(0.079)	
(c) Initial degree of grading						
Starting level (0–100)	0.008***	0.007***	0.009***	0.005***	0.006***	0.004**
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Stage 1: Ψ_i^s (propensity starting level)	34.68***		31.27***		33.40***	
	(1.436)		(2.481)		(2.581)	

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.
The results are based on 3971 observations with general medical conditions, 1947 with musculo-skeletal conditions, and 3380 with conditions related to mental health.
Clustered (case manager–year) standard errors between parentheses.
*/**/*** indicate significance levels of 10/5/1%, respectively.

offering a solution to the problem.[22] 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 indicates that graded work can be meaningful for individuals with chronic illnesses or individuals that recover from medical treatments. Finally, we find that starting at a higher initial level of graded work resumption results in higher probabilities to return to work for all medical conditions and with comparable coefficient magnitudes.

5.4. Sensitivity tests: specialization and case manager quality

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 results 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. To analyze the robustness of our findings to this assumption, we run a set of sensitivity analyses which are reported in Table A11 of Appendix B. First, we re-run the regressions while excluding specific sets of covariates. That is, we exclude sick types in column (2), sick weeks until application in column (3), and all covariates expect time dummies in column (4). We next exclude case managers with abnormal client group compositions from our sample, which gives the results shown in columns (5) and (6). To this end, we define a group composition to be abnormal if the group average of the characteristics of the clients per case manager-year combination is more than three (column 4) or two (column 5) standard deviations away from the overall mean. For both tests, the overall results for both graded work and the initial level of work resumption are similar to the baseline.[23]

We also conduct sensitivity tests that consider the importance of the overall case manager's quality. For the validity of our approach, case managers' propensity to grade should not be correlated with overall case manager quality. This exclusion restriction may not hold when, for example, high quality case managers are also more likely to use graded work arrangements. We therefore extend our baseline models with two proxies for case manager quality – see columns (2) and (3) of Table A12. The first proxy is the lagged 'propensity to cure', which measures the return to work rate of a particular case manager in the previous year. The second proxy for the quality of case managers is derived from the group of sick-listed workers that already participated in graded work trajectories before entering case management and therefore were excluded from the sample. Knowing that the graded work status is given for these observations, we calculate the case managers' propensity to cure for this auxiliary sample for each case manager–year combination with at least ten observations. The general picture that emerges is that the inclusion of the two proxies of case manager quality are indeed positively correlated with work resumption, but second stage estimates are close to those for the baseline models.

[22] This mechanism may also explain why employment rates of individuals with mental conditions are usually substantially lower than those of individuals with other conditions (OECD, 2010).

[23] It is important to stress that sick type is the only variable that is determined by the case manager after the client is assigned to him. Additional regressions on propensities reveal that case managers with workers with less severe conditions more often apply graded work, and/or that case managers who tend to label someone's illness as mild also tend to find graded return-to-work more often appropriate. 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. In all cases, the sizes of the difference in point estimates do not lead to concerns about the validity of our approach.

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 resumption 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 a propensity for the starting level of graded work.

We find that graded return-to-work has the potential to speed up the recovery process, but does not does not necessarily help rehabilitate individuals who would otherwise have not rehabilitated. Graded work that is initiated in the first 26 weeks of absence yields an increase in the number of weeks worked during the first two years after sick-listing of 18 weeks, but has no significant effects on the probability to return to work within two years. These findings may stem from the fact that employers in the Netherlands are committed to facilitate return-to-work for all sick-listed workers.

Our evidence also suggests that the initial level of graded work is a crucial determinant of the success of trajectories. 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 5 percentage point. This indicates that the potential positive effects of graded work can only be established if the individual can properly participate in the work process. These effects are permanent, meaning that they affect the likelihood to resume work also after the waiting period of two years.

Appendix A. Additional data descriptives

Tables A1 and A2

Table A1
Selection steps in data of the workplace reintegration provider.

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/ goal other than back to work	−2093
Intervention/graded rtw took place before application	−5030
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	−1534-1,534
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 A2
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%

Fig. A1

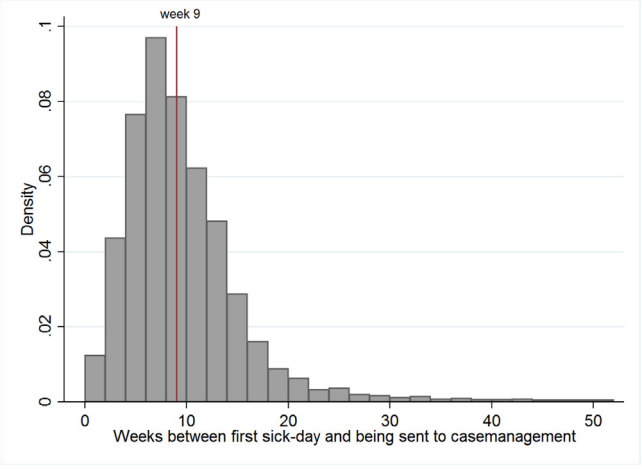

Fig. A1. Histogram of application moments of newly assigned sick-listed workers.

Appendix B. Additional estimation results

Tables A3–A12

Table A3
Effect of graded return-to-work on full work resumption 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	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_i^g (prop. graded rtw)	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	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 employees	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	−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 employees	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.
(b) Baseline category: 0–2 week.
(c) Baseline category: 2011.
Cluster robust standard errors in parentheses.
Duration baseline coefficients for stage-two estimates are shown in FigA2, propensities to treat before scaling are shown in Fig. A3.
*/**/*** indicate significance levels of 10/5/1%, respectively.

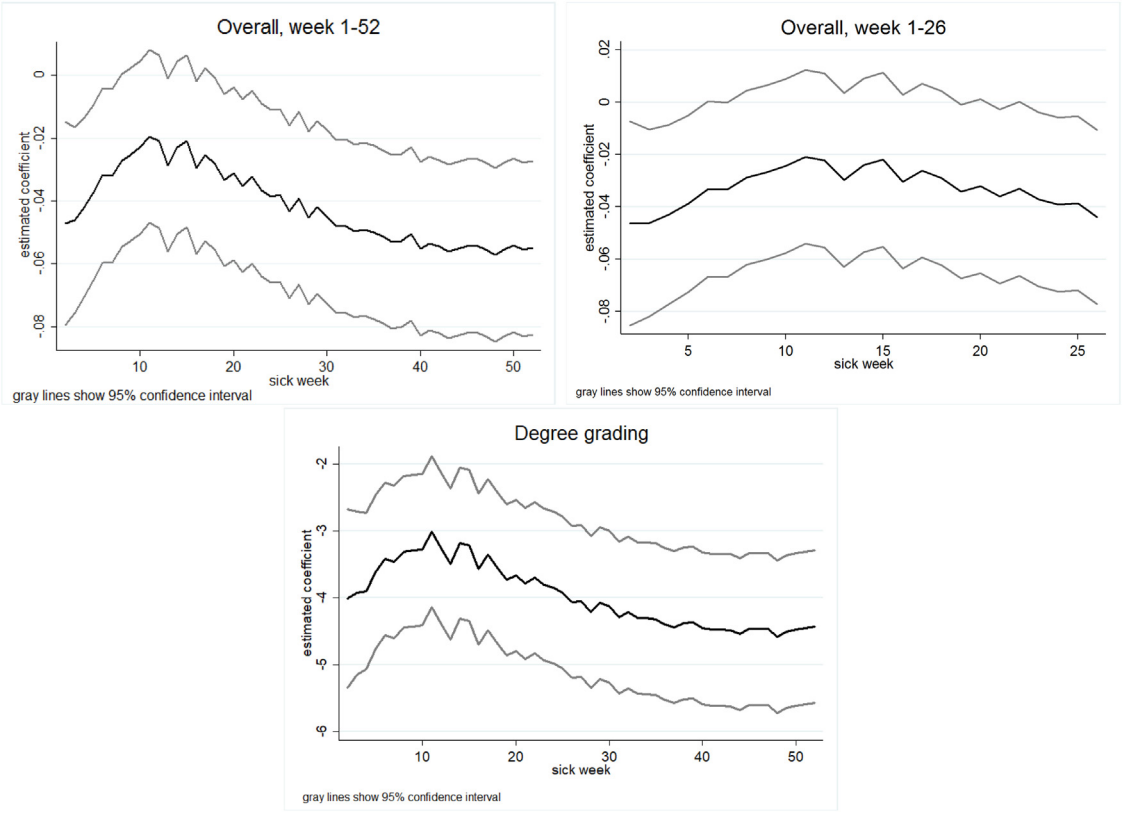

Fig. A2. Duration baseline coefficients for the effect of graded return-to-work starting in weeks 1–52 (Table A3), graded return-to-work starting in weeks 1–26 (Table A4) and the starting degree of grading (Table A5).

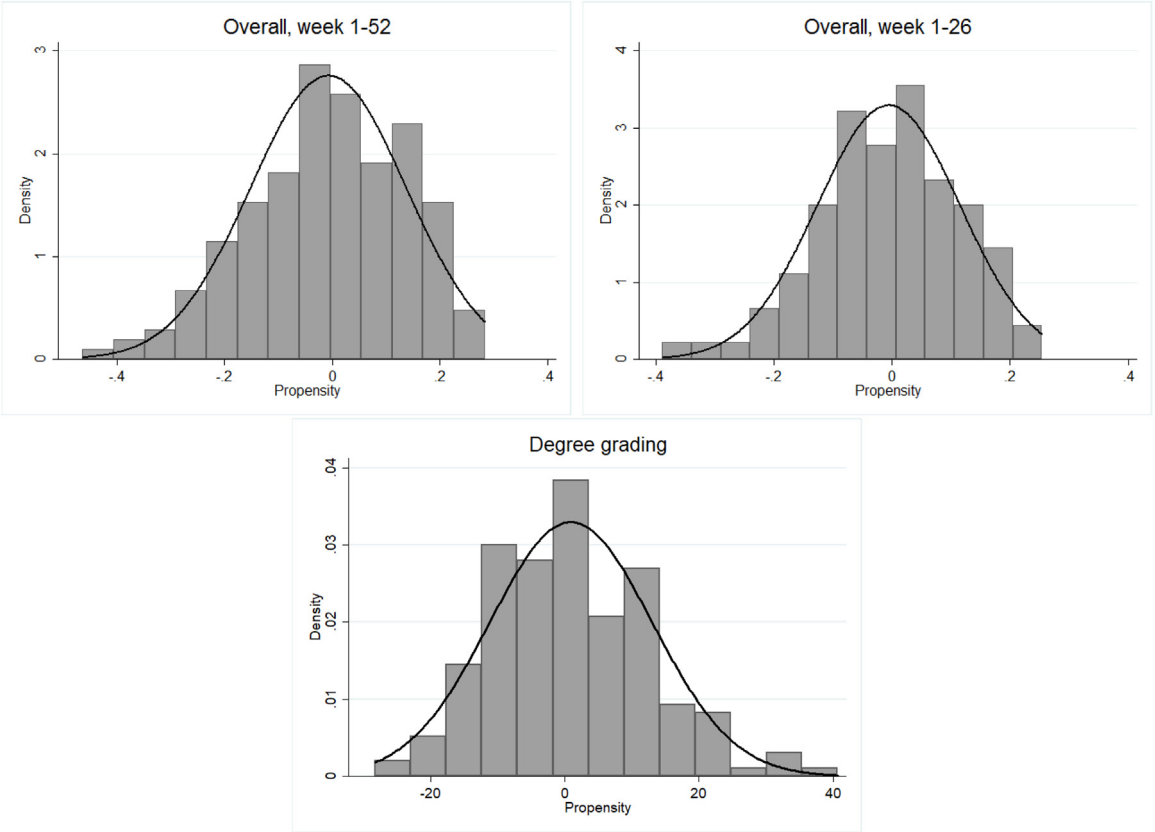

Fig. A3. Propensities to treat before scaling, obtained from models for the effect of graded return-to-work starting in weeks 1–52 (Table A3), graded return-to-work starting in weeks 1–26 (Table A4) and the starting degree of grading (Table A5).

Table A4
Effect of graded return-to-work on full work resumption 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, ankle, 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_i^g (prop. graded rtw)	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, ankle, 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, ankle, 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: Firm size		
2013	0.186***	(0.018)	Psychological – severe	−0.455***	(0.034)	2–10 employees	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.
(b) Baseline category: 0–2 week.
(c) Baseline category: 2011.
Cluster robust standard errors in parentheses.
Duration baseline coefficients for stage-two estimates are shown in Fig. A2, propensities to treat before scaling are shown in Fig. A3.
*/**/*** indicate significance levels of 10/5/1%, respectively.

Table A5
Effect of initial degree of graded return-to-work on full work resumption, including coefficients on control variables.

(a) Stage 0–dependent: starting level (0–100)								
Sex	−0.042*	(0.025)	Condition			Contract type		
Age at application	0.027***	(0.007)	General medical – medium	0.249***	(0.053)	B	0.022	(0.068)
Age at application^2	0.000***	(0.000)	General medical – severe	−0.223***	(0.053)	C	0.081	(0.072)
ln(gross wage)	0.173***	(0.039)	Neck, shoulder, arm complaints	0.111*	(0.062)	D	−0.011	(0.070)
ln(gross wage)^2	−0.012***	(0.004)	Physical – mild	0.234***	(0.061)	E	0.072	(0.069)
Sick weeks until application	0.023***	(0.006)	Physical – severe	−0.122*	(0.073)	F	0.265***	(0.096)
Sick weeks until application^2	0.000	(0.000)	Hip, ankle, knee complaints	0.339***	(0.065)	G	0.110	(0.102)
Application year			Other	−0.476***	(0.090)	H	0.099	(0.073)
2012	0.022	(0.072)	Psychiatric	−0.300***	(0.090)	I	0.165**	(0.081)
2013	0.153	(0.099)	Psychological – mild	−0.116**	(0.054)	Firm size		
2014	0.181	(0.126)	Psychological – severe	−0.280***	(0.077)	2–9 employee	0.078**	(0.035)
Constant	−0.016	(0.088)	Psychosocial – mild	0.031	(0.055)	10–49 employees	0.181***	(0.035)
			Psychosocial – severe	−0.060	(0.096)	50 or more employees	0.397***	(0.082)
			Back complaints	0.089	(0.061)	Unknown	0.049	(0.046)
Observations	290,929		Social problems	0.250***	(0.094)			
R-squared	0.009		Conflict	−0.511***	(0.062)			
(b) Stage 1–dependent: starting level (0–100)								
Ψ_i^s (prop. starting level)	22.50***	(0.613)	Condition			Contract type		
Sex	−0.586	(0.433)	General medical – medium	4.209***	(1.203)	B	0.475	(1.289)
Age at application	0.419***	(0.135)	General medical – severe	−2.103*	(1.140)	C	1.241	(1.351)
Age at application^2	−0.004***	(0.002)	Neck, shoulder, arm complaints	3.149**	(1.323)	D	−0.219	(1.275)
ln(gross wage)	2.932***	(0.781)	Physical – mild	4.389***	(1.194)	E	0.938	(1.292)
ln(gross wage)^2	−0.212***	(0.077)	Physical – severe	−1.126	(1.312)	F	4.005**	(1.684)
Sick weeks until application	−0.091	(0.082)	Hip, ankle, knee complaints	5.874***	(1.215)	G	1.287	(1.851)
Sick weeks until application^2	−0.005***	(0.002)	Other	−11.40***	(1.888)	H	1.284	(1.290)
Application year			Psychiatric	−4.759***	(1.686)	I	2.569*	(1.520)
2012	0.592*	(0.311)	Psychological - mild	−1.247	(1.068)	Firm size:		
2013	0.307	(0.347)	Psychological – severe	−3.203**	(1.470)	2–9 employee	1.577**	(0.648)
2014	−1.334***	(0.379)	Psychosocial – mild	1.221	(1.050)	10–49 employees	3.227***	(0.597)
Constant	−8.696**	(3.655)	Psychosocial – severe	1.694	(2.012)	50 or more employees	5.548***	(1.285)
			Back complaints	2.245*	(1.205)	Unknown	1.278	(0.873)
Observations	11,741		Social problems	4.109**	(1.757)			
			conflict	−12.56***	(1.319)			
(c) Stage 2–dependent: returned to work within 1 year								
Starting level (0-100)	0.007***	(0.002)	Condition			Contract type		
Sex	−0.025***	(0.009)	General medical – medium	−0.173***	(0.020)	B	0.049**	(0.024)
Age at application	−0.001	(0.003)	General medical – severe	−0.504***	(0.020)	C	0.039	(0.026)
Age at application^2	0.000	(0.000)	Neck, shoulder, arm complaints	−0.269***	(0.023)	D	−0.012	(0.027)
ln(gross wage)	0.009	(0.016)	Physical – mild	−0.109***	(0.022)	E	0.000	(0.027)
ln(gross wage)^2	−0.002	(0.002)	−Physical – severe	−0.420***	(0.029)	F	0.016	(0.034)
Sick weeks until application	−0.009***	(0.002)	Hip, ankle, knee complaints	−0.209***	(0.025)	G	0.028	(0.034)
Sick weeks until application^2	0.000	(0.000)	Other	−0.399***	(0.056)	H	−0.023	(0.026)
Application year			Psychiatric	−0.442***	(0.035)	I	0.018	(0.031)
2012	0.161***	(0.017)	Psychological - mild	−0.292***	(0.022)	Firm size		
2013	0.171***	(0.019)	Psychological – severe	−0.479***	(0.033)	2–9 employees	0.004	(0.014)
2014	0.149***	(0.019)	Psychosocial – mild	−0.157***	(0.020)	10–49 employees	0.002	(0.015)
Constant	0.688***	(0.071)	Psychosocial – severe	−0.402***	(0.035)	50 or more employees	0.031	(0.034)
			Back complaints	−0.267***	(0.023)	Unknown	0.011	(0.019)
Observations	11,741		Social problems	−0.080***	(0.025)			
R-squared	0.187		Conflict	−0.061*	(0.035)			

(a) Baseline category: general medical light.
(b) Baseline category: 0–2 week.
(c) Baseline category: 2011.
Cluster robust standard errors in parentheses.
Duration baseline coefficients for stage-two estimates are shown in Fig. A2, propensities to treat before scaling are shown in Fig. A3.
*/**/*** indicate significance levels of 10/5/1%, respectively.

Table A6
IV estimation results for different medical conditions; effects on 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) Overall effect: trajectories started in week 1–52						
Graded return-to-work	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_i^g (propensity graded rtw)	0.191***		0.155		0.1703**	
	(0.072)		(0.095)		(0.074)	
(b) Overall effect: trajectories started in week 1–26						
Graded return-to-work	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_i^g (propensity graded rtw)	0.281***		0.229***		0.266***	
	(0.066)		(0.076)		(0.079)	
(c) Initial degree of grading						
Starting level (0–100)	0.157***	0.486***	0.199***	0.485***	0.139**	0.410***
	(0.050)	(0.111)	(0.071)	(0.148)	(0.061)	(0.142)
Stage 1: Ψ_i^s (prop. starting level)	34.68***		31.27***		33.40***	
	(1.436)		(2.481)		(2.581)	

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.
The results are based on 3971 observations with general medical conditions, 1947 with musculo-skeletal conditions, and 3380 with conditions related to mental health.
Clustered (case manager–year) standard errors in parentheses.
*/**/*** indicate significance levels of 10/5/1%, respectively.

Table A7
IV estimation results for effects of graded work starting in first 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 return-to-work	0.093	0.086	−0.784	4.090
	(0.117)	(0.103)	(3.489)	(8.245)
Stage 1: ψ_i^g (propensity graded rtw)	0.385***			
	(0.033)			
(b) 20 clients or more per caseworker (N = 12,258)				
Graded return-to-work	0.129	0.079	0.487	5.821
	(0.115)	(0.109)	(3.375)	(8.189)
Stage 1: ψ_i^g (propensity graded rtw)	0.343***			
	(0.032)			
(c) 25 clients or more per caseworker (N = 11,741)				
Graded return-to-work	0.127	0.075	1.173	6.642
	(0.122)	(0.109)	(3.581)	(8.531)
Stage 1: ψ_i^g (propensity graded rtw)	0.270***			
	(0.027)			
(d) 30 clients or more per caseworker (N = 11,343)				
Graded return-to-work	0.145	0.041	1.243	5.922
	(0.121)	(0.108)	(3.626)	(8.469)
Stage 1: ψ_i^g (propensity graded rtw)	0.268***			
	(0.029)			
(e) 35 clients or more per caseworker (N = 10,810)				
Graded return-to-work	0.188	0.054	2.757	7.682
	(0.124)	(0.110)	(3.683)	(8.734)
Stage 1: ψ_i^g (propensity graded rtw)	0.271***			
	(0.030)			

Clustered (case manager–year) standard errors between parentheses.
*p<0.1.
**p<0.05
*** p < 0.01.

Table A8
IV estimation results for effects of graded work starting in first 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 return-to-work	0.344***	0.0736	7.605**	16.02**
	(0.115)	(0.100)	(3.596)	(8.113)
Stage 1: ψ_i^g (propensity graded rtw)	0.386***			
	(0.034)			
(b) 20 clients or more per caseworker ($N = 12{,}258$)				
Graded return-to-work	0.348***	0.061	7.331**	15.56*
	(0.111)	(0.103)	(3.355)	(7.945)
Stage 1: ψ_i^g (propensity graded rtw)	0.386***			
	(0.035)			
(c) 25 clients or more per caseworker ($N = 11{,}741$)				
Graded return-to-work	0.380***	0.070	8.901**	18.30**
	(0.125)	(0.104)	(3.759)	(8.624)
Stage 1: ψ_i^g (propensity graded rtw)	0.268***			
	(0.027)			
(d) 30 clients or more per caseworker ($N = 11{,}343$)				
Graded return-to-work	0.337***	0.031	7.803**	14.84*
	(0.119)	(0.103)	(3.699)	(8.227)
Stage 1: ψ_i^g (propensity graded rtw)	0.268***			
	(0.028)			
(e) 35 clients or more per caseworker ($N = 10{,}810$)				
Graded return-to-work	0.335***	0.0272	8.125**	13.99
	(0.123)	(0.109)	(3.745)	(8.520)
Stage 1: ψ_i^g (propensity graded rtw)	0.244***			
	(0.026)			

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

Table A9
IV estimation results for the effect of starting degree grading 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 ($N = 12,534$)				
Starting level (0–100)	0.007***	0.006***	0.141***	0.447***
	(0.002)	(0.001)	(0.047)	(0.107)
Stage 1: Ψ_i^s (propensity starting level)	26.86***			
	(0.805)			
(b) 20 clients or more per caseworker ($N = 12,258$)				
Starting level (0–100)	0.007***	0.005***	0.138***	0.440***
	(0.002)	(0.001)	(0.046)	(0.108)
Stage 1: Ψ_i^s (propensity starting level)	22.61***			
	(0.638)			
(c) 25 clients or more per caseworker ($N = 11,741$)				
Starting level (0–100)	0.007***	0.005***	0.132***	0.421***
	(0.002)	(0.001)	(0.049)	(0.113)
Stage 1: Ψ_i^s (propensity starting level)	22.50***			
	(0.613)			
(d) 30 clients or more per caseworker ($N = 11,343$)				
Starting level (0–100)	0.007***	0.005***	0.138***	0.428***
	(0.002)	(0.002)	(0.053)	(0.122)
Stage 1: Ψ_i^s (propensity starting level)	18.24***			
	(0.528)			
(e) 35 clients or more per caseworker ($N = 10,810$)				
Starting level (0–100)	0.008***	0.005***	0.144**	0.447***
	(0.002)	(0.002)	(0.058)	(0.134)
Stage 1: Ψ_i^s (propensity starting level)	17.60***			
	(0.561)			

Clustered (case manager–year) standard errors between parentheses.
** $p < 0.05$.
*** $p < 0.01$.

Table A10
IV first stage estimation results for graded work starting in first 52 weeks, stratified by subcategories of impairments.

Subgroup	Overall: started in		Initial degree of grading	*N*
	Week 1–52	Week 1–26		
(a) Detailed subcategories				
General medical – mild	0.174*	0.175	25.49***	907
	(0.104)	(0.109)	(6.163)	
General medical – medium	0.343***	0.334***	25.24***	1,588
	(0.067)	(0.067)	(3.142)	
General medical – severe	0.198***	0.202***	25.48***	1,350
	(0.075)	(0.075)	(4.243)	
Physical – mild	0.177*	0.120	21.16***	838
	(0.093)	(0.092)	(4.748)	
Physical – severe	0.488***	0.495***	35.83***	427
	(0.099)	(0.101)	(5.385)	
Neck, shoulder, arm complaints	0.179	0.228*	23.61***	810
	(0.146)	(0.127)	(7.183)	
Hip, ankle, knee complaints	0.318***	0.391***	23.89***	743
	(0.106)	(0.105)	(4.801)	
Back complaints	0.081	0.036	25.89**	860
	(0.232)	(0.222)	(10.44)	
Psychiatric	0.043	0.082	−1.336	210
	(0.196)	(0.195)	(7.142)	
Psychological – mild	0.140*	0.172**	11.99***	1,338
	(0.077)	(0.085)	(3.349)	
Psychological – severe	0.042	−0.116	8.468	328
	(0.173)	(0.162)	(6.821)	
Psychosocial – mild	0.370***	0.288***	23.67***	1254
	(0.084)	(0.098)	(3.415)	
Psychosocial – severe	0.445***	−0.001	22.72**	209
	(0.156)	(0.169)	(9.219)	
Social problems	0.423***	0.481***	23.71***	244
	(0.096)	(0.101)	(5.001)	
Conflict	0.228	0.440**	30.05***	464
	(0.197)	(0.187)	(10.28)	
Other[a]	0.269**	0.247**	18.28***	171
	(0.110)	(0.106)	(6.950)	
F-test on equality of coefficients				
F(15, 181)	1.98	1.94	2.93	
p-Value	0.0190	0.0220	0.0004	
(b) Rough subcategories				
General medical	0.261***	0.176	25.55***	3,845
	(0.049)	(0.108)	(2.003)	
Musculo-skeletal	0.298***	0.333***	22.91***	2,413
	(0.060)	(0.067)	(2.391)	
Mental	0.246***	0.207***	18.06***	3,373
	(0.048)	(0.075)	(1.984)	
F-test on equality of coefficients				
F(2, 181)	0.21	1.15	2.98	
p-Value	0.8094	0.3186	0.0531	

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

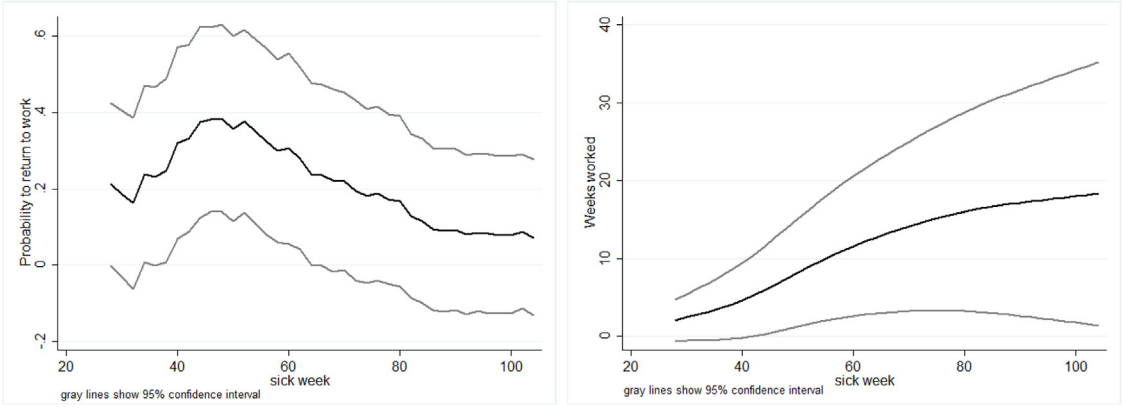

Fig. A4. Cumulative effects of graded return-to-work starting in weeks 1–52 on full work resumption and weeks worked, estimated by elapsed sick weeks.

Table A11
Sensitivity tests for specialization effects for return-to-work within one year.

Dependent: Return-to-work within 1 year	(1) Baseline	(2)	(3)	(4) Exclude covariates	(5)	(6) Exclude abnormal groups
		Sick type	Weeks until application	All expect year dummies	>3 sd from mean	>2 sd from mean
(a) Overall effect: trajectories started in week 1–52						
Graded return-to-work	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_i^g (propensity 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 return-to-work	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_i^g (propensity 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) Initial degree of grading						
Starting level (0-100)	0.007***	0.007***	0.007***	0.005***	0.008***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.003)
Stage 1: Ψ_i^s (propensity starting level)	22.50***	21.70***	23.07***	21.62***	21.76***	20.15***
	(0.613)	(0.568)	(0.688)	(0.658)	(0.687)	(1.058)

Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant.
The results are based upon 11,741 observations of which 8464 remain in column (5) and 3807 in column (6).
Clustered (case manager–year) standard errors between parentheses.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table A12
Sensitivity tests to the inclusion of proxies for case manager quality, with full return-to-work within one year as dependent variable.

Dep.t: Return-to-work within 1 year	(1) Baseline	(2) Propensity to cure	(3)
		Lagged	Graded at start
(a) Overall effect: trajectories started in week 1–52			
Graded return-to-work	0.127	0.266*	0.080
	(0.122)	(0.143)	(0.133)
Ψ_i^c (propensity to cure)		0.202***	0.143***
		(0.071)	(0.038)
Stage 1: Ψ_i^g (propensity graded return-to-work)	0.270***	0.267***	0.268***
	(0.027)	(0.032)	(0.027)
Stage 1: Ψ_i^c (propensity to cure)		−0.091**	−0.008
		(0.037)	(0.023)
(b) Overall effect: trajectories started in week 1–26			
Graded return-to-work	0.380***	0.396***	0.323**
	(0.125)	(0.123)	(0.133)
Ψ_i^c (propensity to cure)		0.188***	0.137***
		(0.060)	(0.035)
Stage 1: Ψ_i^g (propensity graded return-to-work)	0.268***	0.284***	0.257***
	(0.027)	(0.030)	(0.031)
Stage 1: Ψ_i^c (propensity to cure)		−0.036	0.004
		(0.034)	(0.021)
(c) Initial degree of grading			
Starting level (0–100)	0.007***	0.008***	0.006***
	(0.002)	(0.002)	(0.002)
Ψ_i^c (propensity to cure)		0.184***	0.125***
		(0.062)	(0.038)
Stage 1: Ψ_i^s (propensity starting level)	22.50***	21.84***	21.71***
	(0.613)	(0.770)	(0.635)
Stage 1: Ψ_i^c (propensity to cure)		0.422	0.368
		(1.272)	(0.659)

Claimants are excluded when their assigned case managers treated fewer than 25 claimants that year.
The results are based upon 11,741 observations of which 8319 remain in column (3) and 10,244 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) 0.008*** (0.002). Baseline results under the sample of individuals included in column (4): (a) 0.107 (0.138), (b) 0.360*** (0.139), (c) 0.007*** (0.002).
Clustered (case manager–year) standard errors between parentheses.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

References

Andren, D., Svensson, M., 2012. Part-time sick leave as a treatment method for individuals with musculoskeletal disorders. J. Occup. Rehabil. 22, 418–426.

Andren, D., 2014. Does part-time sick leave help individuals with mental disorders recover lost work capacity? J. Occup. Rehabil. 24, 344–360.

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. J. Occup. Environ. Med. 42, 1172–1177.

Bethge, M., 2016. Effects of graded return-to-work: a propensity-score-matched analysis. Scand. J. Work Environ. Health 42, 273–279.

de Jong, P., Lindeboom, M., van der Klaauw, B., 2011. Screening disability insurance applications. J. Eur. Econ. Assoc. 9, 106–129.

de Jong, P., Gielen, M., Haanstra-Veldhuis, V., 2014. Verzekeringsgraad kleine werkgevers – eindrapportage, Research commissioned by the Ministry of Social Affairs.

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

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

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

French, E., Song, J., 2014. The effect of Disability Insurance receipt on labor supply. Am. Econ. J. Econ. Policy 6, 291–337.

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

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

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

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. Scand. J. Work Environ. Health 34, 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., Lindeboom, M., 2015. The rise and fall of disability insurance enrollment in the Netherlands. J. Econ. Perspect. 29, 151–172.

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

Krause, N., Dasinger, L.K., Neuhauser, F., 1998. Modified work and return to work: a review of the literature. J. Occup. Rehabil. 8, 113–139.

Maestas, N., Mullen, K.J., Strand, A., 2013. Does Disability Insurance receipt discourage work? Using examiner assignment to estimate casual effects of SSDI receipt. Am. Econ. Rev. 103, 1797–1829.

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

Markussen, S., Mykletun, A., Røed, K., 2012. The case for presenteeism – evidence from Norway's sickness insurance program. J. Public Econ. 96, 959–972.

Markussen, S., Røed, K., Schreiner, R.C., 2018. Can compulsory dialogues nudge sick-listed workers back to work? Econ. J. 128 (610), 1276–1303, http://dx.doi.org/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., 2018. Labour market effects of activating sick-listed workers. Labour Econ. 53, 15–32.

van Sonsbeek, J.M., Gradus, R.H.J.M., 2013. Estimating the effects of recent disability reforms in the Netherlands. Oxford Econ. Pap. 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. Scand. J. Work Environ. Health 38, 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. Br. J. Gen. Pract. 61, 794–800.