DOI: 10.1002/pam.22560

RESEARCH ARTICLE

Selective exercise of discretion in disability insurance awards

Pilar Garcia-Gomez[1,2] | **Pierre Koning**[2,3,4] | **Owen O'Donnell**[1,2,5] | **Carlos Riumalló-Herl**[1,2]

[1]Erasmus School of Economics, Rotterdam, The Netherlands

[2]Tinbergen Institute, Amsterdam, The Netherlands

[3]IZA, Bonn, Germany

[4]School of Business and Economics, VU Amsterdam, Amsterdam, The Netherlands

[5]Erasmus School of Health Policy and Management, Rotterdam, The Netherlands

Correspondence
Carlos Riumalló-Herl, Erasmus School of Economics at Erasmus University Rotterdam, P. O. Box 1738, 3000 DR Rotterdam, The Netherlands.
Email: riumalloherl@ese.eur.nl.

Funding information
H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 840591

Abstract

Variation in assessor stringency in awarding benefits leaves applicants exposed to uninsured risk that could be systematic if discretion were exercised selectively. Using administrative data on disability insurance (DI) applications in the Netherlands, we show that even in one of the most rule-based DI programs, there is still between assessor variation in awards, and there is systematic variation in assessment across applicants. Discretion is exercised in favor of lower-wage applicants relatively more than it is used to benefit higher-wage applicants. This is evident indirectly from downward discontinuities in pre-disability wages just above benefit entitlement thresholds and directly from wage-related differences in the extent to which assessors intervene in the semi-automated calculation of earnings capacity. While lower-wage applicants benefit on average, they are exposed to greater risk from between assessor variation in the exercise of discretion. Rule-based disability evaluation can reduce, but not eliminate, between-applicant variation in awards.

INTRODUCTION

Variation in the exercise of discretion that officials have in adjudicating social insurance and welfare claims leaves applicants exposed to uninsured risk of assignment to a relatively stringent assessor. For example, applicants to U.S. disability insurance (DI) programs face substantial variation in award propensities across claim assessors (French & Song, 2014; Maestas et al., 2013). Less is known about selective exercise of discretion in response to applicants' non-health characteristics. Assessors may show favor to economically vulnerable applicants for whom DI benefits can be particularly valuable (Deshpande & Lockwood, 2022). In systems that make disability benefits a positive function of lost earnings capacity, assessors may be more stringent with applicants with higher pre-disability earnings because a given disability imposes a greater proportionate reduction in their earnings capacity.

We test for the selective exercise of discretion in benefit awards by examining the processing and outcomes of the universe of applications to the Dutch DI program. This program has a sophisticated claim adjudication process that involves feeding an applicant's functional abilities (determined in a standardized medical examination), education, and basic skills into an algorithm that searches a database to find feasible job matches and calculates the respective earnings capacity (Maestas et al., 2021). The proportionate shortfall of earnings capacity from pre-disability earnings defines a *degree of disability* that determines DI entitlement. This delineated procedure still deliberately leaves scope for assessors to influence awards. Drawing on information, including on mental health and social skills, that is not fed into the algorithm, assessors are responsible for judging whether all algorithm-generated job matches are infeasible. Exclusion of a job match would reduce earnings capacity and may raise benefit entitlement. We test for the selective exercise of this discretion across applicants distinguished by pre-disability wages and earnings.

We first show indirect evidence of the selective exercise of discretion. On crossing degree of disability thresholds that determine entitlement to (higher) benefits, there are discontinuous drops in pre-disability wage rates and earnings, and there are upward jumps in a proxy for disability severity. These discontinuities are consistent with assessors exerting relatively more effort through the selection of job matches to lift lower-wage applicants above the thresholds for receipt of partial benefits and being relatively less lenient with higher-wage applicants. The insurance principle on which the program is founded ensures that for given disability severity and post-disability earnings capacity, benefit entitlement is an increasing function of pre-disability earnings. Consequently, higher-wage applicants can qualify, while more severely disabled lower-wage applicants do not. A sense of unfairness, or an appreciation of the greater value of benefits to the less fortunate, may provoke assessors to intervene to raise the entitlement of lower-wage applicants by excluding job matches based on the information that is available to them but not the algorithm, while being less likely to make such interventions for higher-wage applicants. We then show direct evidence that lower-wage applicants who end up just above the thresholds are relatively more likely than higher-wage applicants to have had their degree of disability raised through assessors deeming algorithm-generated job matches infeasible.

To gauge the consequences of the selective exercise of discretion, we estimate assessor effects on DI awards that are fixed within applicant pre-disability wage groups but are allowed to differ between them. We find that assessors vary more in deciding the benefits of lower-wage applicants and the leniency of many assessors depends on an applicant's wage. Around a quarter of assessors are more lenient than average with applicants in one wage tercile group while being more stringent than average with applicants in another. Using the distributions of assessor fixed effects, we predict that rejected low-wage (bottom tercile) applicants would have a negligible chance of being awarded benefits if randomly reassigned to another assessor. In contrast, rejected high-waged (top tercile) applicants would have a 13% chance. Opportunities to exercise discretion favorably appear to have been exhausted for rejected low-wage applicants but not for rejected high-wage applicants. Low-wage applicants who were awarded full benefits would face a 7% downside risk of getting lower benefits if randomly reassigned to another assessor, while the respective risk for high-wage applicants is negligible.

Evidence that U.S. DI awards are prone to misclassification errors (Benitez-Silva et al., 2004; Low & Pistaferri, 2015, 2019; Nagi, 1969; Wachter et al., 2011) and display substantial between-assessor variation (French & Song, 2014; Maestas et al., 2013) has prompted proposals for a more clearly defined adjudication procedure similar to that used in the Netherlands (Maestas, 2019; Maestas et al., 2021). Limiting discretion afforded to assessors can reduce system noise (Kahneman et al., 2021), and reduce the element of luck in DI awards that produces horizontal inequity, at the cost of reducing sensitivity to relevant but less objective information. We show that even in one of the most rule-based DI programs, there is still between-assessor variation in DI awards, and there is systematic variation in assessment across applicants.

Evidence of systematic bias in DI awards in relation to a non-health characteristic is rare. Low and Pistaferri (2019) found that U.S. DI false rejection rates are substantially higher for female applicants. This appears to be because women are incorrectly assessed to have higher likelihoods of finding work than men with the same observed health conditions, qualifications, and labor market experience. This

is consistent with our hypothesis that, possibly subconsciously, Dutch DI assessors favor lower-wage applicants because they are perceived to have worse labor market opportunities relative to higher-wage applicants.

Deshpande and Lockwood (2022) demonstrated that, somewhat paradoxically, mismatch between receipt of U.S. DI program benefits and realized health risks is likely to be welfare improving because errors in targeting health risks help to fill gaps in insurance of non-health risks. If lower-wage applicants to the Dutch DI program are exposed to greater uninsured non-health risks, which seems likely at least for labor market risks, then they may benefit more from DI benefits, even when disability impacts their earnings potential less than it does for higher-wage applicants. Deshpande and Lockwood (2022) found demand-side evidence of selection on non-health risks that arises through differences in the propensity to apply for DI. We find supply-side evidence of selection through assessors' exercise of discretion in awarding DI.

Between-assessor variation in award rates for DI and other social programs has become a popular instrument to identify effects on labor supply and other outcomes (Bakx et al., 2020; Dahl et al., 2014; Doyle, 2007; French & Song, 2014; Maestas et al., 2013). In various fields, including law (Bhuller et al., 2020; Dobbie et al., 2018; Kleinberg et al., 2018; Kling, 2006), medicine (Doyle et al., 2010; Doyle et al., 2015), and education (Figlio & Lucas, 2004), quasi-random assignment to discretion-exercising officials is increasingly exploited for identification—the "judges design." If assignment to an assessor raises the probability of program entry for some applicants but reduces it for others, as we find for low-wage and high-wage DI applications, then monotonicity—strict (Vytlacil, 2002), average (Frandsen et al., 2023), and probabilistic (Chan et al., 2022)—is violated and the instrumental variable estimator is not consistent for the local average treatment effect. Chan et al. (2022) demonstrated that monotonicity is violated when medical decision makers have heterogeneous skill in minimizing type I and II errors. We look beyond the instrumental use of the exercise of discretion to its consequences for the functioning of a disability insurance program.

This paper primarily contributes by being one of the first to deliver evidence of systematic differences in DI award rates. It shows that even in a predominantly rule-based social program, like Dutch DI, claim assessors exercise discretion selectively, although not necessarily consciously, to benefit lower-wage relative to higher-wage applicants.

In the next section, we describe the Dutch DI program, particularly the procedure for assessing applications. The section "Data" describes the administrative data used. The section "Exercise of Discretion – Evidence" presents indirect and then direct evidence of the selective exercise of discretion in evaluating DI applications. The section "Exercise of Discretion – Consequences" shows the consequences for DI awards of the selective exercise of discretion. The final section concludes.

DISABILITY INSURANCE IN THE NETHERLANDS

The Netherlands has a public DI program with compulsory enrollment of all employees. In the early 1990s, the country had one of the highest DI dependency rates in the world, with about 12% of the insured population receiving disability benefits (Koning & Lindeboom, 2015). Since the turn of the century, major reforms precipitated steep declines in the inflow and stock of DI claimants (Burkhauser et al., 2008; Koning & Lindeboom, 2015). Apparently, a key to this success was obliging employers to take steps to prevent employees from claiming long-term sickness and, subsequently, DI and giving them a financial incentive to make this prevention effective (Godard et al., 2022; Koning & Lindeboom, 2015). Under a so-called Gatekeeper Protocol, employers must organize preventive and reintegration actions while continuing to pay the wages of sick workers during a 2-year mandatory waiting period before application for DI. Together with experience rating that ties an employer's DI contributions to previous claims made by its employees, this protocol has been proposed as a benchmark reform of other DI programs, including Social Security Disability Insurance (SSDI) in the U.S. (Autor, 2015; Burkhauser et al., 2008).

A second distinguishing feature of the Dutch DI program is the extent to which it is designed to deliver insurance. Benefits are proportionate to pre-disability earnings, with the proportion determined by the extent to which disability reduces earnings capacity. The availability of partial benefits is intended to encourage part-time work and avoid labor market detachment—around 60% of those awarded partial benefits work (Deursen et al., 2019). In the U.S., SSDI mainly covers the risk that disability prevents work entirely and gives little opportunity or incentive for beneficiaries to work part-time. This has prompted calls for partial disability benefits (Maestas, 2019). The potential of such a reform depends on the extent to which it is feasible to implement and operate a system that produces detailed and consistent assessments of the impact of disability on earnings capacity, which the Dutch program aims to achieve.

Benefit entitlement is determined by an applicant's *degree of disability*, which is the percentage shortfall of their post-disability earnings capacity from their pre-disability earnings:

$$degree\ of\ disability = \left(1 - \frac{Wage_{post} \times Hours_{post}}{Wage_{pre} \times Hours_{pre}}\right) \times 100, \quad (1)$$

where $Wage_{pre}$ and $Hours_{pre}$ are, respectively, the hourly wage and the hours worked per week prior to applying for DI, and the potential post-disability wage ($Wage_{post}$) and work hours ($Hours_{post}$) are assessed on the basis of health limitations and educational attainment. For given post-disability earnings capacity, applicants with higher pre-disability earnings have greater degrees of disability and benefit entitlements.

To qualify for any DI benefit, the degree of disability must be at least 35%. Entitlement increases discontinuously at thresholds of 45%, 55%, 65%, and 80%. Applicants above the top threshold are classified as fully disabled and are paid benefits at 70% of pre-disability earnings, up to a maximum of approximately 3 times the minimum wage. For applicants classified as less than fully disabled, the replacement rate is 70% of the mid-point of the respective degree of disability interval. For example, an applicant assessed as 38% disabled receives 28% ($0.7 \times 40\%$) of pre-disability earnings (Appendix A[1]).

Post-disability earnings capacity ($Wage_{post} \times Hours_{post}$) is estimated following a medical examination by a physician and a subsequent interview with an occupational assessor. The physicians and occupational assessors are employed by the public insurance agency. An examining physician uses a standardized instrument to identify functional impairments arising from the nature and severity of an applicant's health problem.

An occupational assessor feeds this information, along with an applicant's education, basic skills (e.g., driving license and computer skills), and physician-assessed constraints on working hours into an algorithm that searches the agency's database to identify jobs that the applicant is deemed capable of performing (Uitvoeringsinstituut Werknemersverzekeringen [UWV], 2013).[2] The database contains jobs that are performed in the labor market irrespective of whether there are current vacancies for these jobs. The resulting job matches are ordered from lowest to highest implied degree of disability (Appendix Figure A1). The algorithm cannot match on functional abilities related to mental capacity and social skills. Hence, the assessor reviews the listed jobs, using information provided by the physician, to determine feasibility. The assessor must review the listed jobs in ascending order of the degree of disability and select the first three that are both feasible and for which there are at least three positions (not necessarily vacant) in the database (Appendix Figure A2). If possible, the assessor selects some spare jobs (> 3) that imply higher degrees of disability. If it is not possible to

[1] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

[2] For each job, the database contains the wage and hours, as well as physical and mental capacities required to perform job-related tasks. The insurance agency collects this information by visiting employers throughout the country. Each visit focusses on starting positions that could potentially have been filled by people with functional impairments.

select three jobs, then the applicant is classified as fully disabled with 100% degree of disability. Otherwise, the applicant's potential wage ($Wage_{post}$) is the median wage of the three jobs with the lowest degrees of disability, and potential hours ($Hours_{post}$) are the minimum hours across these three jobs.

If the assessor takes no further action, the resulting degree of disability is used to calculate benefit entitlement. However, the assessor may exclude certain job matches, switching to spares, and re-calculating the degree of disability. This would lead to upward revision of the degree of disability. This process can continue interactively, giving the occupational assessor the opportunity to observe the impact of different job selections on benefit entitlement. The occupational assessor can also ask the physician to reconsider whether a functional limitation score is a complete and accurate reflection of the applicant's capacity. Occasionally, the physician makes changes that facilitate matching of functional impairments to (other) jobs, and this reduces the degree of disability (Appendix A).

While the assessment process is reasonably tightly defined, it intentionally leaves scope for the exercise of discretion. Occupational assessors are responsible for evaluating the algorithm-generated jobs matches and judging whether each is feasible given an applicant's functional, mental, and social capacities. Deeming job matches to be infeasible can raise benefit entitlement. Refraining from doing so can leave an applicant below a threshold at which they would qualify for any, or a higher, benefit. Assessors exercise discretion when they decide to exclude job matches in some cases and when they decide not to do so in others. Each decision can be consequential. Discretion could be used indiscriminately or selectively. The fact that lower-wage applicants must be more functionally impaired to reach a given benefit entitlement may motivate some assessors to intervene to exclude job matches relatively more often for such applicants.

Consider applicant *L* who is assessed to have more functional limitations than applicant *H*, who has a higher pre-disability wage. The wage difference may dominate, such that, even if both applicants were assessed to be capable of performing the same jobs, *H* would have a higher degree of disability because of a greater loss of earnings capacity. Applicant *H* may qualify for (higher) DI benefit, while *L* would not. Potentially, an assessor could respond by classifying better-paying job matches found by the algorithm as infeasible for *L*, which would raise the degree of disability to an extent sufficient for *L* to qualify for (higher) DI benefit. Or the assessor may exert greater effort in finding better-paying job matches for *H,* which would reduce the degree of disability and benefit entitlement of *H*. Since the algorithm finds matches based on education as well as health, if two applicants were similar in these two characteristics but one had lower pre-disability earnings, that applicant would have a lower degree of disability and, hence, lower benefit entitlement.

Assessors need not consciously compare applicants in this way. They may simply learn to identify the type of applicants who qualify for (higher) DI and subconsciously exert greater effort in evaluating the feasibility of the final job set to help those who otherwise would not. Variation in the exercise of discretion across occupational assessors need not reflect differences in consciously motivated actions. It could reflect differences in daily routines or informal guidelines issued by managers.

To summarize, the Dutch DI program leans strongly toward rule-based determination of benefit entitlement. Assessment of post-disability earnings capacity at the intensive margin is more sophisticated and consequential than in systems that determine eligibility by comparing work capacity, assessed subjectively by a physician or committee, with a single minimum threshold. However, there are limits to the objective assessment of earnings capacity. The program deliberately leaves scope for the occupational assessor to exercise discretion in assessing the work that an applicant can do and find, given the nature and severity of their health condition as well as labor market conditions.

DATA

We use administrative records covering the universe of DI applications from January 2006 to July 2017 (Appendix B). We exclude applications that were decided without an occupational assessment (∼30% of all applications). Almost all of these were either rejected based on a medical examination

that was sufficient to establish that there was no disability, or they were halted before completion of the application process. Less than 0.1% of applicants were classified as 100% disabled and awarded full disability benefits based on medical examination only.

The remaining applications ($n \approx 400{,}000$) are those for which the degree of disability was calculated using eq. (1). For each of these applications, we have data on a) demographics, education, pre-disability employment contract (permanent or temporary), industry, wage rate ($Wage_{pre}$), and work hours ($Hours_{pre}$) at the time of application; b) diagnosis and a functional limitations score from the medical examination; and c) potential wage rate ($Wage_{post}$), potential work hours ($Hours_{post}$), and degree of disability from the occupational assessment. All wage values are expressed in terms of 2010 prices. The functional limitations score is a count of the number work-related physical and cognitive tasks (out of 100) that the physician confirms or judges that the applicant cannot perform (Appendix A). This is not used by the algorithm to find potential job matches, and so it does not directly determine benefit entitlement. We use it as a summary proxy of the severity of functional impairment similar to the scores of limitations in daily activities (Katz, 1983). It is positively correlated with the degree of disability. For a given degree of disability, it is lower at higher wages (Appendix Figure C1).

For each application, we also observe the initial selection of jobs from the algorithm-generated job matches, any jobs from this selection that the assessor finally deems to be infeasible and so are excluded from the calculation of degree of disability, and, for each job match, the respective hours and hourly wages that would be used in the calculation if that match were included in the final selection (Appendix Figure A2). Further, for each application, we can use anonymized codes to uniquely identify the examining physician, the occupational assessor, and the assessment office. The number of medical examinations and occupational assessments conducted for each application are observed only from January 2011 until July 2017.

The average age of DI applicants is 46 years (Appendix Table C1). A slight majority is female (52%). Around a quarter (26%) of applicants have no more than compulsory education. Before application, a slight majority (52%) was either on a temporary contract or unemployed, average work hours were about 32 per week, and the average hourly wage was 16.53 euros (2010 prices). The most prevalent main diagnosis is a musculoskeletal condition (35%), followed by a mental health problem (30%). The average functional limitations score indicates difficulty in almost 13 (out of 100) domains of work-related tasks. The average degree of disability (37.5%) is just above the 35% threshold at which an applicant is entitled to the lowest partial DI benefit. Around 59% of applicants do not reach this threshold, while 21% have a degree of disability of at least 80% and qualify for the full benefit.

RESULTS

Exercise of discretion - evidence

We first show indirect evidence of the selective exercise of discretion in the form of discontinuities in applicants' characteristics at benefit entitlement thresholds and bunching or missing mass of applications around those thresholds. We then provide direct evidence of the selective exercise of discretion through differences in the extent to which assessors deem job matches to be infeasible and how this impacts the degree of disability.

Degree of disability is a positive function of pre-disability earnings (eq. 1). The functional limitations score does not directly determine the degree of disability. However, these two variables will be positively related if post-disability potential earnings fall as the limitations count rises. If discretion were not exercised selectively, each of pre-disability earnings and the functional limitations score would be expected to increase smoothly, if not linearly, with the degree of disability through the DI entitlement thresholds. To assess this, first we non-parametrically regress each of the variables on the degree of disability within intervals of the latter separated by the thresholds. Fitted values from these regressions plotted in Figure 1 reveal that there appear to be downward discontinuities

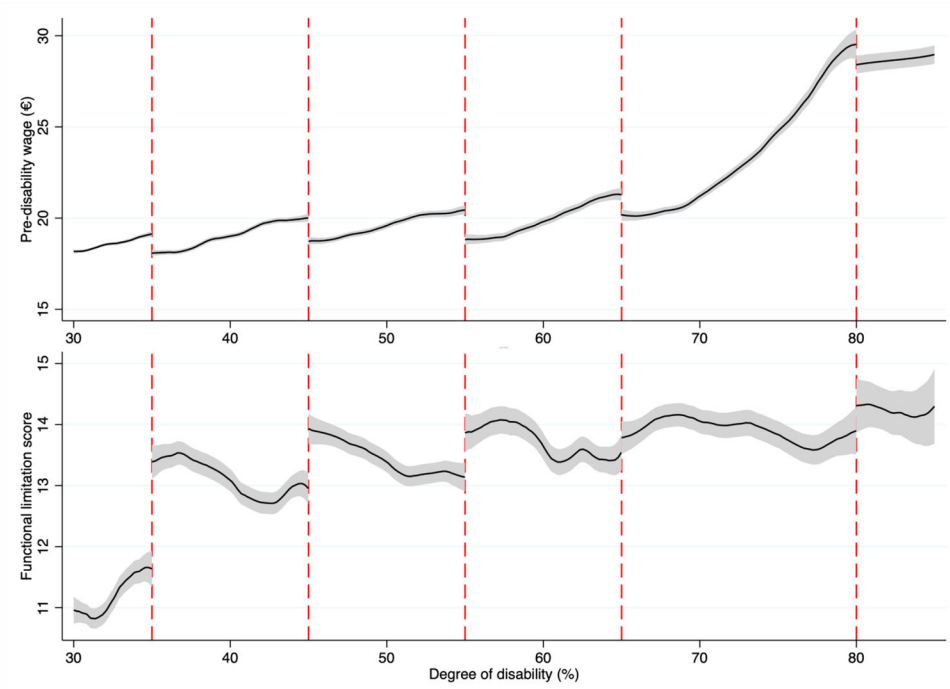

FIGURE 1 Pre-disability wage and functional limitation score by degree of disability.

Notes: Figure plots fitted values from local linear regressions within each degree of disability interval defined by DI entitlement thresholds: [30, 35), [35, 45), [45, 55), [55, 65), [65,80), and [80, 85]. We use the Epanechnikov kernel function and select bandwidths by a pseudo rule-of-thumb plug-in estimator (Fan & Gijbels, 1996). Shaded area shows 95% confidence intervals derived from the square root of the conditional variance of the estimator at each smoothing point.

in pre-disability wages and upward discontinuities in functional limitation scores at the entitlement thresholds.

To formally test for discontinuities at the thresholds, we estimate regressions of this form,

$$Y_i = \beta_\tau \cdot 1\left(DD_i \geq T_\tau\right) + f_{\tau L}\left(DD_i\right) \cdot 1\left(DD_i < \mathrm{T}_\tau\right) + f_{\tau R}\left(DD_i\right) \cdot 1\left(DD_i \geq \mathrm{T}_\tau\right) + \varepsilon_{\tau i}\,, \tag{2}$$

where Y_i is, in turn, the pre-disability hourly wage, hours worked per week, weekly earnings, and the functional limitations score of applicant i; DD_i is the degree of disability; T_τ is a threshold equal to 35%, 45%, 55%, 65%, or 80% in the respective regression; 1(.) is the indicator function; and $f_{\tau L}(DD_i)$ and $f_{\tau R}(DD_i)$ are flexible functions of the degree of disability, which are specified as third order polynomials in the main estimates. We use the same coverage error optimal bandwidth on each side of the threshold and conduct robust biased-corrected inference (Calonico et al., 2020).

Consistent with Figure 1 and with the selective exercise of discretion, Table 1 shows that there are discontinuities at entitlement thresholds. The pre-disability wage drops significantly ($p < 0.01$) on crossing each of the three lower thresholds from below. The estimated €1.37 drop at the 35% threshold implies that applicants given a degree of disability just sufficient to get the lowest benefit had a wage that was 7.0% lower than the wage of those just denied DI. The direction of this effect runs counter to the built-in propensity for the degree of disability to increase with pre-disability earnings. The wage discontinuities are greater—both in absolute and relative terms—at the 45% and 55% thresholds where the benefit increases. At the 65% threshold, the estimated wage drop is smaller and not significant ($p \geq 0.1$). At the top threshold that determines entitlement to full benefit, the point estimate is larger, and the lack of significance could possibly be due to the smaller sample size.

TABLE 1 Discontinuities in determinants of disability insurance awards at entitlement thresholds.

	Degree of disability threshold for DI entitlement				
	35%	**45%**	**55%**	**65%**	**80%**
Wage per hour (€)	-1.365	-1.637	-2.197	-0.644	-1.820
	(0.226)	(0.296)	(0.499)	(0.544)	(1.195)
Hours per week	0.705	0.972	-0.122	-0.058	0.365
	(0.352)	(0.368)	(0.381)	(0.359)	(0.591)
Earnings per week (€)	-30.666	-37.586	-86.993	-22.528	-46.966
	(9.766)	(13.463)	(18.201)	(20.602)	(47.144)
Functional limitations score	1.655	0.984	0.608	0.131	0.697
	(0.306)	(0.387)	(0.418)	(0.384)	(0.751)
N	40,152	20,081	17,529	20,736	6,426

Notes: Bias-corrected robust regression discontinuity estimates of β_τ from model (2) with $f_{\tau L}(DD_i)$ and $f_{\tau R}(DD_i)$ each specified as a different third-order polynomial and the same (coverage error probability) optimal bandwidth on each side of the threshold. Robust standard errors in parentheses. Wage, hours, and earnings are prior to application for DI. See Appendix Figure C2 for corresponding regression discontinuity plots.

There are significant upward discontinuities in hours worked at the 35% and 45% thresholds. But the drops in wages dominate, resulting in significant ($p < 0.05$) discontinuous downward shifts in pre-disability weekly earnings at the 35%, 45%, and 55% thresholds (see also Appendix Figure C3). The pattern of results suggests that the exercise of discretion favors lower-wage applicants working longer hours, and not simply lower-earning applicants.

There are significant ($p < 0.05$) upward discontinuities in the functional limitations score as the degree of disability crosses the 35% and 45% thresholds from below. Applicants given a degree of disability just sufficient to qualify for DI ($\geq 35\%$) were assessed to have 1.66 more functional limitations than those who just failed to qualify, which corresponds to a 14.1% increase. The point estimates indicate upward jumps in limitations at the 55%, 65%, and 80% thresholds, although none of these is close to significance at conventional levels.

While applicants who qualify for (higher) benefits are expected to be more functionally impaired, the discontinuity in the relationship is difficult to explain in the absence of the selective exercise of discretion. Assessors make more interventions in the selection of job matches for applicants who are more functionally impaired but have lower pre-disability wages that are closer to a floor determined by the minimum wage and collective bargaining. This floor constrains the potential fall in post-disability earnings, and so the degree of disability. Assessors may be relatively more likely to accept the job selections for less impaired but higher-wage applicants on the margin of (higher) entitlement, leaving these applicants below the respective threshold. This is consistent with discontinuities in the pre-disability wage and the functional limitations score at the 35% threshold that are evident and significant only among applicants in the top tercile of the wage distribution (Appendix Figure C4). In the lower-wage groups, not only is there no wage discontinuity at the threshold, but there is also no increase in the pre-disability wage with the degree of disability through the threshold. This also suggests the use of discretion. By design, the degree of disability should be a positive function of the pre-disability wage if all else is equal (eq. 1). Assessors may intervene in the selection of job matches to a greater extent towards the bottom than the top of the wage distribution. Within terciles of the distribution, an inverse relationship between the propensity to intervene and the wage can also contribute to the wage discontinuities at the thresholds seen in Figure 1.

The estimates of threshold discontinuities are robust to using different functions of the degree of disability and bandwidths (Appendix Table C2). There are some differences in magnitudes, but the direction and significance of the discontinuities do not change.

One may hypothesize that these discontinuities arise from lumpiness in the wage distribution of the algorithm-generated job matches or some other source of nonlinearity that may exist even without selective exercise of discretion and that is not fully captured by the flexible functions specified on either side of the respective threshold. If this were the case, then discontinuities would emerge also at degrees of disability that are not entitlement thresholds. To test this, we estimate equation (2) for the pre-disability wage (and the functional limitations score) at each percentile point of degree of disability (T_τ) from 20% to 85%. We find only three significant ($p < 0.05$) wage discontinuities out of 66 degree of disability values that are not at DI entitlement thresholds (Appendix Figure C5). This predominance of null placebo effects suggests that the three significant wage discontinuities (and the two significant limitations score discontinuities) that are found across only five entitlement thresholds are unlikely to be attributable to unmodelled nonlinearities that are built into the determination of degree of disability. They are more likely to arise from occupational assessors' selective exercise of discretion in excluding job matches close to the thresholds.[3]

The discontinuities at each threshold can result from assessors deeming algorithm-generated job matches as infeasible relatively more often for lower-wage applicants, who have more functional limitations at any given degree of disability, than for higher-wage applicants with fewer limitations. Such systematic difference by wage in the rate of interventions to rule out job matches for applicants who otherwise would not qualify for (higher) benefits would raise the relative likelihood of hoisting up lower-wage applicants to reach a (higher) degree of disability threshold while higher-wage applicants would be more likely to be held below a threshold. Figure 2 shows that in the top wage tercile and the bottom two functional limitations terciles, applications are indeed bunched below degree of disability thresholds, leaving missing mass in regions where entitlement increases. Non-parametric bunching estimates (Cattaneo et al., 2018) confirm that there is missing mass in the distributions of high-wage and lower-limitations applications at the three lowest thresholds (Appendix Table C4). Around these thresholds, high-wage applicants overlap considerably with those who have lower functional impairment since a low-impairment applicant cannot reach the margin of (higher) entitlement if their pre-disability wage is also low. This explains why the bunching is similar in the bottom-left (high wage) and top-right (low impairment) panels of Figure 2. The fact that high-wage applicants can get to the margin of qualifying for DI even with relatively low impairment may reduce the likelihood that assessors intervene to deem job matches infeasible for such applicants.

In the bottom two wage terciles, there is no missing mass to the right of the thresholds. This suggests that assessors do not reduce their propensity to deem job matches infeasible for lower-wage applicants who would not qualify for (higher) benefits without such intervention. An increase in the intervention propensity for these applicants might be expected to generate excess mass to the right of the thresholds in the lower-wage terciles. While this is not apparent in Figure 2, non-parametric estimation does reveal instances of excess mass above the three lowest thresholds for lower- and middle-wage terciles (Appendix Table C4).[4] This is not evident in the figure partly because ruling out job matches does not necessarily produce marginal changes in the degree of disability of lower-wage applicants that would leave them just above a threshold. Further, selective exercise of discretion can occur marginally within wage terciles and not only between them. Systematic differences in the propensity to rule out job matches that favors *relatively* lower-wage applicants within each tercile group can also contribute to the discontinuities observed in Figure 1 and Table 1 but are not evident from between group comparisons in Figure 2.

Low-wage applicants are more likely to have a musculoskeletal (main) diagnosis and less likely to have a psychiatric diagnosis (Appendix Table C5). A higher percentage of the most function-

[3] The discontinuities do not appear to result from differences in recourse to medical re-examination, which is relatively rare (Appendix Table C3). For example, at the 35% threshold, discontinuities are larger, and only significant, for applications with just one medical examination (Appendix Figure C6).

[4] There is bunching below and missing mass at the three lowest degree of disability thresholds in the unstratified distribution of all applications (Appendix Figure C7 and Appendix Table C4).

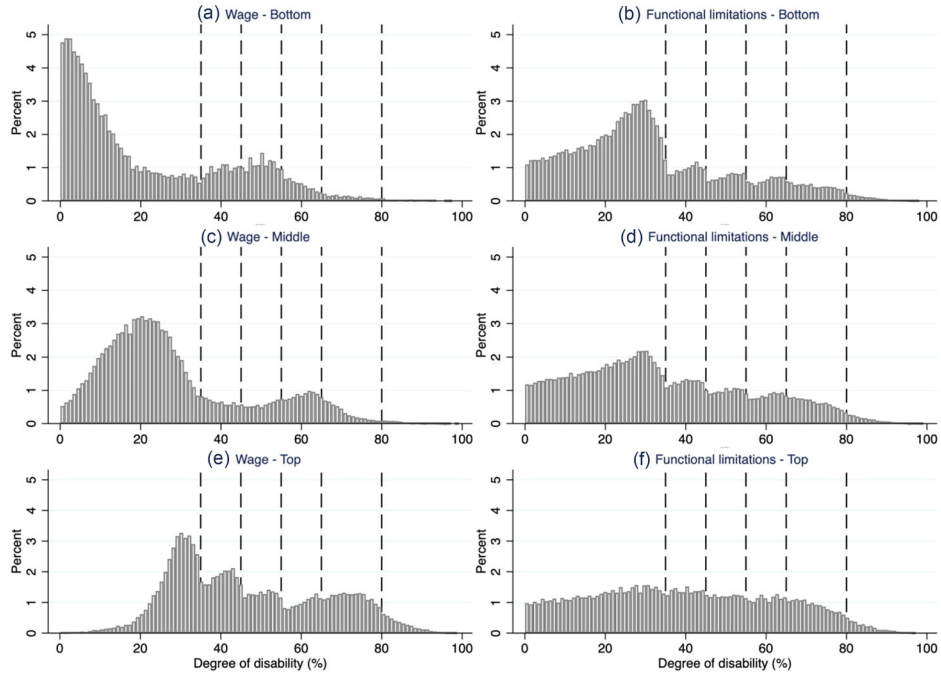

FIGURE 2 DI applications by degree of disability stratified by wage and functional limitations score tercile groups.

Notes: Each panel shows a distribution of DI applications by degree of disability. In panels A, C, and E applications are stratified into the bottom, middle and top third of the pre-disability hourly wage distribution, respectively. In panels B, D, and F applications are stratified into the bottom, middle, and top third of the functional limitation score distribution, respectively. To focus on thresholds critical to benefit entitlement, in this figure we exclude applications with degree of disability of 0 (~30% of applications) and 100 (~20% of applications). Sample sizes: 35,008, 78,542, and 78,825 for panels A, C, and E, respectively; 53,164, 97,559, and 41,815 for panels B, D, and F, respectively. The number of applications is not the same across groups because the terciles are calculated including those with degree of disability of 0 and 100. Dashed lines represent the DI entitlement thresholds. See Appendix Table C4 for bunching estimates and Appendix Figure C7 for the unstratified density of all applications.

ally impaired applicants have musculoskeletal conditions and a lower percentage have psychiatric conditions (Appendix Table C5). These differences in the diagnosis composition of groups do not explain the wage and functional limitations score discontinuities at the entitlement thresholds observed in Figure 1. After stratifying by main diagnosis, there are still drops in wages and upward jumps in functional limitations at the thresholds (Appendix Figure C8). For those diagnosed with a psychiatric condition, there is less bunching below (and missing mass above) the 45%, 55%, and 65% degree of disability thresholds than is observed for other diagnoses, including musculoskeletal conditions (Appendix Figure C9). Since high-wage applicants are more likely to have a psychiatric condition, the differential bunching across wage groups in Figure 2 cannot be due to differences in the diagnosis composition of wage groups.

We now present more direct evidence of the selective exercise of discretion by comparing an applicant's final degree of disability—the one used to calculate their benefit entitlement—with the potential degree of the degree of disability they would have been given if the assessor had not intervened to deem (higher-paying) job matches infeasible. The final degree of disability cannot be lower than the potential degree of disability. For this analysis, we use applications that end up with a final degree of disability that is up to five percentage points (pp) above each threshold, i.e. 35% to 40%, 45% to 50%, 55% to 60%, 65% to 70% and 80% to 85%. We calculate four measures of the discrepancy between the final and potential degrees of disability. We average each measure over

applications within the respective final degree of disability interval and in each wage tercile group. This allows us to compare the propensity to exclude job matches and the consequences of doing so across wage terciles among applicants who just qualify for each level of benefit. We do not use regression discontinuity design because there is no reason to expect any discontinuity in the discrepancy between the potential and final degrees of disability when the latter, or indeed the former, crosses a threshold.

Panel A of Table 2 shows, for applications within a five percentage point interval above the respective threshold that gives entitlement to (higher) benefits, the proportion for which there is a discrepancy between the final and potential degrees of disability. For example, among those who just qualify for the lowest benefit by having a final degree of disability of 35% to 40%, assessors raise that measure above the potential degree of disability by deeming at least one job match infeasible for 46.2% of the low-wage applicants and 32.5% of the respective high-wage applicants. The wage gradient is the same at all the other thresholds. Assessors intervene more with the job selections for lower-wage applicants. For around 60% of assessors, the likelihood of intervening is higher when the applicant is low-wage rather than high-wage (Appendix D, Figure D1). Less than 35% of assessors intervene more frequently for high-wage applicants.

Panel B shows the mean difference between the final and potential degrees of disability. This difference cannot be negative. Panel C shows the mean difference conditional on there being any difference. Among those just qualifying for the lowest benefit, assessors raise the degree of disability of low-wage applicants by 3.5 pp, on average, by ruling that some higher-paying job matches are infeasible. The mean increase in the degree of disability for this wage group rises to 7.7 pp when we condition on the exclusion of at least one job match. For high-wage applicants, the respective mean difference and conditional mean difference are only 0.31 pp and 0.94 pp. At all other thresholds, both the mean difference and the conditional mean difference between the final and potential degrees of disability fall monotonically in moving from the low- to middle- to high-wage groups. This is further evidence of the selective exercise of discretion in favor of lower-wage applicants which can produce important jumps in the final degree of disability.

Panel D shows the proportion of applicants with a final degree of disability within five percentage points above the respective threshold who would have been below this threshold, and so receiving a lower or no benefit, if there had been no intervention to deem job matches infeasible. For example, almost 25% of low-wage applicants who were just above the 35% threshold would have been below it if higher-paying job matches found by the algorithm had not been ruled infeasible. By contrast, less than 5% of similarly positioned high-wage applicants would have been denied the partial DI benefit awarded if the potential degree of disability had been used. At each of the other thresholds, lower-wage applicants were more likely to be lifted above the threshold, and so receive higher benefits, because assessors deemed job matches infeasible. The patterns observed in all four panels of Table 2 are robust to conditioning on the functional limitations score, demographics, education, industry, diagnosis, and ZIP code fixed effects (Appendix Table C6).

Figure 3 provides further evidence that exclusion of job matches is more common and consequential for lower-wage applicants on the margin of qualifying for (higher) benefits. This figure restricts attention to applicants with a *potential* degree of disability within five percentage points *below* each threshold. For each threshold, the discrepancy between the final degree of disability and the potential degrees of disability among applicants who would not qualify for (higher) benefits without intervention decreases as pre-disability wage increases. For the lowest-wage applicants, the impact of the exclusion of job matches on the final degree of disability can be substantial and lead to an applicant who would not otherwise qualify for (higher) benefit being more than just above the respective threshold.

Table 2 and Figure 3 provide direct evidence that discretion is exercised through the exclusion of job matches, that it is selective—benefiting lower-wage applicants relatively more than higher-wage applicants—and that it is potentially consequential for benefit entitlement.

TABLE 2 Discrepancies between final and potential degrees of disability by pre-disability wage tercile, marginal applicants awarded (higher) DI.

	Degree of disability threshold				
	35%	**45%**	**55%**	**65%**	**80%**
A: Proportion with final degree of disability > potential degree of disability					
Low wage	0.462	0.509	0.467	0.484	0.449
	(0.013)	(0.011)	(0.015)	(0.028)	(0.060)
Middle wage	0.377	0.356	0.338	0.405	0.429
	(0.009)	(0.010)	(0.009)	(0.010)	(0.028)
High wage	0.325	0.335	0.335	0.345	0.386
	(0.006)	(0.007)	(0.008)	(0.007)	(0.011)
N	11,124	9,020	7,690	7,447	2,292
B: Mean (final degree of disability – potential degree of disability)					
Low wage	3.547	3.225	2.608	5.545	5.247
	(0.210)	(0.167)	(0.217)	(0.709)	(1.515)
Middle wage	0.770	0.955	1.045	1.023	2.722
	(0.060)	(0.085)	(0.084)	(0.086)	(0.422)
High wage	0.306	0.386	0.528	0.621	0.670
	(0.018)	(0.027)	(0.041)	(0.042)	(0.056)
N	11,124	9,020	7,690	7,447	2,292
C: Mean (final degree of disability – potential degree of disability | final DD > potential DD)					
Low wage	7.671	6.332	5.590	11.447	11.678
	(0.402)	(0.297)	(0.431)	(1.306)	(2.994)
Middle wage	2.043	2.685	3.086	2.526	6.344
	(0.151)	(0.226)	(0.237)	(0.202)	(0.895)
High wage	0.943	1.155	1.576	1.797	1.737
	(0.054)	(0.077)	(0.117)	(0.118)	(0.138)
N	3,974	3,410	2,740	2,759	902
D: Proportion of those at or above threshold who would be below if used potential degree of disability					
Low wage	0.250	0.262	0.204	0.275	0.261
	(0.007)	(0.007)	(0.008)	(0.016)	(0.038)
Middle wage	0.076	0.081	0.079	0.107	0.192
	(0.005)	(0.006)	(0.005)	(0.006)	(0.018)
High wage	0.046	0.052	0.060	0.066	0.099
	(0.003)	(0.004)	(0.005)	(0.004)	(0.007)
N	11,124	9,020	7,690	7,447	2,292

Notes: Final degree of disability is the value used to decide the DI benefit awarded. Potential degree of disability is the value that would have resulted if the three highest-paying job matches had been used without deeming any infeasible. Each column calculated using applicants with a final degree of disability within five percentage points at or above the respective threshold. Panels A, B, and D include all such applicants. Panel C includes the subset with a final degree of disability greater than potential degree of disability. Rows are stratified by pre-disability wage tercile groups: Low wage = bottom third of wage distribution, etc.

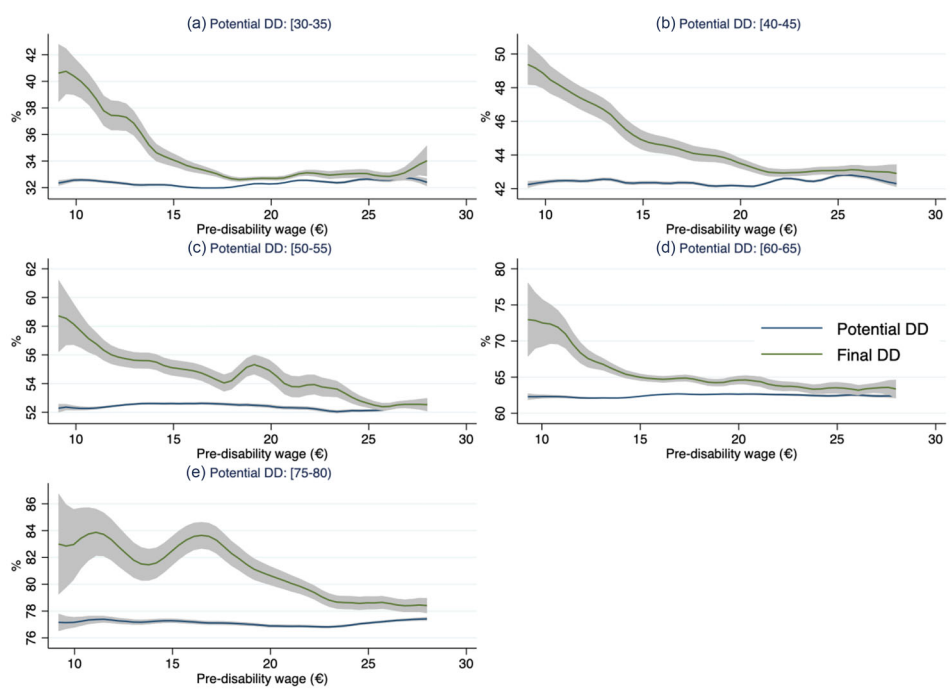

FIGURE 3 Final and potential degrees of disability by pre-disability wage, marginal applicants who would not have been awarded (higher) DI.

Notes: Final degree of disability is the value used to decide the DI benefit awarded. Potential degree of disability is the value that would have resulted if the three highest-paying job matches had been used without deeming any infeasible. Calculations use applicants with a potential degree of disability between 30 and 34.9 (Panel A), 40 and 44.9 (Panel B), 50 and 54.9 (Panel C), 60 and 64.9 (Panel D), and 75 and 79.9 (Panel E).

Exercise of discretion - consequences

To further gauge the importance of the exercise of discretion and of systematic differences in its prevalence, we estimate assessor fixed effects (FE) in determining the degree of disability and use them to simulate DI awards under counterfactual assignments of applicants to assessors. To allow assessor behavior to depend on applicants' pre-disability wages, we stratify on that variable and estimate,

$$DD_{igjrt} = \alpha^g + \lambda_j^g + X_i\beta^g + \delta_r^g + \tau_t^g + \varepsilon_{ijrt}^g, g \in \{low,\ middle,\ high\}, \quad (3)$$

where DD_{igjrt} is the final degree of disability of applicant i in wage tercile group g and ZIP code area r who is assigned to assessor j in year t. λ_j^g are wage group specific assessor FE, X_i is a vector of controls that includes the applicant's age, sex, employment contract at the time of application, functional limitations score, limitations on daily or weekly work schedule, main diagnosis, educational attainment, and hours worked at application. δ_r^g and τ_t^g are ZIP code and year FE, respectively, and ε_{ijrt}^g is a stochastic error. The inclusion of year FE allows for variation in general economic conditions that may affect the degree of disability awarded directly or through the composition of applicants. We estimate these regressions by OLS using applications that were assigned to assessors who handled at least 10 applications per wage group and 50 in total.[5]

[5] On average, an assessor handled 116 applications (Table C7). About half (50.4%) of the 2,635 assessors meet the inclusion criteria of handling at least 10 applications within each wage tercile group and at least 50 applications in total. Most of the excluded assessors handled fewer than 10 applications in total (Figure C10).

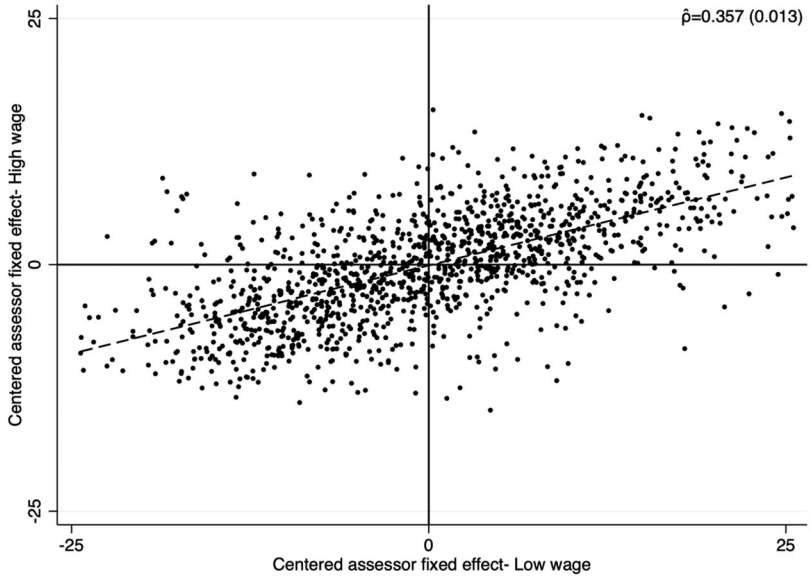

FIGURE 4 Scatter of assessor fixed effects for high- and low-wage applicants.

Notes: Each dot represents one assessor. Estimates of fixed effects are obtained from eq. (3) and then centered at the respective wage tercile group mean. Positive values indicate assessors who are more lenient than average. $\hat{\rho}$ is the estimated correlation coefficient, with the standard error in parentheses.

Figure 4 shows a scatter plot of mean-centered assessor FE estimated using applicants in the high-wage group against the respective FE obtained from the low-wage group. Greater horizontal than vertical dispersion of the effects indicates greater variation between assessors in determining the degree of disability of low-wage applicants.[6] For a low-wage applicant, assignment to an assessor carries more risk, in the sense that it is more consequential for their benefit entitlement. There is an upside risk of being assigned an assessor who systematically exercises discretion in favor of lower-wage applicants, and a downside risk of getting an assessor who tends to use the three highest-paying job matches with little sensitivity to additional information on mental health and social functioning that is not available to the algorithm that generates the matches. There is a moderate correlation ($\hat{\rho} = 0.357$) between the assessor FE for low-wage and high-wage applicants. Around a quarter (24.7%) of the assessors are more lenient (higher degree of disability) than average when evaluating applicants from one wage group while being more stringent than average when evaluating applicants from the other group. This discordance is not the mechanical result of compensatory decisions to remain within a cap on the proportion of applications an assessor can award because there is no such cap.

We predict each applicant's degree of disability if assigned to a different assessor by substituting the estimated FE of that assessor ($\hat{\lambda}_k^g$) for the estimated FE of the assessor to which the applicant is assigned ($\hat{\lambda}_j^g$),

$$\widetilde{DD}_{igkrt} = DD_{igjrt} + \left(\hat{\lambda}_k^g - \hat{\lambda}_j^g\right) k \neq j. \tag{4}$$

Repeated prediction under assignment to each assessor gives a distribution of counterfactual degrees of disability for each applicant. We use this to calculate the probabilities of being awarded no benefit

[6] See Appendix Figure C11 for clearer graphical evidence of the greater variance of the FE in assessments of lower-wage applicants and Appendix Table C8 for confirmation that the standard deviation and the inter-quartile range of the FE are both larger for lower-wage groups.

($Pr(\widetilde{DD}_{igkrt} < 35)$), partial benefit ($Pr(35 \leq \widetilde{DD}_{igkrt} < 80)$), and full benefit ($Pr(\widetilde{DD}_{igkrt} \geq 80)$) if randomly assigned to an assessor. We average these probabilities within each wage tercile group crossed with actual DI award category (no benefit, partial benefit, and full benefit) and show the results in panel A of Table 3.[7]

Low-wage applicants who were denied DI (DD < 35) had a negligible chance (<0.01%) of being awarded partial benefit if each had been assigned to a different assessor. When assessing these applicants, assessors appear to have exhausted opportunities to exercise discretion by ruling higher-paying job matches infeasible. This is not true for rejected high-wage applicants. On average, they would have a 13% chance of getting partial benefit if assigned to another assessor.

For those actually paid partial benefits, only the low-wage applicants are effectively at risk of an outcome other than a different level of partial benefits if randomly assigned to another assessor. For this group, the downside risk of losing entitlement (7.3%) is substantially greater than the negligible upside risk of acquiring full benefits. Similarly, among those awarded full benefits, around 7% of the low-wage applicants would have received only partial benefits if assigned to another assessor, compared with a negligible risk in the high-wage group. The fact that benefits paid to accepted low-wage applicants are more contingent on the assessor to which they are assigned is consistent with some assessors being more likely to rule higher-paying job matches as infeasible for this group.

The bottom two panels of Table 3 give, for each wage and actual DI award category, the predicted percentages of applicants who would be awarded no benefit, partial benefit, and full benefit if all were assigned to the most lenient assessor within the respective wage group ($\hat{\lambda}_k^g = max(\hat{\lambda}_1^g \ldots, \hat{\lambda}_K^g)$, panel B) and to the most stringent assessor ($\hat{\lambda}_k^g = min(\hat{\lambda}_1^g, \ldots, \hat{\lambda}_K^g)$, panel C). Since there is only a small probability that an applicant would be assigned to the most extreme assessor in each direction, these counterfactuals overstate the consequences of the exercise of discretion. We show them to demonstrate that our finding of differential effects across wage groups is robust to restricting attention to extreme scenarios.[8]

If all applicants were assigned to the most lenient (within wage group) assessor, then around 33% of rejected low-wage applicants and 90% of rejected high-wage applicants would get partial benefits. This implies that the fraction of *never takers* among rejected applicants is much larger for the low-wage than it is for the high-wage. In contrast, assignment to the most stringent assessor would be most consequential for the lower-wage groups. Among those awarded partial benefits, assignment to the most stringent assessor would have resulted in rejection of around 79% of the low-wage applications but only around 39% of the high-wage applications. About 87% of low-wage applicants, but only 31% of high-wage applicants, who were awarded full benefits would have received partial benefits if assigned to the most stringent assessor.

The downside risk of losing entitlement if assigned either randomly to another assessor or specifically to the most stringent assessor is larger for low-wage applicants. The upside risk of gaining entitlement if reassigned either randomly or specifically to the most lenient assessor is larger for high-wage applicants. Both patterns are consistent with selective exercise of discretion in favor of the low-wage group.

CONCLUSION

As with many other social programs, disability insurance awards benefits by applying objective rules in combination with judgement to take account of characteristics and circumstances that are difficult

[7] Appendix Figure C12 gives cumulative distributions of counterfactual degrees of disability—$Pr(\widetilde{DD}_{igkrt} < X)$ at each $X \in \{1, \ldots, 100\}$—averaged within each wage tercile group crossed with actual DI award category.

[8] For their main estimate of the fraction of US DI applicants whose awards could be changed by assessor assignment, Maestas et al. (2013) used the difference in award probability that would arise from the full range of difference in assessor FEs (from most stringent to most lenient). They acknowledged (their footnote 43) an order of magnitude smaller estimate of the fraction whose DI awards would be changed by eliminating assessor variation in award rates.

TABLE 3 Actual and counterfactual disability insurance awards by wage tercile.

	No benefit (DD<35%) %	Counterfactual award partial benefit (DD = 35-79%) %	Full benefit (DD≥80%) %	Number of applicants
Panel A) Random assignment to an assessor				
Actual award				
No benefit				
Low wage	100.00	0.00	0.00	61,260
Middle wage	97.70	2.30	0.00	54,429
High wage	87.05	12.95	0.00	25,868
Partial benefit				
Low wage	7.15	92.85	0.00	9,147
Middle wage	0.00	100.00	0.00	18,863
High wage	0.00	100.00	0.00	43,556
Full benefit				
Low wage	0.00	7.07	92.93	25,508
Middle wage	0.00	1.11	98.89	24,029
High wage	0.00	0.00	100.00	25,530
Panel B) All assigned to most lenient assessor				
Actual award				
No benefit				
Low wage	67.01	32.97	0.01	61,260
Middle wage	32.60	67.40	0.00	54,429
High wage	9.77	90.23	0.00	25,868
Partial benefit				
Low wage	0.00	66.48	33.52	9,147
Middle wage	0.00	62.79	37.21	18,863
High wage	0.00	68.43	31.57	43,556
Panel C) All assigned to most stringent assessor				
Actual award				
Partial benefit				
Low wage	78.29	21.71	0.00	9,147
Middle wage	46.37	53.63	0.00	18,863
High wage	39.40	60.60	0.00	43,556
Full benefit				
Low wage	0.00	86.37	13.63	25,508
Middle wage	0.00	66.72	33.28	24,029
High wage	0.00	30.78	69.22	25,530

Notes: In panel A, each cell gives the average percent chance of a degree of disability (DD) in the interval indicated by the respective column heading. The percent chance is obtained for each applicant from the cumulative distribution of $\widehat{DD}_{igkrt}$, which is calculated using eq. (4) for all assessors (k). The individual level percent chances are then averaged over all applicants in the same actual DI award category and wage tercile group indicated by the row heading. In panel B, each cell gives the row percent of applicants with $\widehat{DD}_{igkrt}$ calculated from eq. (4) with $\hat{\lambda}_k^g = max(\hat{\lambda}_1^g, \ldots, \hat{\lambda}_K^g)$ in the interval given by the respective column heading. Panel C cells are constructed analogously with $\hat{\lambda}_k^g = min(\hat{\lambda}_1^g, \ldots, \hat{\lambda}_K^g)$. Low wage, middle wage, and high wage are applicants in the bottom, middle, and top third of the pre-disability wage distribution, respectively. DD = degree of disability. The extreme right-hand column gives the number of applicants in each DI category-wage group.

to measure and codify. Giving assessors discretion makes use of additional information they can glean from applicants at the inevitable cost of inconsistency and horizontal inequity. We find that even in a largely rule-based DI program there is still variation in awards resulting from discretion assessors are given to deal with heterogeneity in the impact of disability on potential earnings. Inconsistency across assessors is a price paid to reduce unfairness that would arise from strict application of rules without sensitivity to the particularities of each case, including those arising from temporal variation in the correspondence between system-generated job matches and vacancies in the prevailing job market.

Between assessor variation in the exercise of discretion exposes applicants to welfare-depleting uninsured risk over benefit receipt. However, reducing this risk by making the system even more rule-based, with less scope for the exercise of discretion, would not be costless. Welfare gains from reducing inconsistencies in awards would have to be set against losses from not exploiting all the relevant (soft) information that assessors have available to make a nuanced assessment of earnings capacity given the applicant's health condition, skills, and experience, as well as current labor market conditions.

Our analysis reveals that judgements do not only generate random between-assessor variation in award propensities—system noise (Kahneman et al., 2021)—but can also produce systematic differences in awards across applicants distinguished by labor market characteristics. Discretion is more likely to be exercised in favor of lower-wage applicants relative to higher-wage applicants. Discontinuous drops in pre-disability wages just above entitlement thresholds and greater differences between potential and final disability degrees reflect upward jumps in the fraction of lower-wage applicants who qualify for higher benefits and in the fraction of higher-wage applicants who are prevented from reaching benefit entitlement. This is contrary to what would be expected given that, all else equal, benefit entitlement is an increasing function of the pre-disability wage. It occurs because assessors are more likely to use their discretion to deem jobs infeasible when they are evaluating lower-wage applicants and to refrain from doing so for higher-wage applicants. While this selective exercise of discretion benefits lower-waged applicants, on average, these applicants are exposed to greater uninsured risk—upside as well as downside—due to larger between assessor variation in awards and the impact that small changes in the job composition can have on the degree of disability of lower-wage applicants.

Selective exercise of discretion in favor of lower-wage relative to higher-wage DI applicants may arise from claim assessors—the street-level bureaucrats (Lipsky, 2010) who implement disability insurance policy—acting with a sense of fairness that conflicts with the insurance principle of the DI program. It can also result from assessors drawing on their experience and softer information they are able to glean from the applicants to complement the estimates of earnings capacity generated by feeding hard data into the algorithm.

We reported preliminary findings of this study to a non-random selection of occupational assessors. In follow-up unstructured discussions that did not follow a formal study protocol, the assessors confirmed that they find it more difficult to assess lower-wage applicants close to entitlement thresholds and are more likely to review the feasibility of algorithm-generated job matches in such cases. They argued that it is easier to find feasible jobs for higher-wage applicants because they are more skilled and less functionally impaired at any given degree of disability. They regarded the review of cases near thresholds to be part of their job and believed this makes the system fairer. While these somewhat anecdotal reactions should be interpreted cautiously, they are consistent with findings from a more formal qualitative study that concludes that occupational assessors consider it their responsibility to exclude infeasible job matches when the consequences of not doing so are most substantial, which is more likely to be the case for lower-wage applicants (de Jong et al., 2013).

One study limitation is that we cannot distinguish the scenario of assessors using discretion selectively to correct inaccuracies that would particularly disadvantage lower-wage applicants from the scenario of assessors perceiving unfairness in unequal entitlements that arise from differences in pre-disability earnings. Our results are consistent with both motivations for the selective exercise of discretion.

Welfare implications can be hypothesized but not confirmed. If lower-wage applicants value DI benefits more, possibly because they are exposed to more uninsured non-health risks related to their disadvantaged labor market position (Deshpande & Lockwood, 2022), then the selective discretion exercised by assessors could increase the social value of the program. While the relatively high social safety net in the Netherlands would be expected to limit the scope for DI to cover otherwise uninsured non-health risks, denial of program benefits to lower-wage applicants can still result in substantial relative income losses, particularly for those without working partners. On the other hand, systematic restraint in the exercise of discretion in favor of higher-wage applicants undermines achievement of the program's objective to deliver insurance against disability-related loss of earnings to all workers. Assessors are given discretion because blind adherence to a guideline—to use the three highest-paying job matches the algorithm finds using objective but incomplete information—would produce imperfect assessment of the heterogeneous impact of disability on earnings capacity. If the opportunity to correct such inaccuracy is utilized less for higher-wage applicants, then they will be less fully insured. There would be a welfare loss in comparison with a system that used discretion to provide more effective coverage also of higher-wage workers. The overall welfare impact of the selective exercise of discretion depends on the magnitude of any gain to lower-wage workers through greater insurance of non-health risks compared with the magnitude of any loss to higher-wage workers through less insurance of health risks.

ORCID

Pilar Garcia-Gomez https://orcid.org/0000-0002-5634-4609
Pierre Koning https://orcid.org/0000-0002-8808-9497
Owen O'Donnell https://orcid.org/0000-0002-6289-1924
Carlos Riumalló-Herl https://orcid.org/0000-0001-6268-0759

REFERENCES

Autor, D. H. (2015). The unsustainable rise of the disabilitty rolls in the United States: Causes, consequences and policy options. In J. K. Sholz, H. Moon, & S.-H. Lee (Eds.), *Social policies in an age of austerity* (pp. 107–136). Edward Elgar Publishing.

Bakx, P., Wouterse, B., Van Doorslaer, E., & Wong, A. (2020). Better off at home? Effects of nursing home eligibility on costs, hospitalizations and survival. *Journal of Health Economics*, *73*, 102354. https://doi.org/10.1016/j.jhealeco.2020.102354

Benitez-Silva, H., Buchinsky, M., & Rust, J. (2004). *How large are the classification errors in the social security disability award process?* [Working paper 10219]. National Bureau of Economic Research. https://doi.org/10.3386/w10219

Bhuller, M., Dahl, G. B., Løken, K. V., & Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, *128*(4), 1269–1324. https://doi.org/10.1086/705330

Burkhauser, R., Daly, M., & de Jong, P. (2008). Curing the Dutch disease: Lessons for United States disability policy. *SSRN*. http://dx.doi.org/10.2139/ssrn.1337652

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210. https://doi.org/10.1093/ectj/utz022

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, *18*(1), 234–261. https://doi.org/10.1177/1536867X180180011

Chan, D. C., Gentzkow, M., & Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, *137*(2), 729–783. https://doi.org/10.1093/qje/qjab048

Dahl, G. B., Kostøl, A. R., & Mogstad, M. (2014). Family welfare cultures. *The Quarterly Journal of Economics*, *129*(4), 1711–1752. https://doi.org/10.1093/qje/qju019

de Jong, P., Everhardt, T., & Schrijvershof, C. (2013). *Duurzaam niet-duurzaam? Onderzoek naar niet-duurzaam volledig arbeidsongeschikt verklaarden*. https://www.eumonitor.eu/9353000/1/j4nvgs5kjg27kof_j9vvik7m1c3gyx/vjanel7lp6pv/f=/blg232604.pdf

Deshpande, M., & Lockwood, L. M. (2022). Beyond health: Nonhealth risk and the value of disability insurance. *Econometrica*, *90*(4), 1781–1810. https://doi.org/10.3982/ECTA19668

Deursen, C. V., Koning, P., García Gomez, P., & Riumallo Herl, C. (2019). *Op de drempel van arbeidsongeschiktheid. De gevolgen voor werk en inkomen van wel of geen uitkering*. https://www.arbeidsdeskundigen.nl/kennis/document/akc/1688

Dobbie, W., Goldin, J., & Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, *108*(2), 201–240. https://doi.org/10.1257/aer.20161503

Doyle, J. J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, *97*(5), 1583–1610. https://doi.org/10.1257/aer.97.5.1583

Doyle, J. J., Ewer, S. M., & Wagner, T. H. (2010). Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of Health Economics*, *29*(6), 866–882. https://doi.org/10.1016/j.jhealeco.2010.08.004

Doyle, J. J., Graves, J. A., Gruber, J., & Kleiner, S. A. (2015). Measuring returns to hospital care: Evidence from ambulance referral patterns. *Journal of Political Economy*, *123*(1), 170–214. https://doi.org/10.1086/677756

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66* (Vol. 66). CRC Press. https://doi.org/10.1201/9780203748725

Figlio, D. N., & Lucas, M. E. (2004). Do high grading standards affect student performance? *Journal of Public Economics*, *88*(9-10), 1815–1834. https://doi.org/10.1016/S0047-2727(03)00039-2

Frandsen, B., Lefgren, L., & Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, *113*(1), 253–277. https://doi.org/10.1257/aer.20201860

French, E., & Song, J. (2014). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy*, *6*(2), 291–337. https://doi.org/10.1257/pol.6.2.291

Godard, M., Koning, P., & Lindeboom, M. (2022). Application and award responses to stricter screening in disability insurance. Journal of Human Resources, 1120-11323R11321. https://doi.org/10.3368/jhr.1120-11323R1

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Hachette UK.

Katz, S. (1983). Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, *31*(12), 721–727. https://doi.org/10.1111/j.1532-5415.1983.tb03391.x

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, *96*(3), 863–876. https://doi.org/10.1257/aer.96.3.863

Koning, P., & Lindeboom, M. (2015). The rise and fall of disability insurance enrollment in the Netherlands. *Journal of Economic Perspectives*, *29*(2), 151–172. https://doi.org/10.1257/jep.29.2.151

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

Low, H., & Pistaferri, L. (2015). Disability insurance and the dynamics of the incentive insurance trade-off. *American Economic Review*, *105*(10), 2986–3029. https://doi.org/10.1257/aer.20110108

Low, H., & Pistaferri, L. (2019). *Disability insurance: Error rates and gender differences* [Working paper 26513]. National Bureau of Economic Research. https://doi.org/10.3386/w26513

Maestas, N. (2019). Identifying work capacity and promoting work: A strategy for modernizing the SSDI program. *The ANNALS of the American Academy of Political and Social Science*, *686*(1), 93–120. https://doi.org/10.1177/0002716219882354

Maestas, N., Mullen, K., & Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review*, *103*(5), 1797–1829. https://doi.org/10.1257/aer.103.5.1797

Maestas, N., Mullen, K. J., & Ravesteijn, B. (2021). *Applying aspects of disability determination methods from the Netherlands in the US* [RDRC Center Papers]. National Bureau of Economic Research. https://www.nber.org/sites/default/files/2022-01/NB21-08%20Maestas%20Mullen%20Ravesteijn.pdf

Nagi, S. (1969). *Disability and rehabilitation: Legal, clinical, and self-concepts and measurement*. Ohio State University Press.

Uitvoeringsinstituut Werknemersverzekeringen [UWV]. (2013). *Basisinformatie CBBS*. https://www.ndsz.nl/content/p1-506432

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341. https://www.jstor.org/stable/2692171

Wachter, T. v., Song, J., & Manchester, J. (2011). Trends in employment and earnings of allowed and rejected applicants to the Social Security Disability insurance program. *American Economic Review*, *101*(7), 3308–3329. https://doi.org/10.1257/aer.101.7.3308

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Garcia-Gomez, P., Koning, P., O'Donnell, O., & Riumalló-Herl, C. (2025). Selective exercise of discretion in disability insurance awards. *Journal of Policy Analysis and Management*, *44,* 816–835. https://doi.org/10.1002/pam.22560

AUTHOR BIOGRAPHIES

Pilar Garcia-Gomez is a Professor in the Erasmus School of Economics at Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (email: garciagomez@ese.eur.nl).

Pierre Koning is a Professor in the School of Business and Economics, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands (email: p.w.c.koning@vu.nl).

Owen O'Donnell is a Professor in the Erasmus School of Economics and Erasmus School of Health Policy and Management at Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (email: odonnell@ese.eur.nl).

Carlos Riumalló-Herl is an Assistant Professor in the Erasmus School of Economics at Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (email: riumalloherl@ese.eur.nl).