Appendix 2: Determining overall generalizability [posted as supplied by author]

To avoid a global rating of generalizability based on a rule of thumb estimate with consensus by discussion, we decided on a sequential analytic approach from medical decision making, the Multi-Attribute Utility (MAU) model by Edwards(1) that would make the implicit judgements of each reviewer explicit. Using this approach, three reviewers (JüBa, JaBu, WdB) individually allotted weights to each criterion. Criteria that are more important should get higher weights than less important criteria. The weights could take values between zero (i.e. item of no importance) to one (maximum importance), and the sum of the weights must not exceed one. The allocated weights ranged between 0.05 and 0.2 (two reviewers), resp. 0.05 and 0.25 (one reviewer). Then, ordinal utilities were attributed to the values of each criterion and multiplied by the weight allocated to this criterion. Finally, the products on the eight criteria were summed up to a final value where higher values indicated higher generalizability.

To classify the studies according to their level of generalizability, we divided the range from maximum to minimum possible score (i.e. from 5 to 0) into four equidistant levels: 'generalizable' (5.0 to 3.75), 'probably generalizable' (3.70 to 2.5), 'probably not generalizable' (2.4 to 1.25), 'not generalizable' (1.20 to 0). To explore to what extent the generalizability rankings of the reviewers differed from each other, we calculated the reviewers' concordance in ranking using Kendall's W (coefficient of concordance) which can take on values between zero and one. Kendall's W on the reviewers' concordance in ranking generalizability was 0.93 with a rank correlation of 0.89, which confirmed very high agreement among the raters' rankings (p=0.009).

Based on the very high concordance, we determined the final level of generalizability for each study by allocating the studies of each reviewer to the corresponding levels. Since all studies ended either in the same level (i.e., all three reviewers rated the same study as 'generalizable') or in adjacent levels (i.e. the same study was rated as 'generalizable' by two reviewers and 'probably generalizable' by one reviewer), a minimum of two ratings with the same level determined the final level. In addition, we marked those studies, which did not achieve unanimous ratings across reviewers (see table below).

Final judgement on overall generalizability based on the rankings of the three reviewers (insurance setting only)

Author, year	Reviewer 1	Reviewer 2	Reviewer 3	Generalizability of global disability rating to real world assessments in insurance medicine
Ingravallo, 2008(2)	T	T	T	Yes
Spanjer, 2008(2)	T	T	T	Yes
Spanjer, 2009(3)	T	T	T	Yes
Dell-Kuster, 2014(4)	T	T	T	Yes
de Kort, 1992(5)	T	T	pT	Yes
Spanjer, 2010(6)	T	pT	T	Yes
Lax, 2004(7)	T	pT	T	Yes
Lederer, 1998(8)	pT	pT	pT	Probably yes
Ikezawa, 2010(9)	pnT	pT	pT	Probably yes
Okpaku, 1994(10)	pnT	pT	pT	Probably yes
Dickmann, 2007(11)	pnT	pnT	pT	Probably no
Schellart, 2013(12)	pnT	pnT	pT	Probably no
Elder, 1994(13)	pnT	pnT	pnT	Probably no
Rudbeck, 2011(14)	pnT	pnT	pnT	Probably no
Schreuder, 2012(15)	pnT	pnT	pnT	Probably no
Slebus, 2010(16)	pnT	pnT	pnT	Probably no

Legend Generalizability: T= generalizable; pT = probably generalizable; pnT = probably not generalizable; nT = not generalizable.

1. Edwards W. The theory of decision making. Psychological bulletin. 1954;51(4):380-417.
2. Ingravallo F, Vignatelli L, Brini M, Brugaletta C, Franceschini C, Lugaresi F, et al. Medico-legal assessment of disability in narcolepsy: an interobserver reliability study. Journal of sleep research. 2008;17(1):111-9.
3. Spanjer J, Krol B, Popping R, Groothoff JW, Brouwer S. Disability assessment interview: the role of detailed information on functioning in addition to medical history-taking. Journal of rehabilitation medicine. 2009;41(4):267-72.
4. Dell-Kuster S, Lauper S, Koehler J, Zwimpfer J, Altermatt B, Zwimpfer T, et al. Assessing work ability--a cross-sectional study of interrater agreement between disability claimants, treating physicians, and medical experts. Scandinavian journal of work, environment & health. 2014;40(5):493-501.
5. de Kort WL, Uiterweer HW, van Dijk FJ. Agreement on medical fitness for a job. Scandinavian journal of work, environment & health. 1992;18(4):246-51.
6. Spanjer J, Krol B, Brouwer S, Popping R, Groothoff JW, van der Klink JJ. Reliability and validity of the Disability Assessment Structured Interview (DASI): a tool for assessing functional limitations in claimants. Journal of occupational rehabilitation. 2010;20(1):33-40.
7. Lax MB, Manetti FA, Klein RA. Medical evaluation of work-related illness: evaluations by a treating occupational medicine specialist and by independent medical examiners compared. Int J Occup Environ Health. 2004;10(1):1-12.
8. Lederer P, Pfaff G, Walter K, Weihrauch M, Weber A. [Quality circles in expert assessment as an instrument in quality management]. Gesundheitswesen. 1998;60(7):415-9.
9. Ikezawa Y, Battie MC, Beach J, Gross D. Do clinicians working within the same context make consistent return-to-work recommendations? Journal of occupational rehabilitation. 2010;20(3):367-77.
10. Okpaku SO, Sibulkin AE, Schenzler C. Disability determinations for adults with mental disorders: Social Security Administration vs independent judgments. American journal of public health. 1994;84(11):1791-5.
11. Dickmann JR, Broocks A. [Psychiatric expert opinion in case of early retirement--how reliable?]. Fortschr Neurol Psychiatr. 2007;75(7):397-401.
12. Schellart AJ, Zwerver F, Anema JR, Van der Beek AJ. The influence of applying insurance medicine guidelines for depression on disability assessments. BMC research notes. 2013;6:225.
13. Elder AG, Symington IS, Symington EH. Do occupational physicians agree about ill-health retiral? A study of simulated retirement assessments. Occupational medicine. 1994;44(5):231-5.
14. Rudbeck M, Fonager K. Agreement between medical expert assessments in social medicine. Scand J Public Health. 2011;39(7):766-72.
15. Schreuder JA, Roelen CA, de Boer M, Brouwer S, Groothoff JW. Inter-physician agreement on the readiness of sick-listed employees to return to work. Disabil Rehabil. 2012;34(21):1814-9.
16. Slebus FG, Kuijer PP, Willems JH, Frings-Dresen MH, Sluiter JK. Work ability assessment in prolonged depressive illness. Occupational medicine. 2010;60(4):307-9.