Inter-rater agreement in evaluation of disability: systematic review of reproducibility studies

Jürgen Barth,1,2 Wout E de Boer,1 Jason W Busse,3,4,5 Jan L Hoving,6,7 Sarah Kedzia,1 Rachel Couban,4 Katrin Fischer,8 David Y von Allmen,1 Jerry Spanjer,5,10 Regina Kunz1

ABSTRACT

OBJECTIVES
To explore agreement among healthcare professionals assessing eligibility for work disability benefits.

DESIGN
Systematic review and narrative synthesis of reproducibility studies.

DATA SOURCES
Medline, Embase, and PsycINFO searched up to 16 March 2016, without language restrictions, and review of bibliographies of included studies.

ELIGIBILITY CRITERIA
Observational studies investigating reproducibility among healthcare professionals performing disability evaluations using a global rating of working capacity and reporting inter-rater reliability by a statistical measure or descriptively. Studies could be conducted in insurance settings, where decisions on ability to work include normative judgments based on legal considerations, or in research settings, where decisions on ability to work disregard normative considerations.

Teams of paired reviewers identified eligible studies, appraised their methodological quality and generalisability, and abstracted results with pretested forms. As heterogeneity of research designs and findings impeded a quantitative analysis, a descriptive synthesis stratified by setting (insurance or research) was performed.

RESULTS
From 4562 references, 101 full text articles were reviewed. Of these, 16 studies conducted in an insurance setting and seven in a research setting, performed in 12 countries, met the inclusion criteria. Studies in the insurance setting were conducted with medical experts assessing claimants who were actual disability claimants or played by actors, hypothetical cases, or short written scenarios. Conditions were mental (n=6, 38%), musculoskeletal (n=4, 25%), or mixed (n=6, 38%). Applicability of findings from studies conducted in an insurance setting to real life evaluations ranged from generalisable (n=7, 44%) and probably generalisable (n=3, 19%) to probably not generalisable (n=6, 37%). Median inter-rater reliability among experts was 0.45 (range intraclass correlation coefficient 0.86 to κ−0.10). Inter-rater reliability was poor in six studies (37%) and excellent in only two (13%). This contrasts with studies conducted in the research setting, where the median inter-rater reliability was 0.76 (range 0.91-0.53), and 71% (5/7) studies achieved excellent inter-rater reliability. Reliability between assessing professionals was higher when the evaluation was guided by a standardised instrument (23 studies, P=0.006). No such association was detected for subjective or chronic health conditions or the studies’ generalisability to real world evaluation of disability (P=0.46, 0.45, and 0.65, respectively).

CONCLUSIONS
Despite their common use and far reaching consequences for workers claiming disabling injury or illness, research on the reliability of medical evaluations of disability for work is limited and indicates high variation in judgments among assessing professionals. Standardising the evaluation process could improve reliability. Development and testing of instruments and structured approaches to improve reliability in evaluation of disability are urgently needed.

WHAT IS ALREADY KNOWN ON THIS TOPIC
Social and private disability insurers use medical experts to evaluate claimants with impaired health to determine eligibility for disability benefits
Anecdotal evidence suggests that experts often disagree in their judgment of capacity to work when assessing the same claimant

WHAT THIS STUDY ADDS
This systematic review of 23 reproducibility studies from 12 countries shows a lack of good quality data applicable to the real world of disability assessment
In most studies, medical experts reached only low to moderate reproducibility in their judgment of capacity to work
Studies reported higher reproducibility when experts used a standardised evaluation procedure
These findings are disconcerting and call for substantial investment in research to improve assessment of disability

Introduction
Many workers seek wage replacement benefits on the basis of disabling illness or injury, and over the past decade most countries of the Organisation for Economic Co-operation and Development (OECD) have experienced escalating rates of affected workers.12 Current estimates range from four to eight individuals per thousand per year,2 corresponding to 16 000 newly affected workers/year for smaller countries like Switzerland and 170 000/year for countries like the US.
Both public and private insurance systems provide wage replacement benefits for employees whose impaired health prevents them from working, as long as eligibility criteria are met.1 To inform this decision, insurers often arrange for evaluation of disability claims by medical professionals.3,5 Based on these evaluations, about half of all disability claims are declined.2
Equality before the law requires that claimants with similar health impairments and exposed to similar work demands should receive similar judgments of medical restrictions and limitations. Concerns have been raised, however, regarding low quality evaluations and poor reliability between medical experts. Evaluation of disability is a complex process that is affected by the skillset, attitudes, and beliefs of the expert, and few countries enforce standards of practice, which presents considerable challenges to reliability (see box 1). We conducted a systematic review of reproducibility studies to summarise empirical evidence regarding the inter-rater reliability of global judgments on work disability and examined the hypothesis that studies using standardised assessments would show higher reliability.

Methods
We followed the standards set by the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) and Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) for the reporting of our study.

Eligibility criteria
We included reproducibility studies conducted in an insurance setting (evaluation of claimants) or in a research setting (evaluation of patients for work disability outside of actual assessments) in which two or more health professionals evaluated the work capacity of individuals claiming disability and reported inter-rater reliability on a global rating of work disability. Studies that reported only the inter-rater reliability of experts' evaluation of specific physical or mental activities (such as lifting, conflict management) were excluded. All types of “subjects” qualified: real claimants, records of claimants, videotaped actors, vignettes, short case summaries.

Search strategy
We searched Medline, Embase, and PsycInfo from inception to 16 March 2016, without language restrictions. An experienced medical librarian (RC) developed database specific search strategies combining the following subject terms: reproducibility of results (MeSH, including reliability) and reliability statistics, disability or work capacity evaluation, and sick leave (see appendix 1 for the detailed search strategy). We screened the bibliographies of all included studies for additional relevant articles.

Study process
Three teams of paired reviewers (WdB, JWB, JH, SK, JS, RK) with expertise in medical evaluations and training in research methodology independently screened titles, abstracts, and full texts for eligibility, assessed generalisability, and collected data from each eligible study using standardised pilot tested forms with detailed instructions. Reviewers resolved disagreement through discussion or, if required, adjudication by a third reviewer (RK or WdB).

Quality appraisal
Quality appraisal of reproducibility studies includes methodological quality and generalisability to the setting in which the instrument will be used. To address the former, we assessed the blinding of raters to each other’s findings, the risk of order effects, and appropriateness of the statistical analyses following Quality Appraisal for Reliability Studies (QAREL) guidance. To address generalisability, we evaluated whether claimants, raters, and the performance of the disability evaluation were similar to the insurance context in which such evaluations take place.

As reliability is a product of the interaction between the performance of the test, the subjects/objects, and the context of the assessment, and as its estimate is affected by various sources of the variability in the measurement setting (that is, rater and subject characteristics, performance of the test, box 1), we used an explicit and transparent process to evaluate generalisability. Based on the checklist of QAREL, GRRAS, and expert guidance, we identified four claimant items and four expert items for defining greater generalisability:

- The recruitment strategy captures diverse cases as would present in actual evaluation of disability (in declining order: random, consecutive, other recruitment; not applicable to written cases or videos)
- Recruitment success (in declining order: >80%, 80-50%, <50%; not applicable to records of patients, videos, or written cases)
- Verisimilitude—that is, the extent that cases reflect the population in real life (in declining order: real claimants, including videotapes/audiotapes of real
We calculated the reviewers’ concordance in generalisability (for example, “generalisable,” “probably generalisable,” or “not generalisable”). This approach allows ranking using Kendall’s W (coefficient of concordance) and one (perfect agreement) to the measurement error.

Five reviewers (JB, RK, WdB, JS, JanHo), blinded to the study results, assessed generalisability of each study, independently and in duplicate. Given the lack of empirical evidence about the relative importance of each item we used a sequential approach from medical decision making to make the weighting of each item explicit (see appendix 2 for detailed description). This approach facilitated judgments regarding overall generalisability (that is, “generalisable,” “probably generalisable,” “probably not generalisable,” and “not generalisable”).

We calculated the reviewers’ concordance in generalisability ranking using Kendall’s W (coefficient of concordance), which generates values between zero (no agreement) and one (perfect agreement).

We limited assessment of generalisability to studies performed in an insurance setting because studies conducted in a research setting, by definition (“normative or legal considerations not part of the judgment”, see data analysis), lack generalisability to real life assessments of disability.

Data collection
We extracted the following information from each eligible study:

- **Study context**—background and setting (insurance, rehabilitation, research)
- **Patients’ characteristics** (“cases”)—number of cases per study; presenting disorder(s) (mental disorder, musculoskeletal disease, mixed); course of disease or injury (acute, chronic)
- **Expert characteristics** (“raters”)—number of raters per study; number of cases per rater; number of raters per case; profession (primary or secondary care, occupational physician, insurance physician)
- **Procedures**—time frame before the evaluation for judging current health status and work disability; time frame for predicting global work disability (for both time frames, short term refers to less than six months; long term refers to more than six months; mixed); instrument (professional expertise with or without specific rating instrument) to support global rating of work disability and the related categories (for example, fully limited, partially limited, no limitations) or scales (for example, scale 0-100)
- **Outcomes**—global rating of work disability (for example, work capacity, sick leave, readiness for return to work, reduction in working hours); decisions on suitability for a specific job; occupational functioning; measure of reliability or agreement (intraclass correlation coefficient (ICC), κ statistic, or percentage agreement), including measure of precision, or descriptive measure (for example, frequency of judgments).

Data analysis
For three studies that reported on reduction of working hours, we calculated the κ statistics and intraclass correlation, based on the raw data provided by authors.

We distinguished between studies conducted in an insurance setting or a research setting. In an insurance setting, health professionals make judgments on disability for work based on functional limitations that includes normative judgments from a societal perspective. An insurance setting does not imply any specific format of the claimant’s presentation in the study, which can range from a real patient to a written case (see also “generalisability, verisimilitude”). Researchers in a research setting who develop and/or validate instruments tend to standardise their research environment when judging occupational functioning. Normative (legal) considerations or a societal perspective are not part of their judgments.

We used studies conducted in a research setting to investigate the association between level of standardisation in the evaluation process and inter-rater reproducibility. Level of standardisation was considered as “not standardised” when medical experts in the insurance setting used only their professional expertise to elicit information and rate findings from the claimant; as “semi-standardised” when they used a structured instrument as one component of the evaluation; and as “fully standardised” when occupational functioning was primarily evaluated with a structured instrument.

Lack of information on variation associated with reproducibility statistics and heterogeneity of statistical measures and outcomes precluded pooling of the data across studies. Using a two tailed Fisher’s exact test, we explored whether objective (versus subjective) and acute (versus chronic) health conditions as well as higher levels of generalisability and/or higher levels of standardisation in the evaluation process were associated with a higher inter-rater reproducibility. We defined mental disorders as “subjective complaints” and somatic disorders as objective complaints, though we acknowledge the crude nature of this classification, and acute conditions shorter than six months and chronic
conditions longer than six months. We excluded from our analysis three studies that did not specify the chronicity. Fisher’s exact test does not provide a test statistic, only whether the difference is significant or not.

For clinical interpretation of reliability measures, we used the thresholds established by Fleiss in 1981\(^2\) to distinguish between poor, fair, good, and excellent inter-rater reliability.\(^{26-28}\) For \(\kappa\), weighted \(\kappa\), and intraclass correlation, the cut-off levels were <0.40 (poor), 0.40-0.59 (fair), 0.60-0.74 (good), and 0.75-1.00 (excellent); for percentage agreement, the levels were <70% (poor), 70-79% (fair), 80-89% (good), and 90-100% (excellent), taking into account that percentage agreement does not account for an agreement of raters by chance. Biometricians acknowledge that these guidelines are broadly accurate with some arbitrariness. Though at times they might come up with conflicting results, they have proved valuable in clinical application.\(^{24}\)

**Patient involvement**

No patients were involved in setting the research question, in developing plans for design, interpretation, reporting or implementation of the study. We plan to disseminate the results of this study to organisations supporting patients with disabilities.

**Results**

**Study characteristics**

Of 4562 potentially relevant citations identified, 101 reports proved potentially eligible after we had screened titles and abstracts. On full text screening, 23 studies,\(^9\) 11 22-26 29-46 including four non-English studies,\(^9\) 39-41 proved eligible for analysis (fig 1). All studies were published from 1992 onwards and enrolled disability claimants from 12 countries in Europe, North America, Australia, the Middle East, and northeast Asia. Seven studies (30%) were conducted in the Netherlands. Seventy percent of the studies (16/23) were conducted in an insurance setting, with the remainder in a research setting. Investigators used a broad spectrum of designs, ranging from real life disability evaluations, videotapes with actors, and records of claimants to 10 line case vignettes, to perform reliability studies. Study size varied considerably with number of raters from two to 103 and number of patients from one and 3562 per study (tables 1 and 2).

**Methodological quality and generalisability**

Assessment of methodological quality included blinding of raters to each other’s findings, presence of order effects, and appropriateness of the statistical analyses (table 3; appendix 2). The studies on the reproducibility between medical experts conducted in an insurance setting met 80% (31/39) of these items, 15% (6/39) remained unclear, and 5% (2/39) were not applicable. The methodological quality items did not fit the median of the studies that looked at the reproducibility between medical experts and health professionals. Studies conducted in a research setting met 52% (11/21) of the quality items; 33% (7/21) remained unclear and 14% (3/21) were not met (table 3).

With regards to generalisability of the findings to real life disability evaluation, 44% (7/16) of studies in the insurance setting were rated as “generalisable,” 19% (3/16) as “probably generalisable,” and 37% (6/16) as “probably not generalisable” (table 4). Kendall’s \(W\) for reviewers’ concordance in ranking generalisability was 0.93, with a rank correlation of 0.89, confirming high agreement among the raters’ rankings.

**Studies conducted in insurance setting**

In the insurance setting, 13 studies including 463 patients and 367 raters explored agreement between medical experts (two or more experts assessing the same patient) (table 1; appendix 4).\(^9\) 11 22-26 32 34 37 39 43-46 Three studies including 3729 patients (with 3562 patients from a single centre\(^33\)) and eight raters (information was lacking from one study\(^33\)) explored agreement between medical experts and claimant’s treating physicians\(^33\) or independent rehabilitation or occupational health teams with a mandate to care.\(^{26-42}\) The median number of patients per study was 13.5 (range 1-3562), and the median number of raters per study was 12 (2-103, excluding one study that did not report the number of raters\(^33\)). All but three studies\(^24\) 34-42 used a fully crossed design (that is, all raters assessed a single patient).\(^34\) 36-42 In the insurance setting, 13 studies including 463 patients and 367 raters explored agreement between medical experts (two or more experts assessing the same patient) (table 1; appendix 4).\(^9\) 11 22-26 32 34 37 39 43-46 Three studies including 3729 patients (with 3562 patients from a single centre\(^33\)) and eight raters (information was lacking from one study\(^33\)) explored agreement between medical experts and claimant’s treating physicians\(^33\) or independent rehabilitation or occupational health teams with a mandate to care.\(^{26-42}\) The median number of patients per study was 13.5 (range 1-3562), and the median number of raters per study was 12 (2-103, excluding one study that did not report the number of raters\(^33\)). All but three studies\(^24\) 34-42 used a fully crossed design (that is, all raters assessed a single patient).\(^34\) 36-42

Table 5 summarises claimants’ characteristics. Studies focused on mental health (n=6), musculoskeletal disease (n=4), and mixed disorders (n=6). They enrolled patients with chronic diseases (n=11), chronic injuries (n=2), or mixed, acute, and chronic conditions (n=3). Most referred to a long term time frame before the evaluation for judging health status and work disability and predicted a long term perspective exceeding six months. Most studies used professional expertise only to generate a global rating of work ability (n=10). Six administered one or more specific rating instruments; five were referenced (appendix 3), and none was reported as validated.

---

**Fig 1** | Identification of studies assessing inter-rater agreement of evaluation of disability

Records identified through database searching (Medline, Embase, PsycINFO) March 2016 (n=4562)

Records screened after removal of duplicates (n=4562)

Excluded (n=3718)

Full text articles assessed for eligibility (n=101)

Excluded (n=78):

- No work disability (n=55)
- Rater not qualified (n=9)
- Outcome not appropriate (such as functional capacity) (n=7)
- Statistical information insufficient (n=6)
- Other reasons (n=1)

Studies included in qualitative synthesis (n=23)
| Study ID | Country | Sample Size | Methodology | Disease | Context of Study | Retrospective P/F | Prospective P/F | Disease Course | Outcome | Follow-up | Quality | Comments |
|---------|---------|-------------|-------------|---------|-----------------|-------------------|------------------|----------------|---------|-----------|---------|----------|
| de Kort, 1992, Netherlands | 32 | Random selection of 180 applicants from national database of 10175 individuals | Mixed (individuals judged fit and unfit for job/chronic disease) | Medical expert deciding work ability in a job setting | Single centre study on 162 real disability assessments about reliability of medical experts and family | Unclear, not reported/unclear not reported | Unclear, not reported/unclear not reported | Mixed conditions/chronic disease | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Dell-Kuster, 2014, Switzerland | 90 | Applicants (30 from each job category) had been judged (temporarily) unfit and 90 cases were used as reference cases | Mixed conditions/chronic disease | Long term history/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Long term history/long term ≥6 months | Long term history/mix of short term and long term ≥6 months | Long term history/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Elder, 1994, UK | 16 | 4 medical commissions with 4 experts each, re-evaluated 16 patients from database of 150 patients with different diseases | Mixed conditions/chronic disease | Long term history/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Long term history/long term ≥6 months | Long term history/mix of short term and long term ≥6 months | Long term history/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Iype, 2008, Italy | 37 | In survey with 3 case vignettes from real cases, 36 clinicians made return to work recommendations. | Mixed conditions/chronic disease | Long term history/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Long term history/long term ≥6 months | Long term history/mix of short term and long term ≥6 months | Long term history/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Okpaku, 1994, US | 11 | 4 medical commissions with 4 experts each, re-evaluated 16 patients from database of 150 patients with different diseases | Mixed conditions/chronic disease | Long term history/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Long term history/long term ≥6 months | Long term history/mix of short term and long term ≥6 months | Long term history/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Spanjer, 2008, Netherlands | 23 | 12 insurance physicians used disability assessment structured interview reports of 12 claimants with mental or physical disorder. Reports assessed on functional information system and mental ability list | Mental and somatic disorders/chronic disease | Not reported/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Spanjer, 2009, Netherlands | 24 | 16 insurance physicians from Dutch social insurance office trained partly in disability assessment structured interview assessed 62 real cases. Subanalysis of randomised controlled trial on training | Mental and somatic disorders/chronic disease | Not reported/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
| Spanjer, 2010, Netherlands | 25 | 25 insurance physicians compared assessment of work ability in patients with major depressive disorders and chronic diseases. | Mental and somatic disorders/chronic disease | Not reported/long term ≥6 months | Single centre study on 3562 real life disability assessments about reliability of medical experts and family | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Not reported/long term ≥6 months | Long term | Long term | Quality | Included studies on evaluation of disability from insurance setting |
Table 2 | Included studies on evaluation of disability in research setting*

| Study / Country | Context of study | Disease/course | Retrospective time frame/prospective time frame for judgment |
|-----------------|------------------|----------------|-------------------------------------------------------------|
| Berns, 2007, US | 2 practitioners rated 29 of larger sample with bipolar disorders with newly developed multidimensional scale of independent functioning (MSIF). Study conducted in single centre | Mental (bipolar disorders)/chronic disease | Short term/not reported |
| Chopra, 2002, Australia | 2 clinicians assessed feasibility and reliability of international classification of impairments, disability, and handicap (ICIDH-II) in 20 patients with psychotic disorders. Multicentre study | Mental (psychosis)/acute disease | Short term/short term <6 months |
| Daradkeh, 1994, United Arab Emirates | 2 psychiatrists with experience in rating disability assessment schedule (DAS), based on axis V of ICD-10, with one dimension of work) reviewed 42 psychotic patients with different informants (such as family). Single centre study | Mixed mental inpatient and outpatient/mixed | Short term/not reported |
| Hannula, 2006, Finland | Group of researchers developed rating scale based on social adjustment scale with focus on social and occupational functioning (SOAPS). Four clinically trained professionals administered rating scale to 39 videotaped interviews of consecutive patients from Helsinki Psychotherapy Study | Mental: anxiety and mood disorders/mixed | Short term/short term <6 months |
| Hill, 1989, UK | Authors developed adult personality functioning assessment (APFA) with work as one subdomain. 3 raters used APFA for assessment 21 audiotaped standardised interviews with client | Mental/chronic disease | Unclear, not reported, mixed/unclear, not reported, mixed |
| Mundo, 2010, Italy | 18 raters assessed 180 inpatients with Kennedy Axis V (K Axis), which is equivalent to global assessment of functioning (GAF). One subscale covers occupational skills | Mixed mental disorders/unclear/no information | Short term/unclear, not reported, mixed |
| Nozu, 1995, Japan | 3 experts assessed schizophrenic outpatients who started occupational therapy at Tokyo Metropolitan Chubu Comprehensive Mental Health Centre with newly developed Work-Personality Insufficiency Rating Scale | Patients with schizophrenia/chronic disease | Unclear/unclear, not reported, mixed |

*In research setting, researchers who develop instruments tend to standardise their research environment when judging occupational functioning. Normative considerations or societal perspective are not part of their judgments.

Table 3 | Methodological quality of included studies

| Study / Country | Raters blinded to findings of others | Risk for order effect (sequence of examination) | Appropriate statistical measure of agreement |
|-----------------|-------------------------------------|-----------------------------------------------|---------------------------------------------|
| Reproducibility among experts in insurance setting | | | |
| de Kort, 1992 | Yes | No risk | Yes |
| Dickmann, 2007 | Yes | No risk | NA |
| Elder, 1994 | Unclear | No risk | Yes |
| Ikezawa, 2010 | Yes | No risk | Yes |
| Ingravallo, 2008 | Yes | No risk | Yes |
| Lederer, 1998 | Yes | No risk | NA |
| Rudbeck, 2011 | Yes | No risk | Yes |
| Schellart, 2013 | Unclear | Unclear | Yes |
| Schreuder, 2012 | Unclear | No risk | Unclear |
| Schreuder, 2012 | Yes | No risk | Yes |
| Spanjer, 2008 | Yes | No risk | Yes |
| Spanjer, 2009 | Yes | No risk | Yes |
| Spanjer, 2010 | Yes | No risk | Yes |

| Reproducibility among experts and health professionals in insurance setting | | | |
| Dell-Kuster, 2014 | NA | NA | Yes |
| Lax, 2004 | NA | NA | NA |
| Okpaku, 1994 | NA | NA | NA |

| Research setting | | | |
| Berns, 2007 | Unclear | Yes | Yes |
| Chopra, 2002 | Probably yes | No | Yes |
| Daradkeh, 1994 | Unclear | Unclear | Yes |
| Hannula, 2006 | No | Yes | Yes |
| Hill, 1989 | Yes | Unclear | Yes |
| Mundo, 2010 | Yes | Unclear | Yes |
| Nozu, 1995 | Unclear | Unclear | Yes |
Work disability outcomes varied considerably between studies and included a broad spectrum of domains, definitions, and measurement approaches, ranging from work ability to the employee’s readiness and ability to return to work, the degree of disability or handicap, or reduction in working hours. Measurement approaches included scales, scores, and categories (table 6).

Studies conducted in research setting
Studies conducted in a research setting included 371 patients and 32 raters (table 2; appendix 4). Four studies reported on instrument development, and three studies validated existing instruments. The median number of patients per study was 39 (range 20-180), and the median number of raters per study was three (2-18). All but two studies used a fully crossed design, with a median of 21 patients (11-42) per rater and a median of two raters (2-4) per patient.

All studies were conducted with actual patients and focused on acute and chronic mental health conditions. Most used a short term time frame before the evaluation for judging occupational functioning, two provided a long term period and four used a mixed time frame. Most studies used professional expertise only to rate work disability but five studies used at least one rating or reporting instrument.

Table 4 | Generalisability of study findings to real world of insurance medicine*  

| Recruitment strategy (for claimants) | Recruitment success | Verisimilitude | Range of experience in raters | Specific training for work capacity assessment | Training for study purposes | No of cases | No of raters | Generalisability |
|-------------------------------------|--------------------|---------------|-----------------------------|----------------------------------------------|----------------------------|-------------|-------------|----------------|
| de Kort, 199222                      | Random sample      | NA            | Records of real patients    | Narrow                                       | Yes                        | No          | 180         | 5              | Yes            |
| Dei-Kuster, 201423                   | Consecutive sample | >80%          | Real patients               | Wide                                         | No                         | No          | 3562        | Unclear        | Yes            |
| Dickmann, 2007                       | NA                 | NA            | Video case scenario         | Narrow                                       | Yes                        | No          | 1           | 2              | Probably no    |
| Elder, 199424-25                     | NA                 | NA            | Written case scenarios      | Unclear                                      | Yes                        | No          | 10          | 35             | Probably no    |
| Ikezawa, 201026                      | NA                 | NA            | Written case scenarios      | Wide                                         | Yes                        | No          | 3           | 36             | Probably yes   |
| Ingravillo, 200827                   | Random sample      | >80%          | Real patients               | Narrow                                       | Yes                        | Yes         | 15          | 16             | Yes            |
| Lax, 200428-29                      | Random sample      | >80%          | Records of real patient     | Narrow                                       | No                         | 23          | 2           | Yes            |
| Lederer, 199829-31                   | Any other recruitment | NA          | Records of real patients    | Wide                                         | Yes                        | No          | 1           | 103            | Probably yes   |
| Okpaku, 199432                       | Unclear, not reported | NA      | Records of real patients    | Narrow                                       | Yes                        | 144         | 6           | Yes            |
| Rudbeck, 201134                      | NA                 | NA            | Written case scenarios      | Unclear                                      | Yes                        | No          | 8           | 11             | Probably yes   |
| Schellart, 201335                    | NA                 | NA            | Video case with actor       | Wide                                         | Yes                        | Yes         | 4           | 40             | Probably no    |
| Schreuder, 201236                    | NA                 | NA            | Written case scenarios      | Unclear                                      | Yes                        | No          | 132         | 5              | Probably yes   |
| Siebus, 201037                       | NA                 | NA            | Written case scenarios      | Narrow                                       | Yes                        | Mixed, unclear | 5           | 51             | Probably no    |
| Spanjer, 200838                      | Random sample      | NA            | Records of real patient     | Wide                                         | No                         | 12          | 12          | Yes            |
| Spanjer, 200939                       | Random sample      | NA            | Records of real patients    | Narrow                                       | No                         | 30          | 27          | Yes            |
| Spanjer, 201040                      | Any other recruitment | >50-80%       | Real patients               | Narrow                                       | Yes                        | No          | 62          | 16             | Yes            |

NA=not applicable.
*44% of studies rated as generalisable, 19% as probably generalisable, and 37% as probably not generalisable (see appendix 2 for details).

Table 5 | Characteristics of studies investigating eligibility for work disability benefits

| Insurance setting (n=16) | Research setting (n=7) |
|--------------------------|------------------------|
| Health conditions:       |                        |
| Mental disorders         | 38%21-23,34-36,38-41   |
| Musculoskeletal disorders| 25%,26-28,34-36         |
| Mixed (somatic and mental disorders) | 38%,21-23,34-36 |
| Course of disease or injury: |
| Acute diseases           | —                      |
| Acute and chronic diseases| 6%34,26-28,34-36       |
| Chronic diseases         | 75%,21-23,26-28,34-36 |
| Chronic injuries         | 13%,21-23,26-28        |
| No information/unclear   | 6%,21-23,26-28         |
| Composition of patient population:
| Single disorders (such as narcolepsy, stroke, depression, low back pain, psychosis, depression, anxiety, schizophrenia) | 31%,21-23,26-28,34-36 |
| Mixed disorders          | 69%,21-23,26-28,34-36 |
| Reference time frame before evaluation for judgments on health condition: |
| Short term period        | 6%,21-23,26-28,34-36   |
| Long term period         | 69%,21-23,26-28,34-36  |
| Not reported             | 25%,21-23,26-28        |
| Prognostic time frame:   |                        |
| Short term (≤6 months)   | 13%,21-23,26-28,34-36  |
| Long term (≥6 months)    | 6%,21-23,26-28,34-36   |
| Mixed                    | 6%,21-23,26-28,34-36   |
| Not reported             | 13%,21-23,26-28,34-36  |
| Use of tools to facilitate rating of work disability: |
| Professional expertise only | 63%,21-23,26-28,34-36  |
| ≥1 rating or reporting instruments | 37%,21-23,26-28,34-36 |
short term prognostic judgement on occupational functioning, and this information was missing in five studies (table 2). All seven studies used instruments of varying complexity to elicit or to report capacities or limitations to determine a global rating for occupational functioning (appendix 4). All studies generated global ratings on a range of outcomes for occupational functioning, such as “occupational functioning” or “remunerative employment” (table 7).

Inter-rater reliability of ratings on disability for work and occupational functioning—insurance setting

Overall, across all conditions and outcomes, the median inter-rater reliability was 0.45, ranging from ICC of 0.86 (musculoskeletal disorders; reduction in working hours)22 to κ of −0.10 (narcolepsy; disability benefit)13 (table 8). Six studies reported excellent or good inter-rater reliability for a global rating of work disability, with ICCs of 0.6446 and 0.6546 percentage agreement 82.4% (“return-to-work” recommendations37), or κ of 0.8023 and 0.8622 for reduction in working hours. One study presented mixed judgments in a single case, which we considered overall as “good agreement” based on the relative importance of the outcomes of functional ability to work (91.2% agreement on remaining work ability) and for work recommendations (86% agreement in limitations on work performance) over the outcome of readiness and ability to return to work (56% agreement on reduction in working hours).39 All Dutch studies used one or more rating instruments for determining functional limitations,22 23 44 46 Two studies qualified as “generalisable,”21 32 two as “probably generalisable,”37 39 and two as “probably not generalisable.”44 46

Seven studies reported fair or poor inter-rater reliability across all global ratings of work disability outcomes. All but one26 based their judgments exclusively on professional expertise. One study presented discordant judgment on a single case9 (one third of experts each rated “full,” “partial,” or “no work ability” for the same

---

**Table 6 | Outcomes used in insurance setting to assess global rating of disability for work**

| Outcome measure | Quantification |
|-----------------|----------------|
| **Global rating of work ability** (n=5) | Scale from 100-0% |
| Work ability | 3 categories: >6 hours; 6-3 hours; <3 hours |
| Health related work ability | 4 categories: intact or slightly/much/extremely reduced |
| List of functional abilities | Sum score; range not reported |
| Global rating of work ability | Scale from 100% (status as before depressive disorder) to 0% (“inability to work”) |

**Fit for work recommendations (n=3)**

- Global rating of fit for work | Scale from 100% (status as before depressive disorder) to 0% (“inability to work”) |
- Readiness and ability of employee to return to work | 2 categories: high v low |
- Recommend return to work | 3 categories: return to previous work/return to modified work/no return to work |
- Recommend fit for work | 2 categories: yes v no |

**Decisions on disability benefits (n=3)**

- Approval or decline of application for early retirement because of ill health | 4 categories: accept/reject/other action/no response |
- Approval for social security benefit | Social security administration—2 categories: yes v no. Team—4 categories: yes/maybe/no/undecided |

**Degree of disability or handicap (n=2)**

- Severity of handicap | 3 categories: no handicap/handicap/severe handicap |
- Agreement among occupational health professional and medical expert on 4 disability items | 3 categories: full/partial/disagreement |

**Reduction in working hours (n=3)**

- Reduction in working hours | Hours/day |

---

**Table 7 | Outcomes used in research setting to assess global rating of disability for work**

| Outcome measure | Quantification |
|-----------------|----------------|
| **Functioning within work environment; occupational skills** (n=5) | |
| Global rating about functioning within work environment | 7 item Likert scale, 1 (normal functioning)-7 (total disability) |
| Adult personality functioning assessment | 6 point scale, 0-5, higher values indicate worse functioning |
| Occupational functioning | 6 point scale, “no dysfunction” to “maximum dysfunction” |
| Occupational functioning | Scale 100-0, higher values indicate better functioning |
| Occupational skills | Scale 100-0, higher values indicate better occupational skills |
| **Remunerative employment, employability** | |
| Global rating for remunerative employment | 5 item scale: no to complete or extreme problem |
| Employability | No information |
Tables

### Table 8: Reproducibility among experts stratified by level of inter-rater reliability

| Study | Use of rating or reporting instrument | Outcome | Outcome measure and IRR findings | Generalisability to real world disability evaluation |
|-------|---------------------------------------|---------|----------------------------------|-----------------------------------------------|
| **Studies investigating reproducibility of work disability evaluations between experts (insurance setting)** | | | | |
| Schellart39 | Yes | Functional work ability | ICC 0.65 | Probably no |
| Slebus36 | Yes | Functional work ability | ICC 0.64 | Probably no |
| Ikezawa37 | Yes | Recommend return to work | % agreement 82.4% | Probably yes |
| Spanjer 200833 | Yes | Reduction in working hours | κ 0.8 | Yes |
| Spanjer 200932 | Yes | Reduction in working hours | κ 0.86 | Yes |
| Lederer39 | NR | Remaining work ability; limitations in work performance (single case) | Frequency of agreement: 91%; 86% | Probably yes |
| **Fair to poor** | | | | |
| De Kort32 | NR | Fit for work | κ 0.38 | Yes |
| Dickmann7 | NR | Work ability in last job (single case): <3 hours; 3-6 hours; >6 hours | Frequency of agreement: 27%; 36%; 37% | Probably no |
| Elder34 | NR | Early retirement | κ 0.24 | Probably no |
| Ingravallo35 | NR | Disability benefit | κ −0.10−0.35 | Yes |
| Rudbeck43 | NR | Health related work ability | κ 0.33 | Probably no |
| Schreuder45 | NR | Readiness and ability to return to work | κ 0.14 | Probably no |
| Spanjer 201034 | Yes | Reduction in working hours | ICC 0.53 | Yes |
| **Studies investigating reproducibility of work disability evaluations between experts and health professionals with mandate to care (insurance setting)** | | | | |
| Dell-Kuster33 | NR | Work ability: last job; alternative job | Agreement: 51%; 20% | Yes |
| Lax34 | NR | Agreement on 4 disability items: full; partial; no agreement | Frequency of agreement: 4%; 34%; 78% | Yes |
| Okpaku41 | NR | Approval for social security benefits | Frequency of agreement: yes/no decisions 40% | Probably yes |
| **Studies investigating reproducibility of work disability evaluation between researchers (research setting)** | | | | |
| Berns37 | Yes | Functioning in work environment | ICC 0.86 | NA |
| Chopra38 | Yes | Remunerative employment | κ 0.62 | NA |
| Hannula39 | Yes | Occupational functioning | ICC 0.91 | NA |
| Hill36 | Yes | Dysfunctioning in work as social role | ICC 0.76 | NA |
| Mundo36 | Yes | Occupational skills | ICC 0.75 | NA |
| Nozu41 | Yes | Employability | ICC 0.88 | NA |
| **Fair to poor** | | | | |
| Daradkeh31 | Yes | Occupational functioning | κ 0.53 | NA |

NR=not reported; ICC=intraclass correlation; IRR=inter-rater reliability; NA=not applicable.
*Total >100%.

patient). Three studies qualified as “generalisable” and four as “probably not generalisable.”

Reproducibility between experts and health professionals with a mandate to care

Overall, across conditions and outcomes, percentage agreement ranged from 51% (work ability in last job)33 to 4% (somatic occupational disorders; four disability items)38 (table 8). Three studies compared reproducibility of ratings on work disability between experts and health professionals with a mandate to care.33 38 62 One study reported poor agreement between experts and the claimants’ treating physicians.33 Another study reported highly discordant judgments on disability between medical experts and health professionals of an occupational health centre.38 The third study found poor agreement between the decisions of the social security administration and those of an independent rehabilitation team.42

The direction of disagreement was mixed. Medical experts approved higher levels of work ability for claimants33 or their recommendations and decisions favoured the insurer,38 while in the third study, the rehabilitation team was more reluctant to grant disability benefits to patients with mental disorders than the social security administration.42 All studies based their judgments exclusively on professional expertise. Two studies qualified as “generalisable,”33 38 one as “probably generalisable.”42

Inter-rater reliability of ratings on disability for work and occupational functioning—research setting

Overall, across conditions and outcomes, the median inter-rater reliability was 0.76, ranging from an ICC of 0.91 (anxiety and mood disorders; occupational functioning)38 to κ of 0.53 (mixed mental disorders; occupational functioning)38.

Five of seven studies (71%) reported excellent (global) inter-rater reliability on work disability judgements with ICCs ranging from 0.75 to 0.91.35 The remaining two studies30 31 reported agreement on single items: good agreement (κ 0.62) regarding the ability to engage in remunerative employment30 and fair agreement (κ 0.53) for difficulties encountered in day-to-day work (occupational functioning).31

Impact of generalisability and level of standardisation on inter-rater reliability

Testing the relation between inter-rater reliability and objective (versus subjective) and chronic (versus acute)
health conditions as well as the studies’ overall generalisability did not show any association (subjectivity, 23 studies, \(P=0.46\); chronicity, 20 studies, \(P=0.45\); generalisability, 16 studies, \(P=0.65\)). Testing the relation between the level of standardisation and inter-rater reliability in all 23 studies showed a highly significant association \(P=0.006\).

**Discussion**

**Principal findings**

Current evidence regarding reliability of disability evaluation is limited and shows highly variable agreement between medical experts. Higher agreement seems to be associated with the use of a standardised approach to guide judgment and studies in a research (manufactured) setting.

**Strengths and limitations**

Strengths of our study include broad inclusion criteria to define eligibility and inclusion of publications in any language, which increases confidence that we captured all studies eligible for our review. Our outcome—global rating of disability for work—is highly relevant to the practice of medical experts, disability insurers, and employers, which increases the practical implications of our findings. Further, we evaluated the generalisability of evidence by following international guidance for evaluating reliability studies and by using an explicit approach in eliciting reviewers’ judgments on the relative weights of the generalisability items. While the high agreement we found among reviewers strengthens the credibility of the results, this approach requires further validation. Some cut-offs of the generalisability criteria (such as number of raters) are context specific and might not be applicable to settings other than assessment of disability. Furthermore, variability of study designs, measures of agreement, and outcomes precluded statistical pooling across studies.

**Relevant literature**

Disability evaluation is a poorly understood process that lacks any reference standard to confirm the validity of the findings. Health professionals who perform this task assess medical restrictions and limitations of claimants and are often asked to infer consequences on the ability to work. This, however, requires expertise in vocational rehabilitation, as medical restrictions do not correlate well with function and the ability to work. Implicit in the use of evaluations of disability by a third party is the concern that treating clinicians could have difficulty providing impartial assessments of their patients. Indeed, our findings suggest that medical experts (versus treating physicians) are more likely to conclude that claimants are capable of working. Claimant lawyers and patients’ organisations have raised concerns that experts who are paid to assess claimants for insurers might feel pressure to render opinions that favour the referral source.

**Implications for practice**

Our review suggests that use of standardised instruments could improve reliability in expert judgments on work disability. Appropriate instruments should therefore be considered in routine practice of disability evaluations (see table 8 and appendix 3 for examples). To ensure appropriate administration and interpretation of the findings, experts will need appropriate training and calibration on the use of such instruments. As most instruments reported in this review are available only in Dutch, other countries would need to develop their own instruments or translate instruments and accompanying manuals in national languages.

As few countries have standards to guide assessments, standardised instruments that improve reliability could become a target for change and parties ordering assessments should demand their use.

**Unanswered questions and future research**

Given the widespread use of evaluation of disability for work to determine claimants’ eligibility for work replacement benefits, our findings suggest that further research to improve reliability is urgently needed. Promising targets include formal training in evaluation of capacity to work, use of standardised instruments to guide disability evaluations, and addressing the conflict of interest that arises when insurers (or lawyers) select their own experts. Further, there might be greater need for strategies to improve agreement when patients present with subjective complaints. Ikezawa and colleagues found that different medical experts were able to agree on claimant’s ability to return to work in 97% of claims involving a fracture and 94% of claims involving a dislocation, but only 56% of claims...
because of chronic low back pain. Our review further suggests that interventions should be validated in real insurance settings, as experimental settings could artificially inflate agreement.

Improved knowledge of individual factors that contribute to variability in evaluation of capacity to work is also needed. Promising targets could provide a starting point to develop and test focused strategies to reduce variability (for example, appropriate assessment tools, guidelines, standard cases). Guidance is also required to inform the required level of inter-rater reliability to ensure equal treatment of claimants. Any decision on what constitutes an appropriate threshold, which might be similar to thresholds for clinical medical tests, will require societal discussion on what constitutes acceptable differences in the treatment of claimants or align to standards set by professional organisations of psychology or education. To make evaluations on work disability fair and meaningful and thereby qualify for decisions on claimants’ disability benefits, however, we suggest a minimum intraclass correlation coefficient of 0.6 (the cut off between fair and good inter-rater reliability), with a sufficiently narrow 95% confidence interval (0.5 to 0.7) to exclude poor reliability.

Conclusions
Despite their widespread use, medical evaluations of work disability show high variability and often low reliability. Use of standardised and validated instruments to guide the process could improve reliability. There is an urgent need for high quality research, conducted in actual insurance settings, to explore promising strategies to improve agreement in evaluation of capacity to work.

AUTHOR AFFILIATIONS
1Evidence-based Insurance Medicine (EbIM), Research and Education, Department Clinical Research, University Basel Hospital, University of Basel, Spitalstrasse 6 + 12, CH-4031 Basel, Switzerland
2Institute for Complementary and Integrative Medicine, University Hospital Zurich and University of Zurich, Zurich, Switzerland
3Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8S 4K1, Canada
4Department of Anaesthesia, McMaster University, Hamilton, ON L8S 4K1, Canada
5The Michael G DeGroote Institute for Pain Research and Care, McMaster University, Hamilton, ON L8S 4K1, Canada
6Coronet Institute of Occupational Health, Academic Medical Centre, University of Amsterdam, Amsterdam, Netherlands
7Research Centre for Insurance Medicine, AMC-UMC-UWV-Umc, Amsterdam, Netherlands
8School of Applied Psychology, Institute Humans in Complex Systems, Olten, Switzerland
9Dutch National Institute for Employee Benefits Schemes, Groningen, Netherlands
10Department of Health Sciences, Community and Occupational Medicine, University Medical Centre Groningen, Netherlands
11We thank Gordon Guyatt, McMaster University, Hamilton, Canada, for his input in conceptualising the review; Nozomi Takeshima, Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine, Japan, and Rosella Saule and Laura Amato, Department of Epidemiology, Lazio Regional Health Service, Rome (Italy), for extracting the data from the Japanese and Italian studies; and Sacha Roschard for administrative support.

Contributors: RK and WdB developed the idea. RK, JB, WdB, JWB, RF, and GG contributed substantially to the conception and design. RK, JB, WdB, JWB, SC, KF, DvA, and JS drafted or revised the manuscript critically for important intellectual content, approved the final version to be published. RK and JB are guarantors.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coiDisclosure.pdf (available on request from the corresponding author) and declare that JB acts as a consultant to Prisma Health Canada, a private incorporated company funded by employers and insurers that consults on and manages long-term disability claims. The evidence-based medicine unit at the University Hospital in Basel is funded in part by donations from public insurance companies and a consortium of private insurance companies (RK). After the manuscript was finalised, RK took a part time position at the Swiss National Accident Insurance Fund, Suva. RK, JB, WdB, JWB, and JH were initiators of Cochrane Insurance Medicine.

Ethical approval: Not required.

Data sharing: All additional data available.

Transparency: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc-3.0/.

1 International Social Security Association. Country Profiles. https://www.issa.int/country-profiles.
2 Sickness, Disability and Work. Breaking the Barriers. A Synthesis of Findings across OECD Countries. OECD, 2010 Contract No. ISBN 978-92-64-08884-9.
3 Anner J, Kunz R, Boer Wd. Reporting about disability evaluation in European countries. Disabil Rehabil 2014;36:848-54. doi:10.3109/09638288.2013.821180.
4 Anner J, Schweger U, Kunz R, Trezzi B, de Boer W. Evaluation of work disability and the international classification of functioning, disability and health: what to expect and what not. BMC Public Health 2012;12:670. doi:10.1186/1471-2458-12-670.
5 Busse JW, Bruun-Meyer SE, Ebrahim S, Kunz R. A 45-year-old woman referred for an independent medical evaluation by her insurer. CMAJ 2014;186:5627-30. doi:10.1503/cmaj.130363.
6 Pizała H. Evaluation von psychiatrischen Gutachten [Reports of disability evaluations in claimants with mental disorders. A quality assessment.] [Dissertation]. University of Basel, 2010.
7 Manchikanti L. Impairment evaluation in pain management.
8 Peterson KW, Babitsky S, Beller TA, et al. The American Board of Independent Medical Examiners. J Occup Environ Med 1997;39:509-14. doi:10.1097/00007632-198803000-00019.
9 Dickmann JR, BrookS A. [Psychiatric expert opinion in case of early retirement--how reliable?]. Fortschr Neurol Psychiatrie 2007;75:397-401. doi:10.1055/s-2006-944303.
10 Clark WL, Haldeman S, Johnson P, et al. Back impairment and disability determination. Another attempt at objective, reliable rating. Spine (Phila Pa 1976) 1998;13:332-41. doi:10.1097/00007632-199803000-00019.
11 Ingvaldso F, Vignatelli L, Brin M, et al. Medico-legal assessment of disability in narcolepsy: an interobserver reliability study. J Sleep Res 2008;17:111-1. doi:10.1111/j.1365-2869.2008.00630.x.
12 Alexios G. Disabled people are trapped in assessment 'nightmare' by benefits regime, says Dr. Stephen Duckworth. The Independent 2014 22.4. Sect. Home News.
13 Kleinfeld NR. Exams of injured workers fuel mutual mistrust. New York Times 2009. http://www.nytimes.com/2009/04/01/nyregion/01comp.html.
14 Hesse B, Gebauer E. Sozialmedizinische Begutachtung im Rentenverfahren: Stellenwert, Forschungsbedarf und Chancen. Rehbildung (Stuttgart) 2011;50:17-24. doi:10.1515/0030-127043.
15 Spanjers K, Krol B, Brouwer S, Groothoff JW. Sources of variation in work disability assessment. Work 2010;37:405-11.
16 Kobak KA, Brown B, Sharp I, et al. Sources of unreliability in depression ratings. J Clin Psychopharmacol 2009;29:82-5. doi:10.1097/JCP.0b013e3181926ad7.
17 Kottern J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRARS) were proposed. J Clin Epidemiol 2011;64:96-106. doi:10.1016/j.jclinepi.2010.03.002.
18 Libesati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2070. doi: 10.1136/bmj.b2070.

19 Lucas NP, Macaskill P, Higw I, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol 2010;63:554-61. doi:10.1016/j.jclinepi.2009.10.002.

20 Karanicolas PJ, Bhandari M, Kreder H, et al. Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group. Evaluating agreement: conducting a reliability study. J Bone Joint Surg Am 2009;91(Suppl 3):99-106. doi: 10.2106/JBJS.H.01624.

21 Edwards W. The theory of decision making. Psychol Bull 1954;51:380-417. doi:10.1037/h0053870.

22 Spanjer, KJ, Brol, Popping R, Groothoff JW, Broger S. Disability assessment interview: the role of detailed information on functioning in addition to medical history-taking. J Rehabil Med 2009;41:267-72. doi:10.1080/02868186811582.

23 Spanjer, KJ, Brol, S. Brouwer, S. Groothoff JW. Inter-rater reliability in disability assessment based on a semi-structured interview report. Disab Abil Rehabil 2008;30:1885-90. doi:10.1080/09638280701688185.

24 Spanjer, KJ, Brol, 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. J Occup Rehabil 2010;20:33-40. doi:10.1007/s10926-009-9203-2.

25 Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. Wiley, 1981.

26 Cicchetti DV. The precision of reliability and validity estimates revisited: distinguishing between clinical and statistical significance of sample size requirements. J Clin Exp Neuropsychol 2001;23:695-700. doi:10.1076/jcen.23.5.695.1249.

27 Fleiss JL, Levin B, Paik MC. Statistical Measures for Rates and Proportions. 3rd ed. Wiley, 2003 doi:10.1002/0471445428.

28 Cicchetti D, Bronen R, Spencer S, et al. Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. J Nerv Ment Dis 2006;194:537-64. doi:10.1097/01.nmd.0000230392.83607.c5.

29 Bennis S, Uzelac S, Gonzalez C, Jaeger J. Methodological considerations of measuring disability in bipolar disorder: validity of the Multidimensional Scale of Independent Functioning. Bipolar Disord 2007;9:3-10. doi:10.1111/j.1399-5618.2007.00305.x.

30 Chopra P, Couper J, Hermann H. The assessment of disability in patients with psychotic disorders: an application of the KIDDI-2. Aust N Z J Psychiatry 2003;37:127-32. doi:10.1046/j.1440-1614.2002.00976.x.

31 Daradkeh TK, Saud A. The reliability and validity of the proposed axis V (disabilities) of ID-10. Br J Psychiatry 1994;165:683-5. doi:10.1192/bjp.165.5.683.

32 de Kort WL, Uiterweer HW, van Dijk FI. Agreement on medical fitness for a job. Scand J Work Environ Health 1992;18:246-51. doi:10.5271/sjweh.3440.

33 De Kruif-Bruijn K, Lafeber-van den Brandt PA, et al. Assessing work ability—a cross-sectional study of interrater agreement between disability claimants, treating physicians, and medical experts. Scand J Work Environ Health 2014;40:493-501. doi:10.5271/sjweh.3440.

34 Elder AL, Symington IS, Symington EH. Do occupational physicians agree about ill-health retirement? A study of simulated retirement assessments. Occup Med (Lond) 1994;44:231-5. doi:10.1093/occmed/44.5.231.

35 Hannula JA, Lahtela K, Jarvokoski A, Salminen JK, Mäkelä P. Occupational Functioning Scale (OFS)—an instrument for assessment of work ability in psychiatric disorders. Nord J Psychiatry 2006;60:372-8. doi:10.1080/08039480600937140.

36 Hill J, Harrington R, Fudge H, Rutter M, Pickles A. Adult personality functioning assessment (APFA): An investigator-based standardised interview. Br J Psychiatry 1989;155:24-35. doi:10.1192/bjp.155.1.24.

37 Ikezawa Y, Battle MC, Beach J, Gross D. Do clinicians working within the same context make consistent return-to-work recommendations? Occup Rehabil 2010;20:367-77. doi:10.1007/s10926-010-9230-z.

38 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;10:1-12. doi:10.1179/oeh.2004.10.1.1.

39 Lederer P, Pflafl G, Walter K, Wehrauch M, Weber A. [Quality circles in expert assessment as an instrument in quality management]. Gesundheitswesen 1998;60:415-9.

40 Mondo E, Bontalume L, Del Corso F, Maddeddu F, Lang M. L’applicazione dell’Asse V di Kennedy a un campione clinico italiano. Riv Psichiatr 2010;45:214-20.

41 Nozu M. Evaluating work-personality insufficiency of schizophrenic patients, an assessment of employability in psychiatric rehabilitation]. Seishin Shinkeigaku Zasshi 1995;97:217-38.

42 Oskupki SO, Sibulkin AE, Schenitzer C. Disability determinations for adults with mental disorders: Social Security Administration vs. independent judgments. Am J Public Health 1994;84:1791-5. doi:10.2105/APH.84.11.1791.

43 Rudbeck M, Fonager K. Agreement between medical expert assessments in social medicine. Scand J Public Health 2011;39:766-72. doi:10.1177/1403494811418282.

44 Schellart AJ, Zweerer F, Anema JR, Van der Breek A. The influence of applying insurance medicine guidelines for depression on disability assessments. BMC Res Notes 2011;4:225. doi:10.1186/1756-0500-6-225.

45 Schreuder A, 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;34:1814-9. doi:10.3109/09638288.2012.665125.

46 Siebus FK, Kijpe PP, Willems MM, Frings-Dresen MH, Sluiter JK. Work ability assessment in prolonged depressive illness. Occup Med (Lond) 2010;60:107-9. doi:10.1093/occmed/kqp079.

47 Walter SD, Eliaszew M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101-10. doi:10.1002/(SICI)1097-0258(19980115)17:1.0.CO;2-E.

48 Dutch Organisation of Insurance Physicians (Nederlandse Vereniging voor Verzekeringsgeneeskunde). http://www.nvvg.nl/index.php?what-doet-de-verzekeringarts.

49 Meenshoek A, Krumeich A, Vos R. Judging without criteria? Sickness certification in Dutch disability schemes. Social Health Illn 2007;29:497-514. doi:10.1111/j.1476-5567.2007.00109.x.

50 Bachmann M, de Boer W, Schandelmaier S, et al. Use of a structured functional evaluation process for independent medical evaluations of claimants presenting with disabling mental illness: rationale and design for a multi-center reliability study. BMC Psychiatry 2016;16:271. doi:10.1186/s12888-016-0967-6.

Appendix 1: Search strategy
Appendix 2: Determination of overall generalisability
Appendix 3: Studies reporting use of instrument with instrument referenced
Appendix 4: Full details of performance and findings of included studies