UNI
BASEL

Universitätsspital
Basel

Departement Klinische Forschung

EbIM Forschung & Bildung

The RELY - Studies

On the reliability and agreement of medical assessments in patients with mental disorders

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Sozialversicherungen BSV

suva

Overview

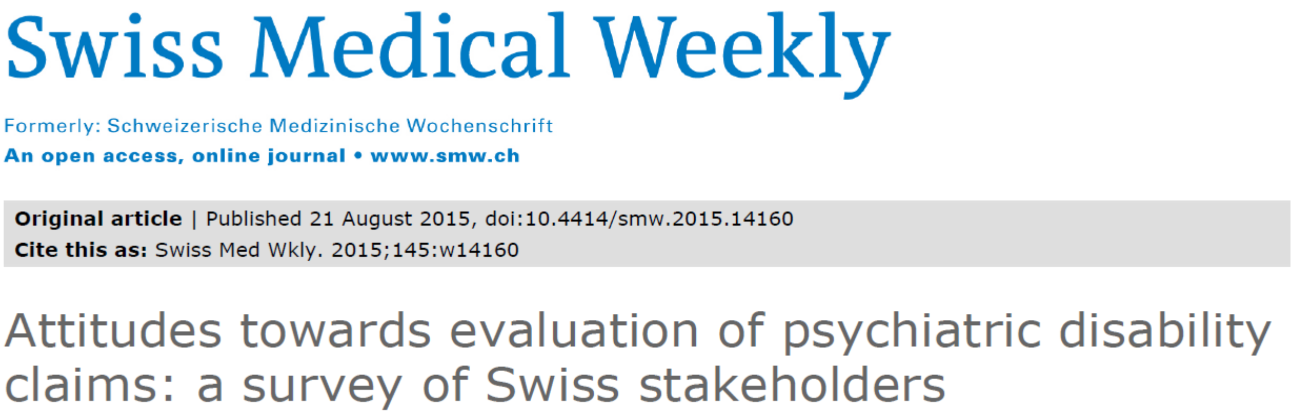
Background to the studies

Study goals

Training in functional assessments

The findings and what they mean

Where next?



Swiss Medical Weekly

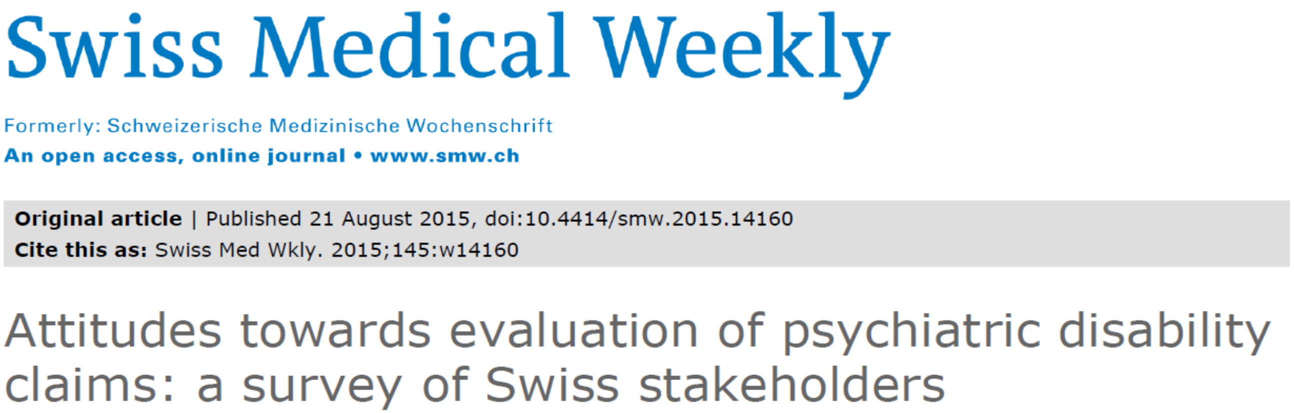
Formerly: Schweizerische Medizinische Wochenschrift
An open access, online journal • www.smw.ch

Original article | Published 21 August 2015, doi:10.4414/smw.2015.14160
Cite this as: Swiss Med Wkly. 2015;145:w14160

Attitudes towards evaluation of psychiatric disability claims: a survey of Swiss stakeholders

What is the maximum acceptable difference in WC ratings between two experts performing a psychiatric evaluation in th same patient ..

	Lawyer (n=81)	Psychiatrists (treating) (n=242)	Psychiatrists (experts) (n=114)	Judges (n=47)	Insurers (n=108)
... in the current procedure with the known restrictions	**15%** (10-20%)	**20%** (10-25%)	**20%** (10-25%)	**15%** (10-20%)	**10%** (10-20%)



Swiss Medical Weekly

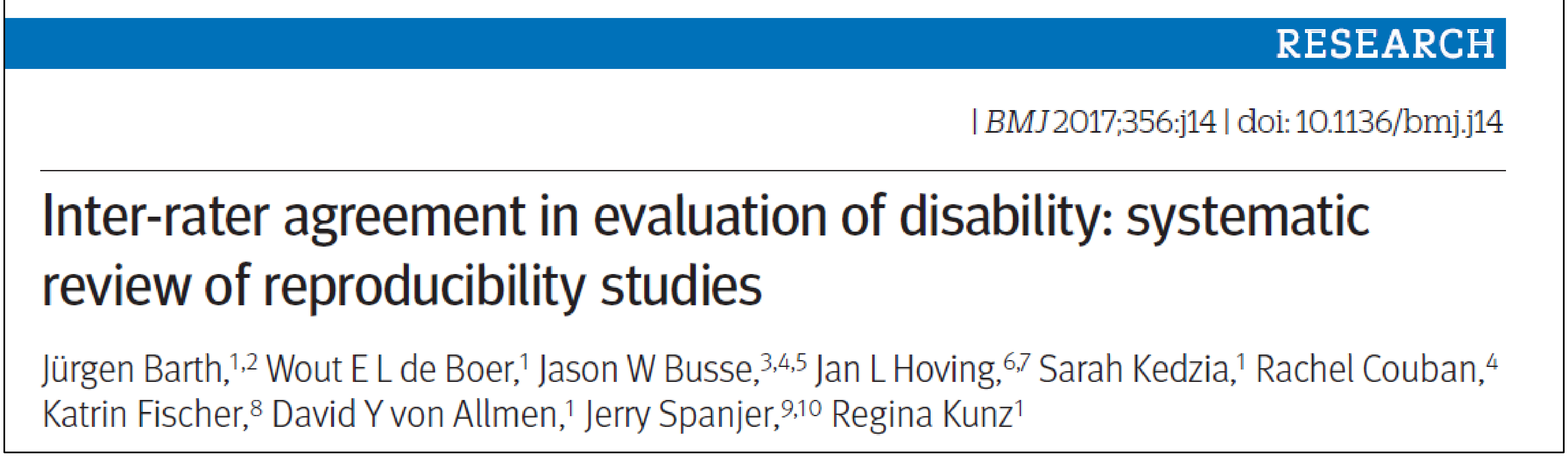
Formerly: Schweizerische Medizinische Wochenschrift
An open access, online journal • www.smw.ch

Original article | Published 21 August 2015, doi:10.4414/smw.2015.14160
Cite this as: Swiss Med Wkly. 2015;145:w14160

Attitudes towards evaluation of psychiatric disability claims: a survey of Swiss stakeholders

What is the maximum acceptable difference in WC ratings between two experts performing a psychiatric evaluation in th same patient ..

	Lawyer (n=81)	Psychiatrists (treating) (n=242)	Psychiatrists (experts) (n=114)	Judges (n=47)	Insurers (n=108)
... in the current procedure with the known restrictions	**15%** (10-20%)	**20%** (10-25%)	**20%** (10-25%)	**15%** (10-20%)	**10%** (10-20%)
... in a process under optimal conditions	**10%** (10-15%)	**10%** (10-20%)	**10%** (10-20%)	**10%** (10-15%)	**10%** (5-10%)

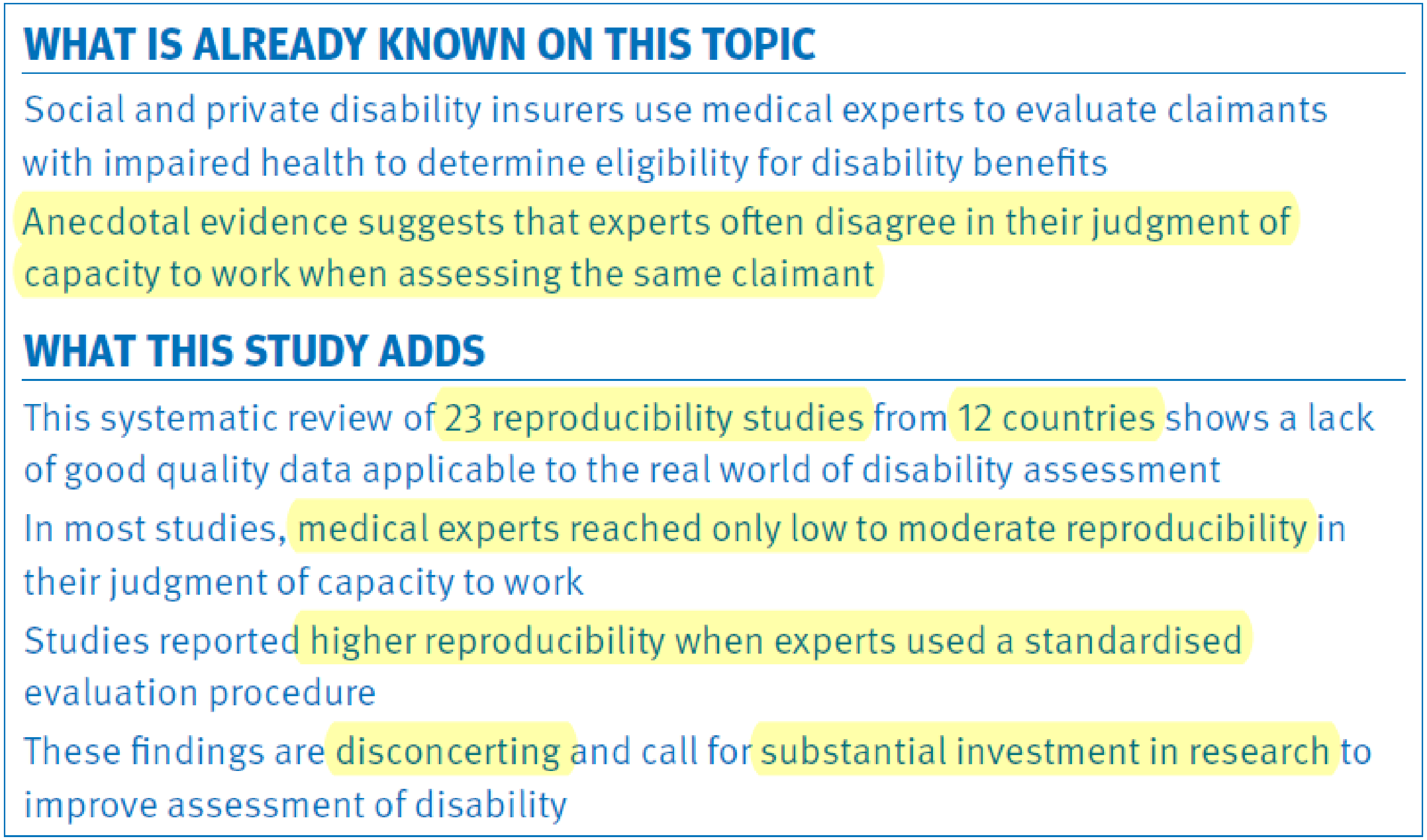


RESEARCH

| *BMJ* 2017;356:j14 | doi: 10.1136/bmj.j14

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

Jürgen Barth,[1,2] Wout E L 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,[9,10] Regina Kunz[1]



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

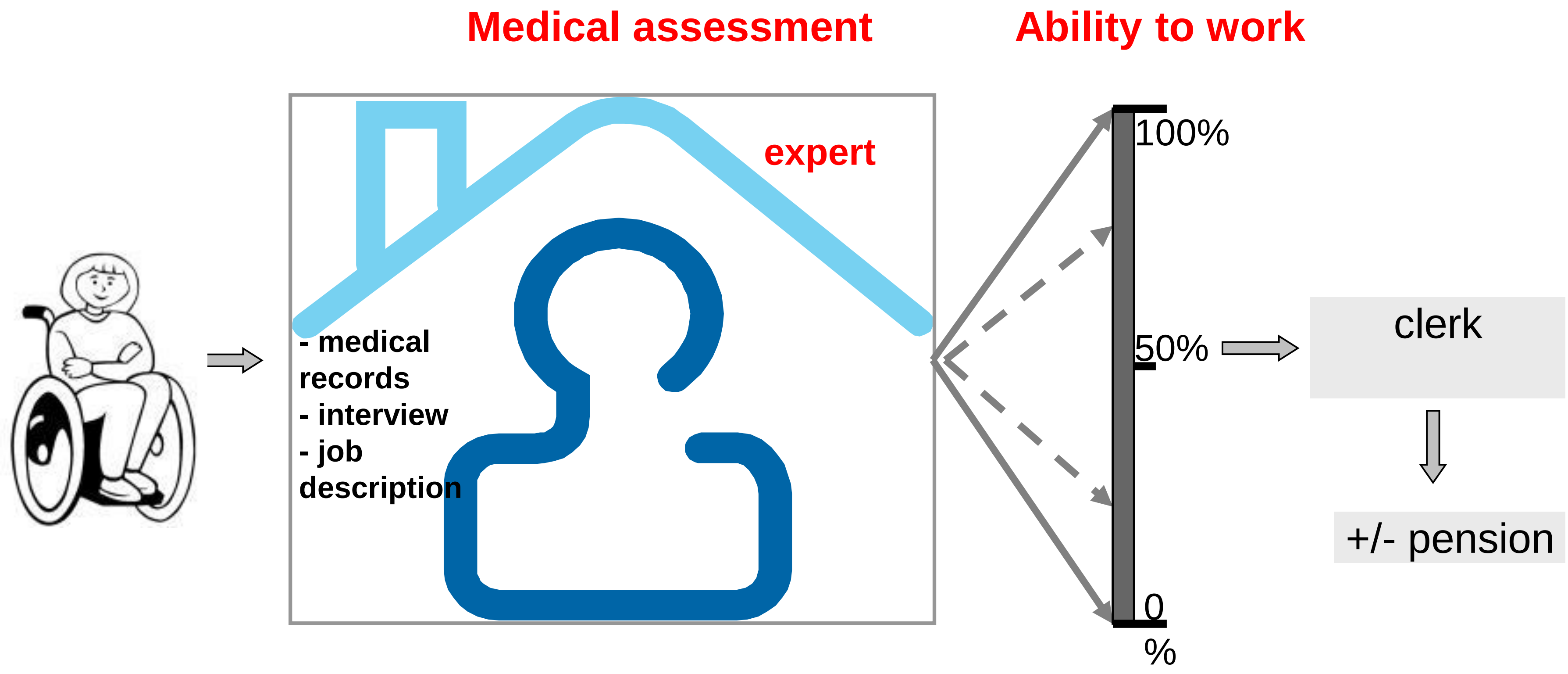
Medical assessments under general criticism

Our goals

1) Increase reproducibility of the evaluation results

2) Improve transparency and comprehension

To understand the process of a medical assessment as an "instrument" (*black box*) for capturing functional capability and ability to work



Medical assessment

Ability to work

expert

- medical records
- interview
- job description

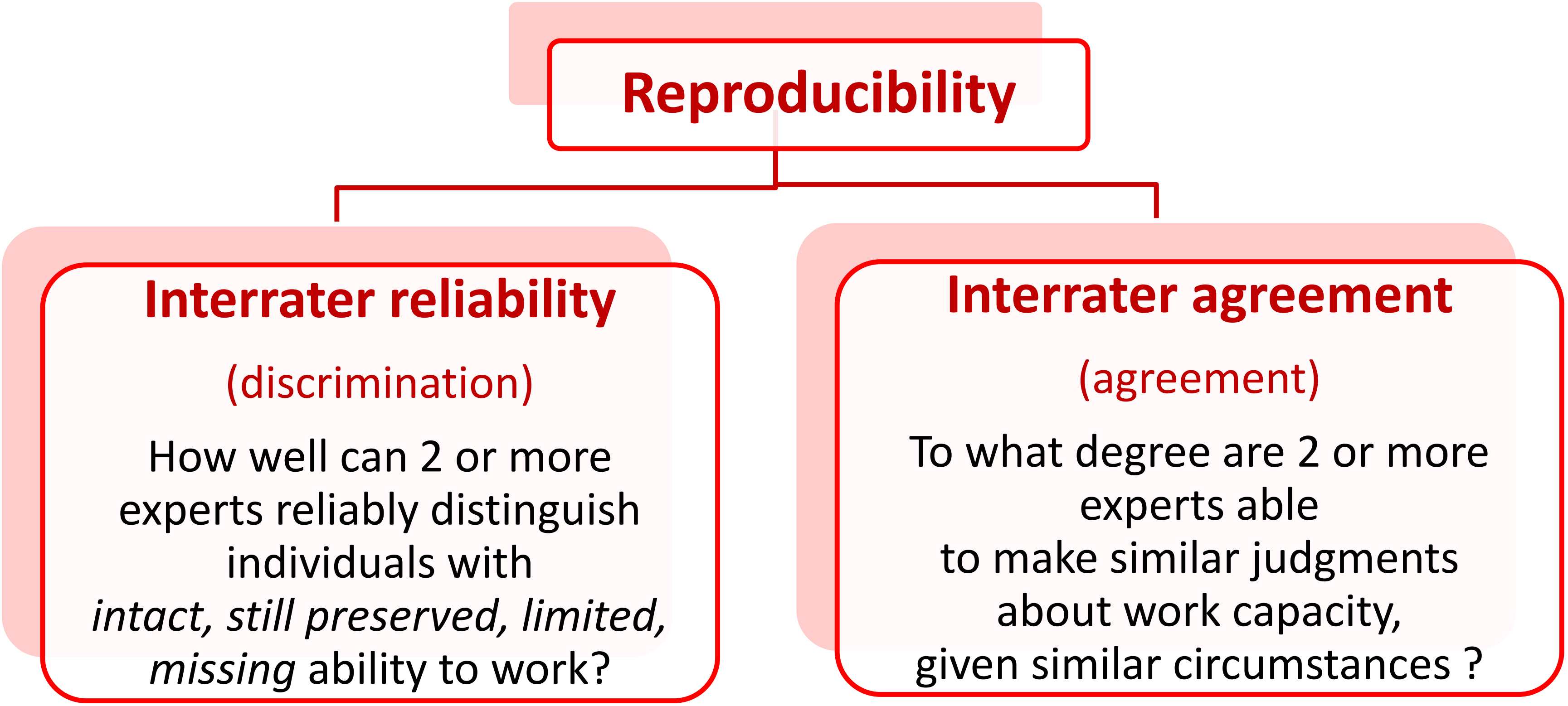
100%

50%

0 %

clerk

+/- pension



Reproducibility

Interrater reliability

(discrimination)

How well can 2 or more experts reliably distinguish individuals with *intact, still preserved, limited, missing* ability to work?

Interrater agreement

(agreement)

To what degree are 2 or more experts able to make similar judgments about work capacity, given similar circumstances ?

Training in functional evaluation

→

Functional Interviewing

semi-structured,
exploring the claimants'
self-reported work limitations

IFAP

Instrument for **F**unctional
Assessment in **P**sychiatry

Procedure

RELY 1

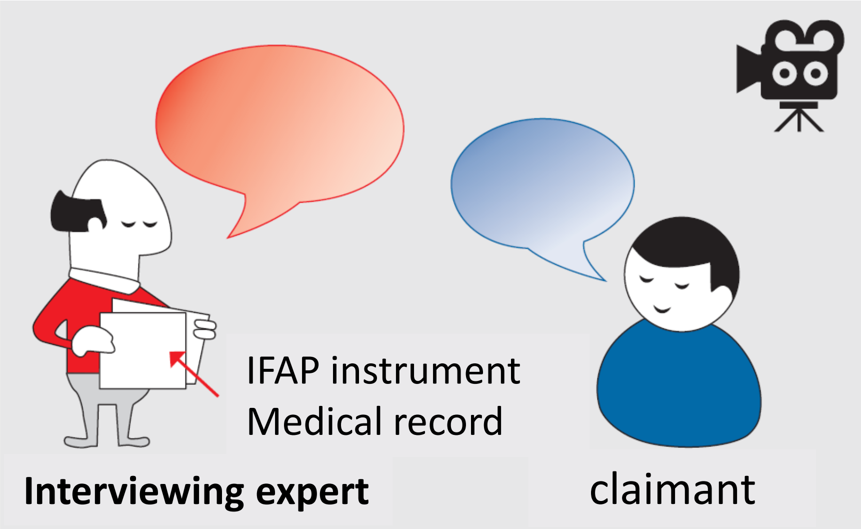
Training

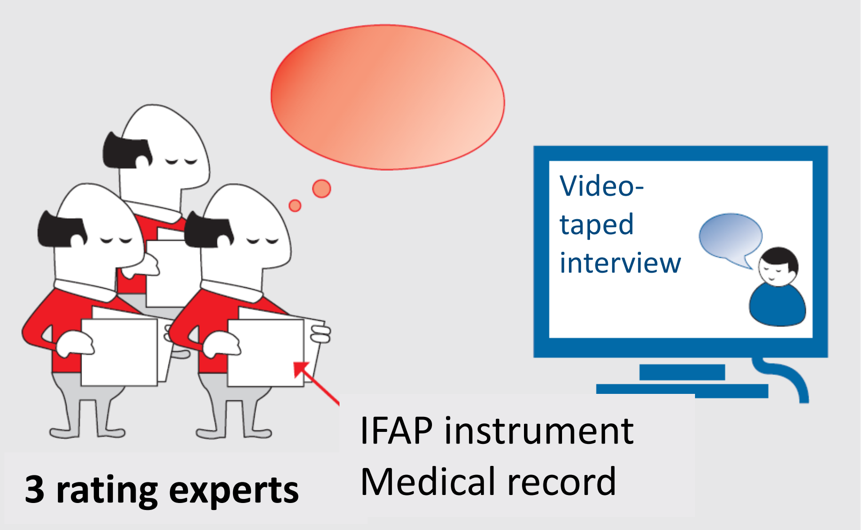
19 psychiatrists

30 claimants

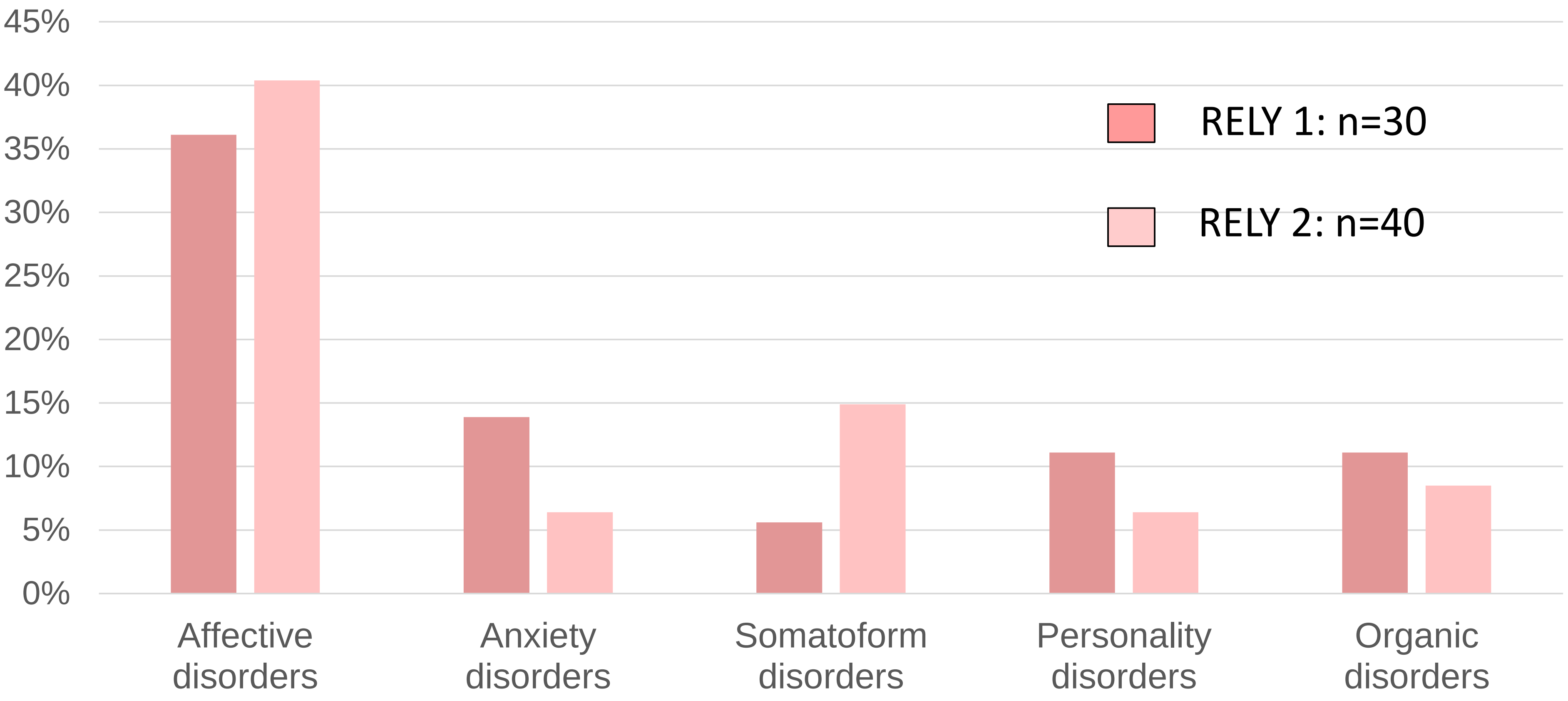
Agreement in

work capacity





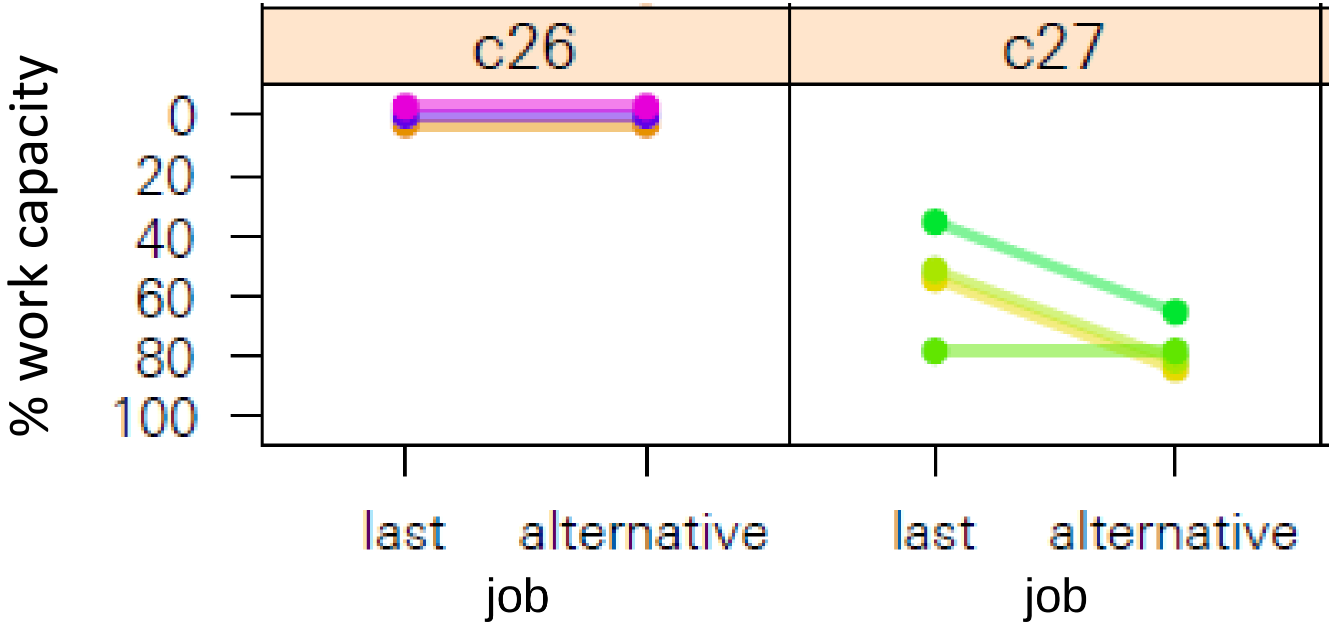
Claimants’ diagnoses



Severity of mental disorders	**RELY 1**	**RELY 2**
scale 0-10	mean 5.34	4.95

How to read the results?

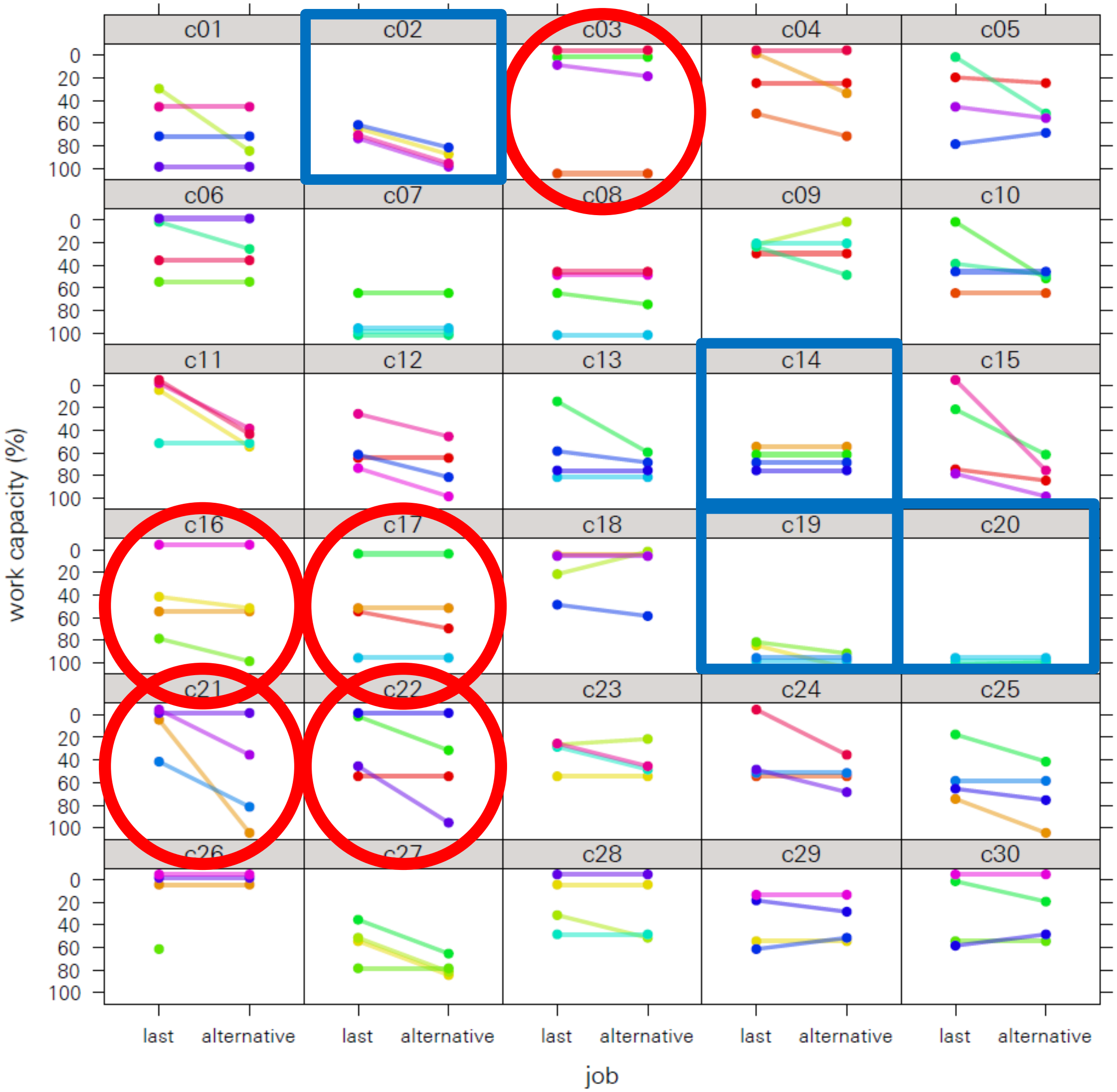
Assessment of … % work capacity



RELY 1

Ratings of the experts

N = 30 applicants



Difference 100% points:

7/60 (12%) Ratings

Our initial explanation for RELY 1 - results

- large time span between training and rating
- 3x3h: training too little intensive

Training RELY 2

Functional Interview

→

IFAP

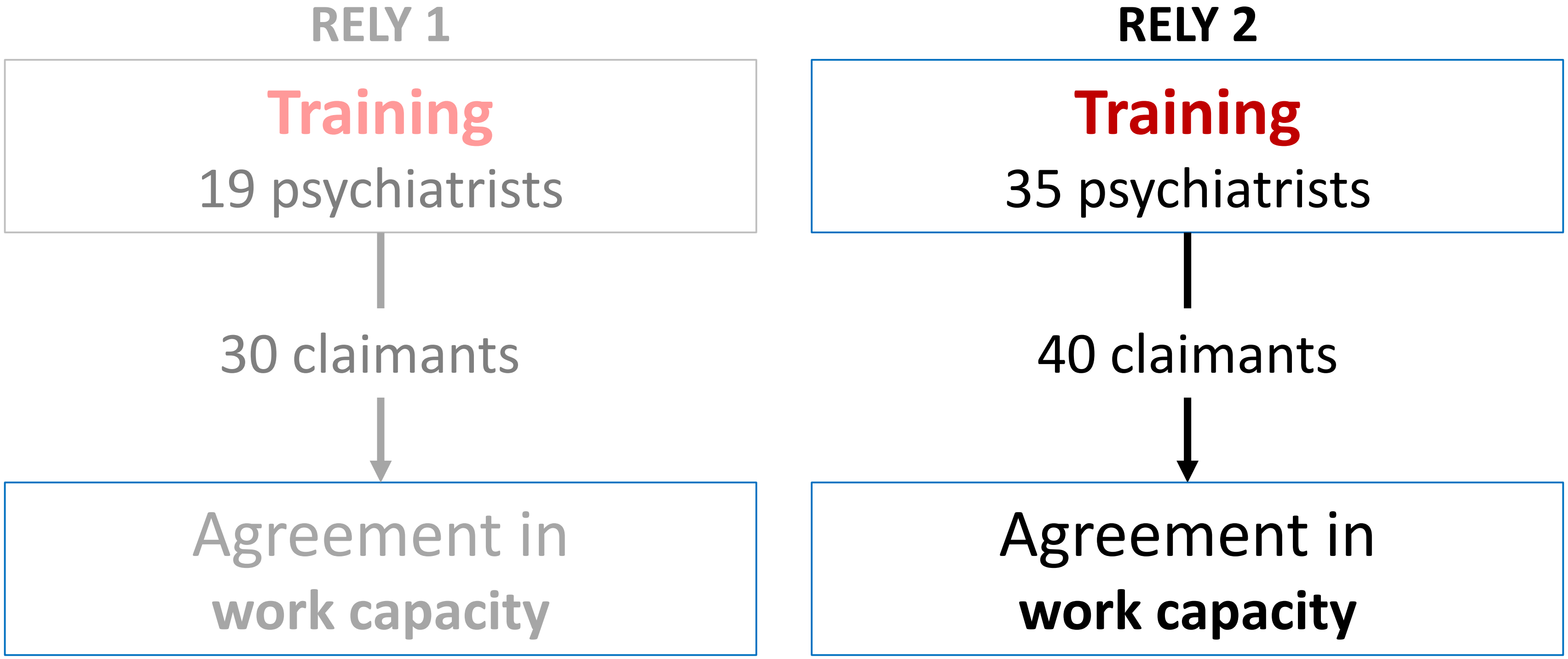
Instrument for **F**unctional **A**ssessment in **P**sychiatry

Enhanced training

- Doubling of face-to-face training time (18 hrs.)
- Revision and enhancement of the manual
- Intensive calibration to the rules

Rating closer to the training

Procedure



RELY 1

Training

19 psychiatrists

30 claimants

Agreement in **work capacity**

RELY 2

Training

35 psychiatrists

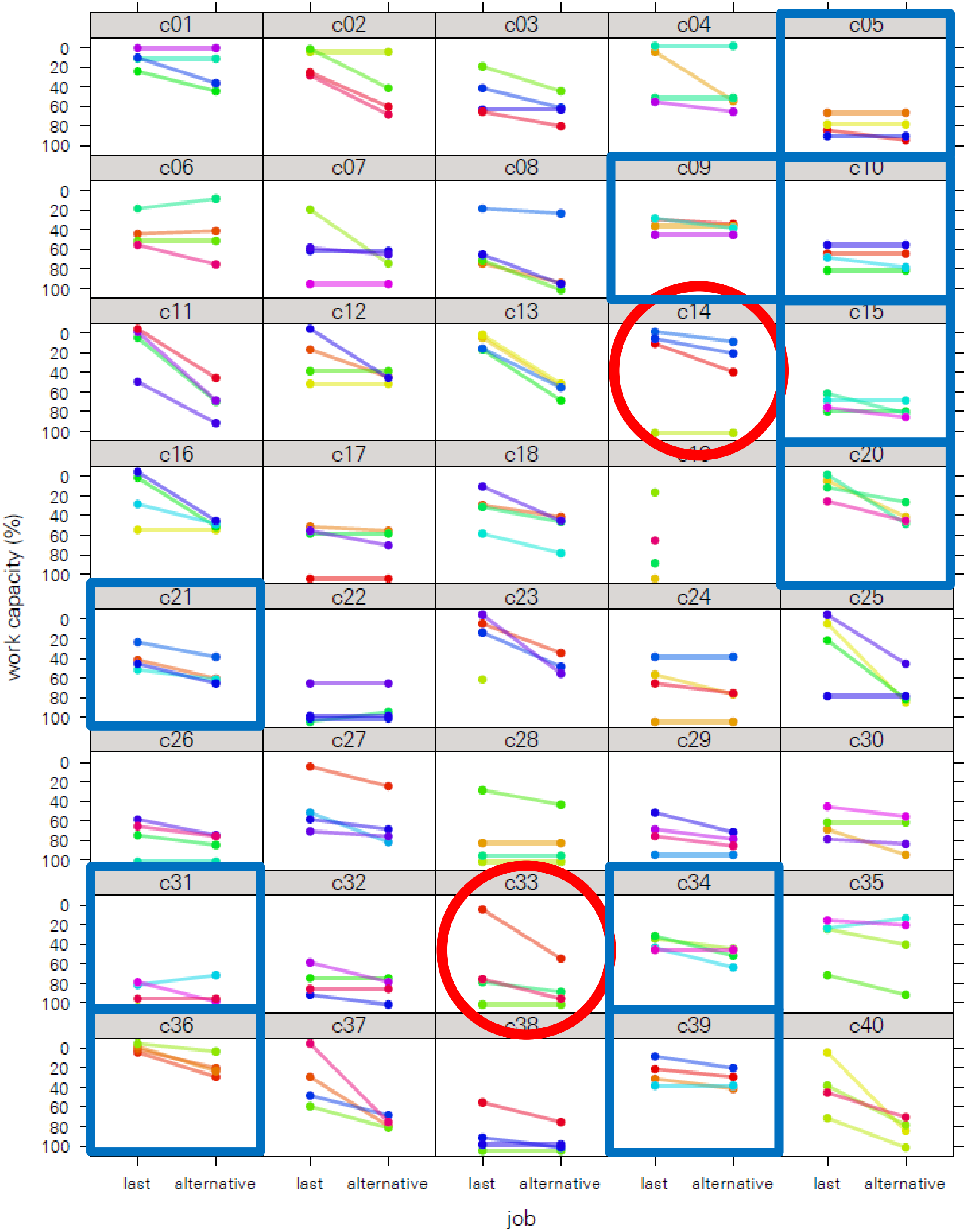
40 claimants

Agreement in **work capacity**

RELY 2

Ratings of the experts

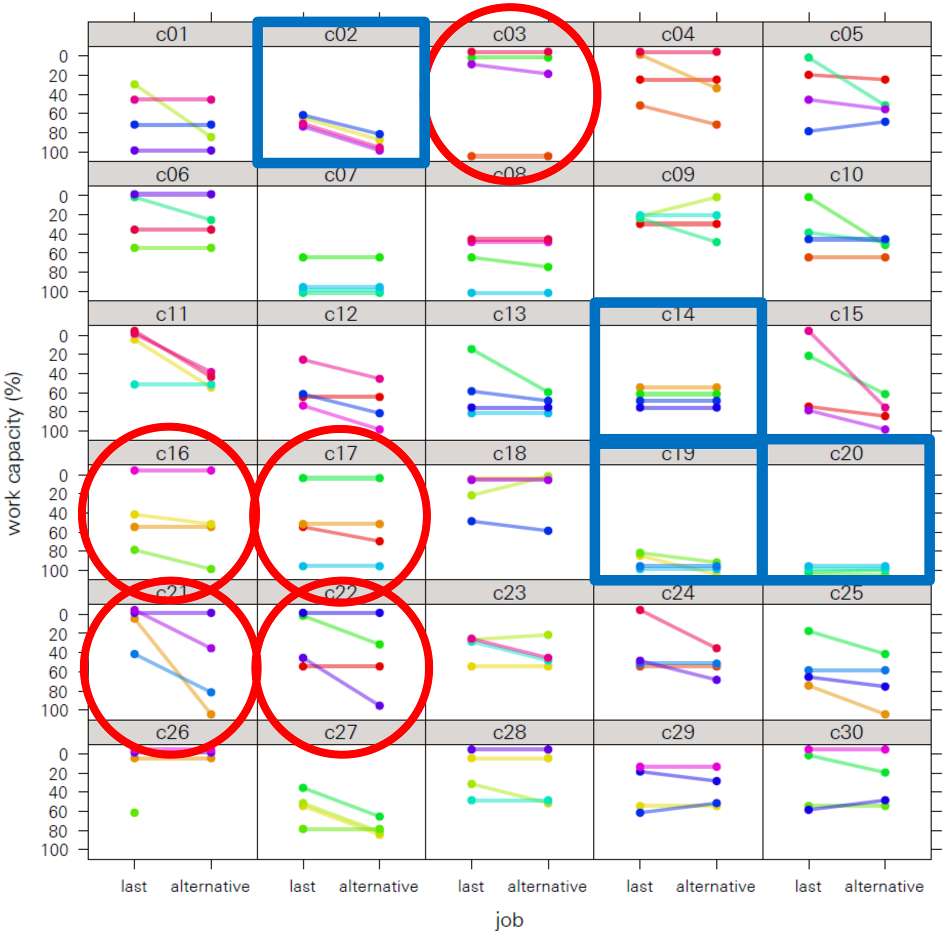
N = 40 claimants



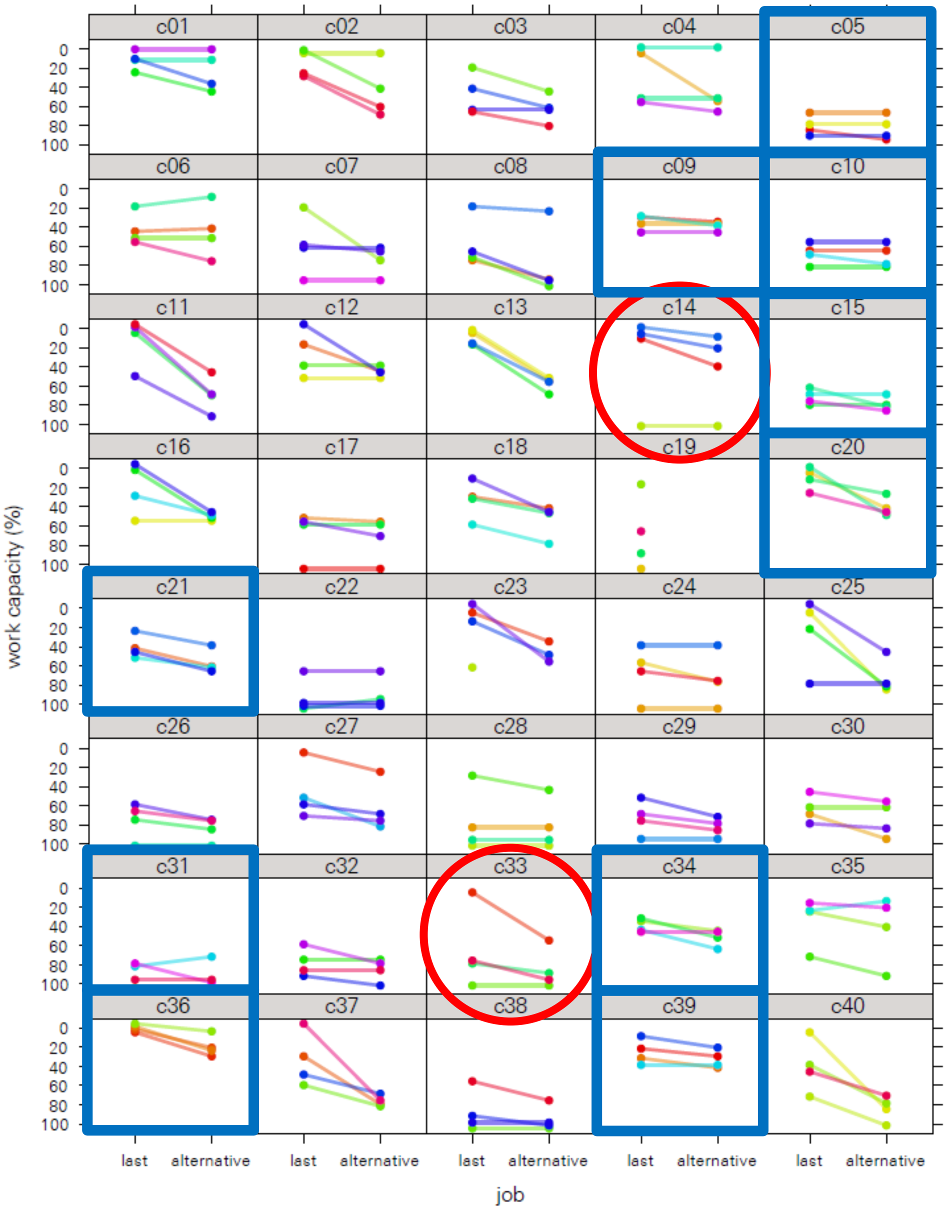
<u>Difference 100% points:</u>

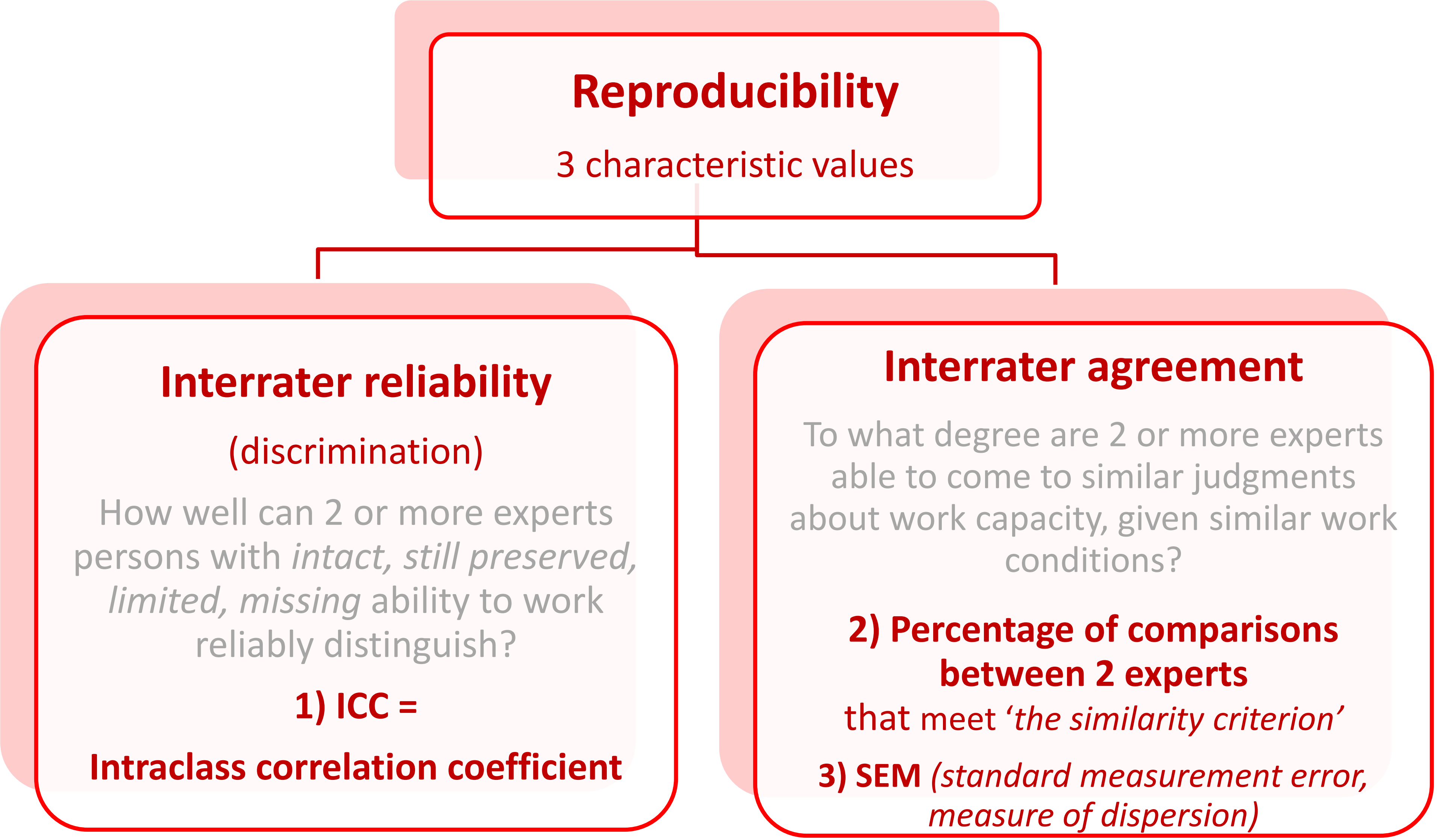
- last job N=2
- alternative job N=0

RELY 1


c01
c02
c03
c04
c05
c06
c07
c08
c09
c10
c11
c12
c13
c14
c15
c16
c17
c18
c19
c20
c21
c22
c23
c24
c25
c26
c27
c28
c29
c30
0
20
40
60
80
100
work capacity (%)
last
alternative
job

RELY 2


c01
c02
c03
c04
c05
c06
c07
c08
c09
c10
c11
c12
c13
c14
c15
c16
c17
c18
c19
c20
c21
c22
c23
c24
c25
c26
c27
c28
c29
c30
c31
c32
c33
c34
c35
c36
c37
c38
c39
c40
0
20
40
60
80
100
work capacity (%)
last
alternative
job



Reproducibility

3 characteristic values

Interrater reliability

(discrimination)

How well can 2 or more experts persons with *intact, still preserved, limited, missing* ability to work reliably distinguish?

1) ICC =

Intraclass correlation coefficient

Interrater agreement

To what degree are 2 or more experts able to come to similar judgments about work capacity, given similar work conditions?

2) Percentage of comparisons between 2 experts
that meet ‘*the similarity criterion*’

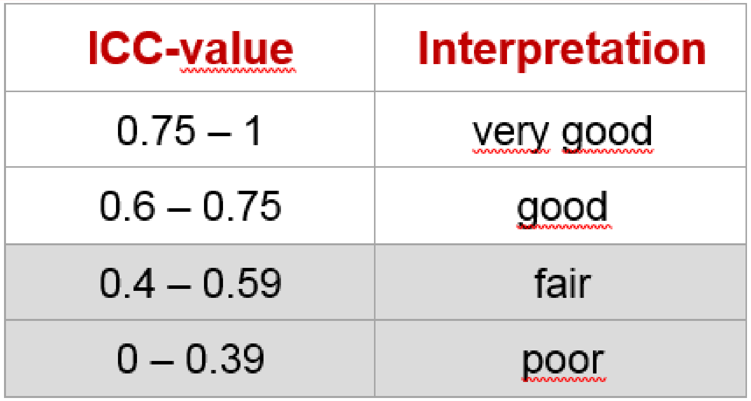
3) SEM *(standard measurement error, measure of dispersion)*

Reproducibility

1) Interrater reliability

(discrimination)

How well can 2 or more experts persons with *intact, still preserved, limited, missing* ability to work reliably distinguish?



ICC-value	Interpretation
0.75 – 1	very good
0.6 – 0.75	good
0.4 – 0.59	fair
0 – 0.39	poor

Results: Reliability and Agreement

	Average ability to work Alternative work
RELY 1 120 reviews	**55%**
RELY 2 160 reviews	**63%**

Results: Reliability and Consistency

1. Reliability values (discrimination)

for last job and alternative work

		Reliability
		ICC
Last job	RELY 1	**0.38**
	RELY 2	**0.47**
Alternative work	RELY 1	**0.43** (0.22-0.60)
	RELY 2	**0.44** (0.25-0.59)

		Reliability
		ICC
Last job	RELY 1	**0.38** (0.19-0.55)
	RELY 2	**0.47** (0.29-0.61)
Alternative work	RELY 1	**0.43**
	RELY 2	**0.44**

Reproducibility

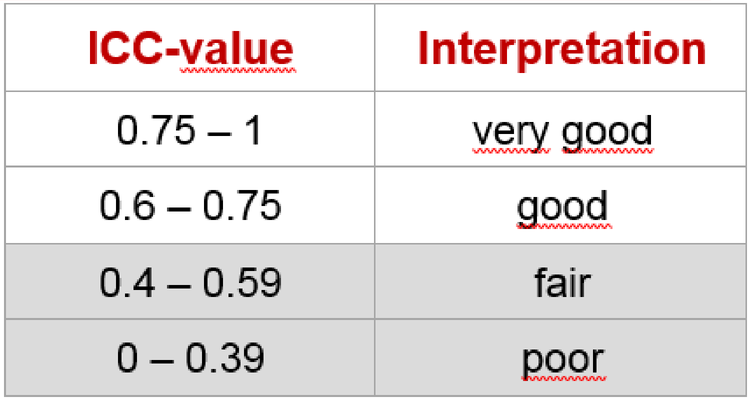
1) Interrater reliability

(Distinctness)

How well can 2 or more experts persons with *intact, still preserved, limited, missing* ability to work reliably distinguish?

ICC

Intraclass correlation coefficient



ICC-value	Interpretation
0.75 – 1	very good
0.6 – 0.75	good
0.4 – 0.59	fair
0 – 0.39	poor

Factors that impact on the %- work capacity

- **Psychiatrists,** e.g.:
  - (Un)structured nature of the procedure
  - Experience as a psychiatrist / medical expert
  - Subjective «strictness / mildness"
  - Political attitude
- **Claimants**, e.g.:
  - Socio-demographic characteristics
  - Diagnosis, severity of disorder
  - Motivation / self-awareness about the ability to work
- **Other factors**
  - Situational factors; interaction psychiatrist* claimant
  - Environmental conditions (e.g. socio-political climate, impact of various federal laws on assessment; staff turn-over in the study)

Interpretation of low reliability in RELY

➔ Claimants are becoming more similar: each claimant has certain limitations, only a few are fully capable (or unable) to perform *(the very sick or healthy ones tend not to come for an assessment ...).*

➔ **Discrimination remains difficult**

It's harder to distinguish people who are relatively similar than people who are very different

(Streiner 2014)

Results: Reliability and Agreement

2. Agreement: How many fulfil '*the similarity criterion*'?

'The similarity criterion'

Maximum acceptable difference
in the assessment of the work capacity
(scale of 100%-0%)

< 25 percentage points of work capacity

Two psychiatric experts independently judging the same claimant in his ability to work

"In your opinion, what would be the maximum acceptable difference in work capacity?"

	Lawyer (n=81)	**Psychiatrists** (treating) (n=242)	**Psychiatrists** (experts) (n=114)	**Judges** (n=47)	**Insurers** (n=108)
... in the current procedure with the known restrictions	**15%** (10-20%)	**20%** (10-25%)	**20%** (10-25%)	**15%** (10-20%)	**10%** (10-20%)

Schandelmaier. Stakeholder Survey Swiss Med Wkly 2015

Please check translation

Results: Reliability and Agreement

2. Agreement: How many fulfil '*the similarity criterion*'?

Maximum acceptable difference in the assessment of the work capacity (scale of 100%-0%)

< 25 percentage points of work capacity

Example

Expert Amann evaluated 30% WC

Expert Bolzli «50% WC» => difference: 20% points WC → similarity

Expert Zapf «70% WC» => difference: 40% points WC → no similarity

Please check translation

Results: Reliability and Agreement

2. Agreement: How many fulfil ‘*the similarity criterion*’?

	Agreement	
	2) Proportion of two experts reaching ‘the similarity criterion’	Measure of dispersion 3) Standard Error of Measurement Smaller is better
RELY 1 n=177 comparisons	**61.6% of agreements** (109/177 comparisons)	**24.6 percentage points WC**
RELY 2 n=231 Comparisons	**73.6% of agreements** (170/231 comparisons)	**19.4 percentage points WC**

Results: Reliability and Agreement

3. Agreement: Dispersion

	Agreement	
	2) Proportion of two experts reaching 'the similarity criterion'	Measure of dispersion b) SEM, Standard Error of Measurement **Smaller is better**
RELY 1 n=177 comparisons	**61.6% of agreements** (109/177 comparisons)	**24.6 percentage points WC**
RELY 2 n=231 Comparisons	**73.6% of agreements** (170/231 comparisons)	**19.4 percentage points WC**

Two psychiatric experts independently judging the same claimant in his ability to work

"In your opinion, what would be the maximum acceptable difference in work capacity?"

	Lawyer (n=81)	**Psychiatrists** (treating) (n=242)	**Psychiatrists** (experts) (n=114)	**Judges** (n=47)	**Insurers** (n=108)
... in the current procedure with the known restrictions	**15%** (10-20%)	**20%** (10-25%)	**20%** (10-25%)	**15%** (10-20%)	**10%** (10-20%)

Schandelmaier. Stakeholder Survey
Swiss Med Wkly 2015

SEM calculation for maximum acceptable differences

a) Expectation by stakeholders	
Expected 'Maximum Acceptable Difference'*	Calculated **Standard Error of measurement**
25% WC	**9.0% WC**

de Vet 2006

Maximum acceptable differences and corresponding SEM

a) Expectation by stakeholders	
Expected 'Maximum Acceptable Difference'*	Calculated **Standard Error of measurement**
25% WC	**9.0% WC**

b) Observed in the RELY studies			
		Observer **‘Standard error of measurement’**	Calculated **'Maximum Acceptable Difference'.**
Alternative work	RELY 1	**24.6% WC**	**68.1% WC**
	RELY 2	**19.4% WC**	**53.9% WC**

de Vet 2006

Summary

Compared to RELY 1, in RELY 2

1) No improvement in reliability

Experts have a low ability to distinguish claimants with mild, moderate, severe limitations in WC

2) Significant improvement in agreement among experts:

- Proportion of experts reaching 'the similarity criterion' increased by 20% ('*similarity criterion*')

- Dispersion between experts reduced by 20% (*SEM*)

3) Nevertheless: The differences in WC judgments between experts remain substantially below the expectations of the stakeholders

Where to go from here?

Engaged cooperation of many people only made the RELY studies possible.

the psychiatrists

the patients

the MEDAS institutes

the employees of the IV office in Zürich

the monitoring Group

the FIP Group

the associations for the disabled

the professional societies

the lawyers and judges

the insurers

A big Thank You goes to my colleagues and friends

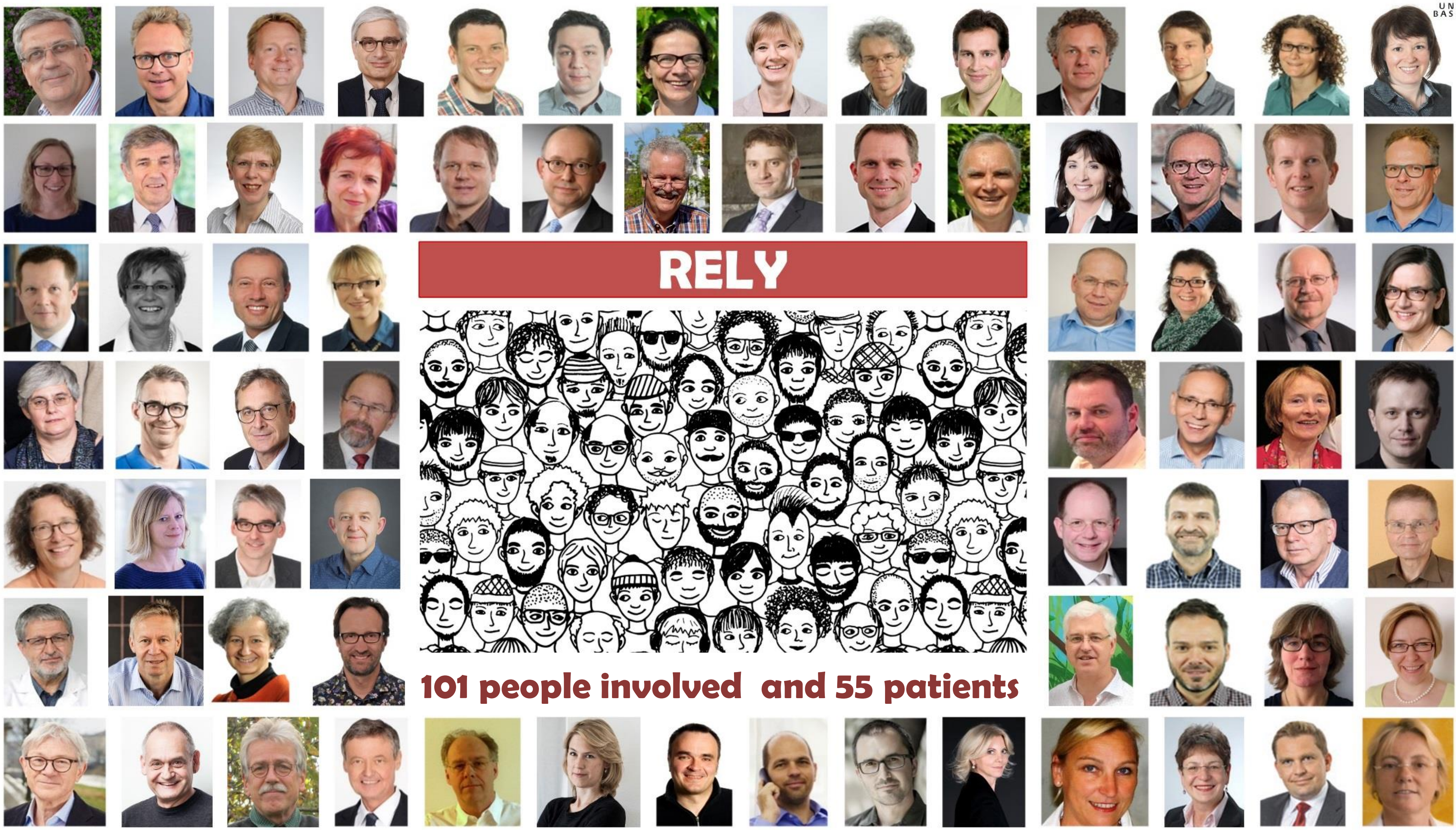
Wout de Boer, Katrin Fischer, David von Allmen, Monica Bachmann, Nicole Vogel, Jason W. Busse, Thomas Zumbrunn

the members of the FIP Group

Renato Marelli, Martin Eichhorn, Ulrike Hoffmann-Richter, Joerg Jeger, Ralph Mager, Etienne Colomb, Heinz J. Schaad

the funders

Swiss National Science Foundation,
Federal Social Insurance Office, Suva



RELY

101 people involved and 55 patients

Prof. Regina Kunz • Dr. David von Allmen • Prof. Katrin Fischer • Dr. med. Wout de Boer • Dr. med. Renato Marelli • Dr. med. Martin Eichhorn • Thomas Zumbrunn • Dr. med. Jörg Jeger • Dr. med. Ulrike Hoffmann-Richter • Prof. Ralph Mager • Dr. med. Etienne Colomb • Dr. med. Heinz Schaad • Dr. Monica Bachmann • Nicole Vogel • Prof. Jason Busse • Dr. med. Oskar Bänziger • Brigitte Walter Meyer • Sacha Röschard • Dr. med. Stefan Schandelmaier • Prof. Gordon Guyatt • lic, iur. Yvonne Bollag • Regina Altermatt • Corinne Schraner • Daniel Hess • Andrea Leibold • Dr. med. Ronald Walshe • Heidrun Demirden • Silvia Joder • Josée Staff • Astrid Palca • Dr. med. Roderich Koesel • Sarah Kedzia • Raphaël Dettwiler • Prof. Wolf Langewitz • Helena Langewitz • Dr. med. Olaf Hentrich • Claudia Bretscher • Dr. Andreas Brunner • Dr. med. Walter Gekle • Martin Schilt • Prof. Ueli Kieser • Dr. Volker Pribnow • Martin Reinert • Dr. med. Fulvia Rota • Dr. med. Rita Schaumann-von Stosch • Michael Stiebel • Dr. Andreas Traub • Marc Gysin • Peter Eberhard • Dr. med. Marco Bachmann • Dr. med. Roman Fischer • Dr. med. Natalie Franke • Dr. med. Jan Felix Hoffmann • Dr. med. Andreas Moldovanyi • Dr. med. Konstantin Moskvitin • Dr. med. Joachim Nelles • Dr. med. Thomas Weber • Dr. med. Peter Keel • Dr. med. Martina Korthal Altes • Dr. med. Tim Niemeyer • Dr. med. Thomas Ihde • Dr. med. Heribert Pizala • Dr. med. Felix Schwarzenbach • Dr. med. Vreni Häller • Dr. med. Karen Fürstenau • Dr. med. Armin Walter • Dr. med. Andreas Linde • Dr. med. Thomas Fellmann • Dr. med. Andres Howald • Dr. med. Christoph Feinendegen • Dr. med. Gerhard Mohr • Dr. med. Arno Zormann • Dr. med. Julia Röseler • Dr. med. Markus Guzek • Dr. med. Bernard Minder • Dr. med. Ueli Blumer • Dr. med. Axel Wallossek • Dr. med. Monika Diethelm-Knoepfel • Dr. med. Lars Hermann • Dr. med. Elmar Meichtry • Dr. med. Andreas Kaldune • Dr. med. Christoph Lenk • Dr. med. Marita Manheim • Dr. med. Stefan Freidel • Dr. med. Beat Schaub • Dr. med. Daniel Thommen • Dr. med. Michael Huber • Dr. med. Helen Klieber • Dr. med. Thomas Cotar • Dr. med. Sabina Kenk Meisser • Frau Beate Martin • Nicole Bruni •