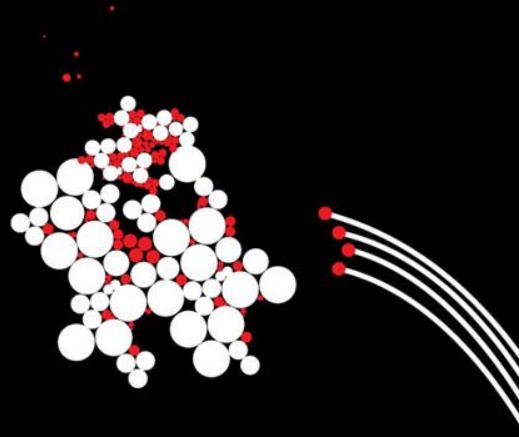
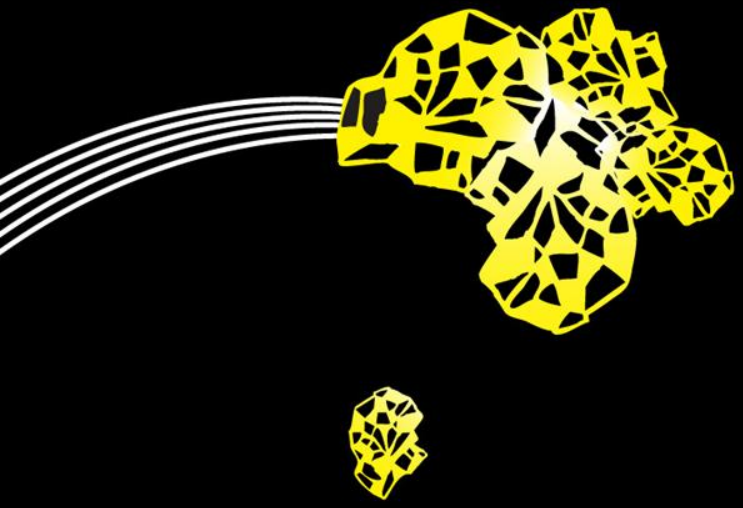
UNIVERSITY OF TWENTE.



USING IRT TO CROSS-CALIBRATE OUTCOME MEASURES

PETER TEN KLOOSTER



LINKING OUTCOME MEASURES

WHY IS IT USEFUL?

- Increasing use of patient-reported outcomes
- Different instruments or subscales intended to measure the same construct, such as physical functioning

  → More than … validated measures of physical functioning in rheumatoid arthritis (Oude Voshaar et al, submitted)
- Difficult to compare scores obtained with different instruments

  → Different items, response categories, scoring procedures, …
- Linking: Converting total scores from one instrument to scores on another instrument

  → Interpretation (patients, populations)

  → Meta-analyses

LINKING, EQUATING, CROSS-CALIBRATING

IS IT ALL THE SAME?

- 5 requirements for equating[1,2]
  - → Equal constructs (e.g., depression)
  - → Equal reliability
  - → Symmetry (linking function A→B = B→A)
  - → Equity (same test difficulty and expected distribution)
  - → Population invariance

- Hierarchy of linking
  - →From 'true' equating (all requirements are met) to prediction (none of the requirements are met)

[1] Dorans, N.J. (2004). *Applied Psychological Measurement, 28*(4), 227-246.

[2] Dorans, N.J. (2007). *Quality of Life Research,16*(Suppl 1), 85-94.

LINKING, EQUATING, CROSS-CALIBRATING

IS IT ALL THE SAME?

- Exact term used probably not so important, however
  - → Type of linking has consequences for degree of interchangeability
  - → Provides key assumptions that should be checked
- IRT has some distinct advantages for 'weaker' forms of linking
  - → Inherent ability to place items on a common underlying metric
  - → Allows integrative / robust checks for different requirements
  - → Can be used with different data collection designs (e.g., incomplete instrument designs)
- However, strong model assumptions!

EXAMPLE: CROSS-CALIBRATING THE PF-10 & HAQ-DI

PETER TEN KLOOSTER; JAKOB BJORNER; BARBARA GANDEK; MATTHIAS ROSE; JOHN WARE, JR

- SF-36 Physical Functioning scale (PF-10) and Health Assessment Questionnaire Disability Index (HAQ-DI)
- Widely used scales for measuring physical functioning in general populations and specific conditions
- PF-10

  → 10 items from the generic SF-36

  → 3-point response scale from 1 (yes, limited a lot) to 3 (no, not limited at all)

  → Summed scores linearly transformed to range between 0 and 100

PF-10

3. **The following questions are about activities you might do during a typical day. Does <u>your health now limit you</u> in these activities? If so, how much?**

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. <u>Vigorous activities</u>, such as running, lifting heavy objects, participating in strenuous sports	☐ 1	☐ 2	☐ 3
b. <u>Moderate activities</u>, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐ 1	☐ 2	☐ 3
c. Lifting or carrying groceries	☐ 1	☐ 2	☐ 3

EXAMPLE: CROSS-CALIBRATING THE PF-10 & HAQ-DI

PETER TEN KLOOSTER; JAKOB BJORNER; BARBARA GANDEK; MATTHIAS ROSE; JOHN WARE, JR

- HAQ-DI
  - → 20 items
  - → 4-point rating scale from 0 (without any difficulty) to 3 (unable to do)
  - → Average scores between 0 and 3

HAQ-DI

HAQ Disability Index:

In this section we are interested in learning how your illness affects your ability to function in daily life. Please feel free to add any comments on the back of this page.

Please check the response which best describes your usual abilities OVER THE PAST WEEK:

	Without ANY difficulty [0]	With SOME difficulty [1]	With MUCH difficulty [2]	UNABLE to do [3]
DRESSING & GROOMING				
Are you able to:				
-Dress yourself, including tying shoelaces and doing buttons?	☐	☐	☐	☐
-Shampoo your hair?	☐	☐	☐	☐

EXAMPLE: CROSS-CALIBRATING THE PF-10 & HAQ-DI

PETER TEN KLOOSTER; JAKOB BJORNER; BARBARA GANDEK; MATTHIAS ROSE; JOHN WARE, JR

- HAQ-DI:
  - → 20 items
  - → 4-point rating scale from 0 (without any difficulty) to 3 (unable to do)
  - → Average scores between 0 and 3
- Highly correlated and sufficiently unidimensional

DEVELOPMENT OF THE CROSSWALK

METHODS

- Large sample of US respondents that completed both scales
- Simultaneously calibrated using the 2-parameter graded response model
- Summed-score linking approach[3]

  → Determine IRT score for each possible observed (summed) score on each scale

  → Match observed scores on both scales to each other using the closest corresponding IRT score
- Check of requirements during IRT calibration
- Crosswalk for transforming total PF-10 scores into HAQ-DI scores and vice versa

[3] Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). *Psychological Assessment, 12*, 354-359.

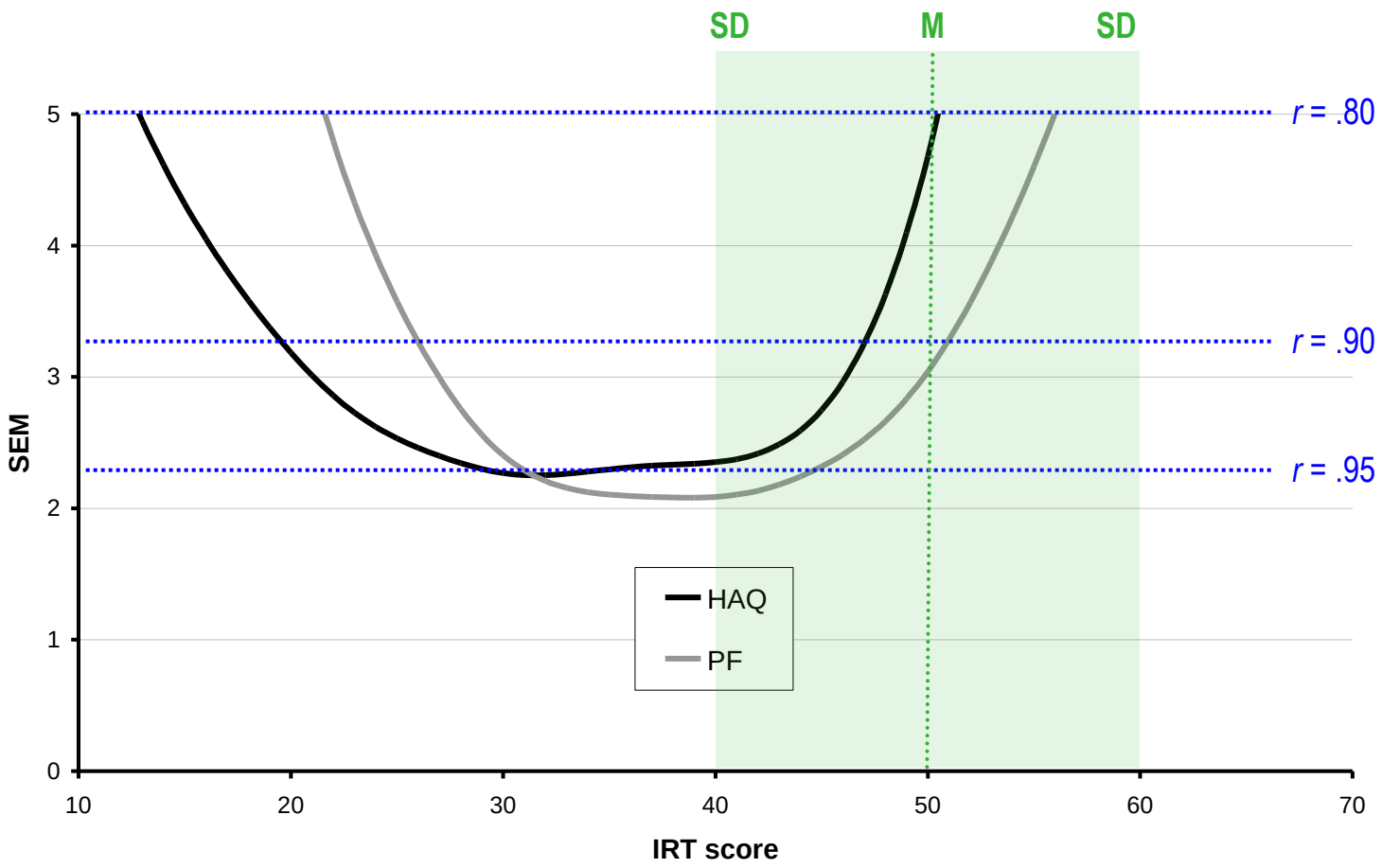
DEVELOPMENT OF THE CROSSWALK

RESULTS

- Items could be well-fitted to the model → unidimensionality
- Reliability and measurement precision (across the underlying construct → test difficulty)

DEVELOPMENT OF THE CROSSWALK

RELIABILITY AND TEST DIFFICULTY



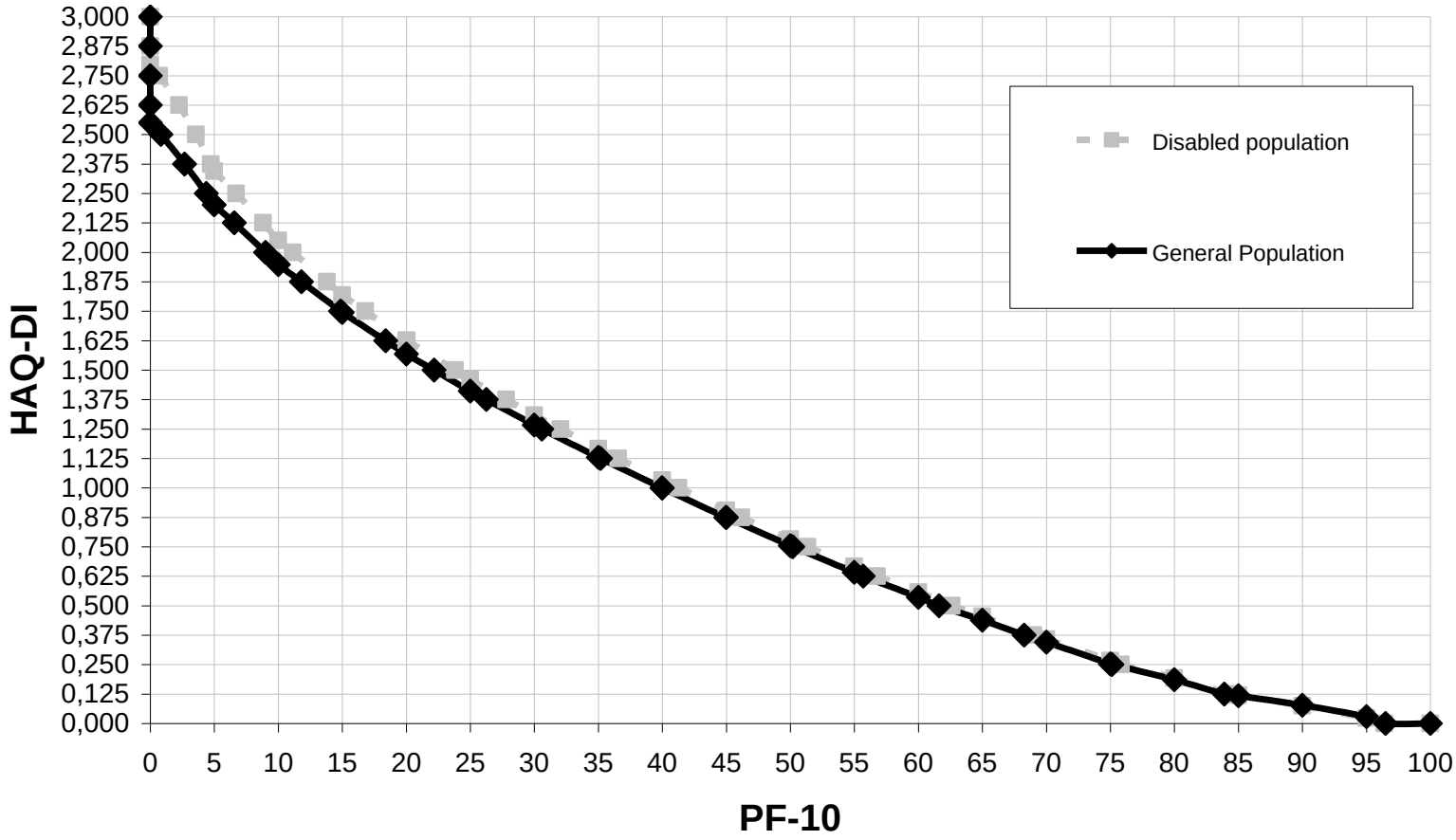
DEVELOPMENT OF THE CROSSWALK

THE CROSSWALK TABLE

PF-10 score	IRT score (θ)	HAQ-DI score	HAQ-DI score	IRT score (θ)	PF-10 score
0	-2.744	2.550	3.000	-3.538	0.000
5	-2.361	2.201	2.875	-3.237	0.000
10	-2.122	1.948	2.750	-3.020	0.000
15	-1.940	1.744	2.625	-2.839	0.000
20	-1.787	1.569	2.500	-2.682	0.819
25	-1.650	1.412	2.375	-2.540	2.669
30	-1.524	1.266	2.250	-2.409	4.373
35	-1.403	1.130	2.125	-2.286	6.563
40	-1.286	0.999	2.000	-2.169	9.015
45	-1.171	0.874	1.875	-2.056	11.817
50	-1.055	0.754	1.750	-1.945	14.859
55	-0.938	0.641	1.625	-1.836	18.400
…	…	…	…	…	…
…	…	…	…	…	…

DEVELOPMENT OF THE CROSSWALK

CONVERSION GRAPH



CROSS-VALIDATION OF THE CROSSWALK

METHODS

- Baseline data of 650 Dutch patients with RA that completed both scales
- Agreement of observed and estimated scores:

  → Intraclass Correlation Coefficient

  → Bland-Altman plots
- Relative validity in discriminating between levels of health

  → ANOVA, F-statistics

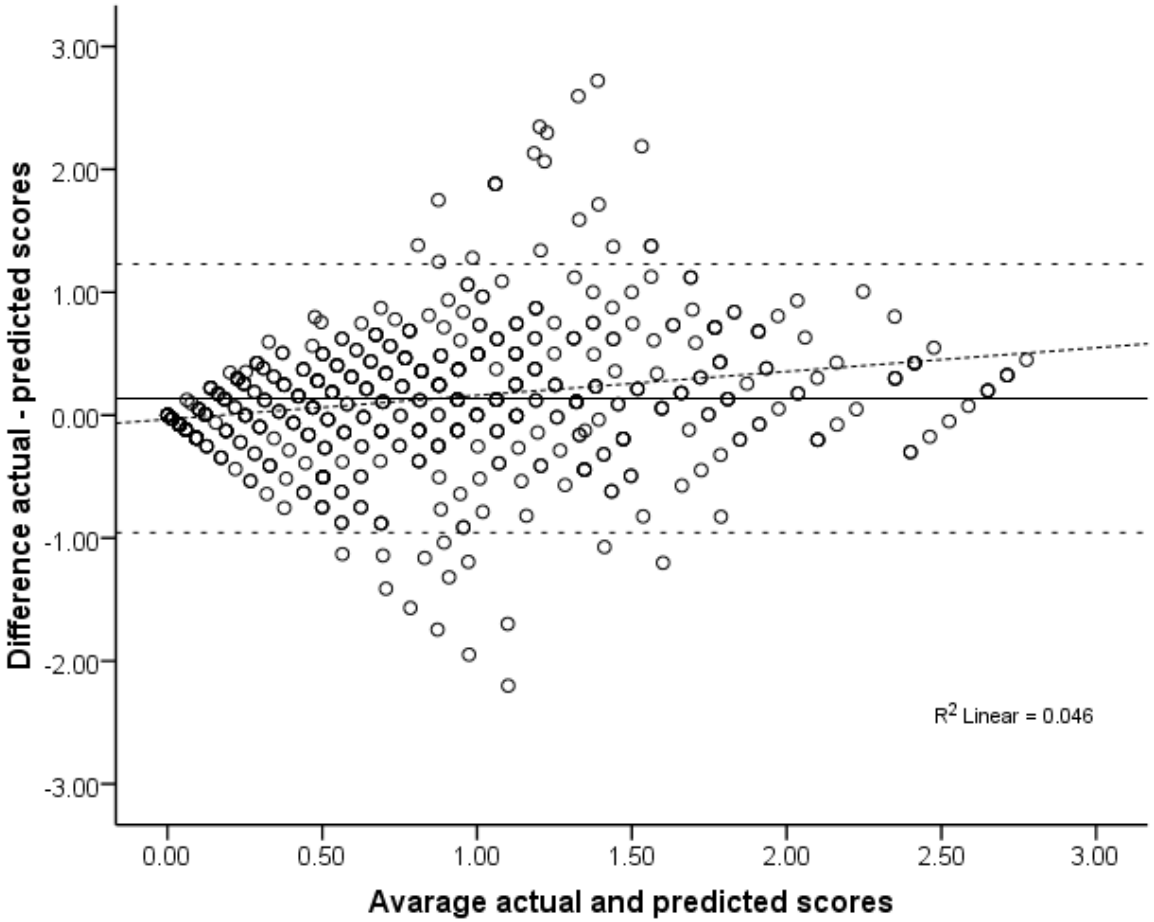
CROSS-VALIDATION OF THE CROSSWALK

AGREEMENT

- ICC = 0.67
- Bland-Altman analyses

CROSS-VALIDATION OF THE CROSSWALK

AGREEMENT



CROSS-VALIDATION OF THE CROSSWALK

RELATIVE VALIDITY

	Excellent (n = 27)	Very good (n = 188)	Good (n = 228)	Fair (n = 42)	Poor (n = 27)	F*	RV
HAQ-DI							
Actual score	0.29	0.38	0.64	1.19	1.69	62.06	1.00
Predicted score	0.34	0.44	0.58	1.03	1.50	51.09	0.82
PF-10							
Actual score	77.50	71.79	62.26	42.01	27.24	58.53	1.00
Predicted score	78.44	73.18	60.85	37.16	21.39	69.41	1.19

* All F-values significant at p<0.001
RV = relative validity (ratio of F-statistic compared with actual score)

CONCLUSIONS

USING IRT TO CROSS-CALIBRATE OUTCOME MEASURES

- Possible to develop a reasonably accurate and valid crosswalk for converting HAQ-DI scores into PF-10 scores and vice versa
- Crosswalk can be used to estimate group-level scores, but should not be used to directly convert individual patient scores
- IRT offers a powerful and integrative approach to link scores from different patient-reported outcomes measuring the same construct