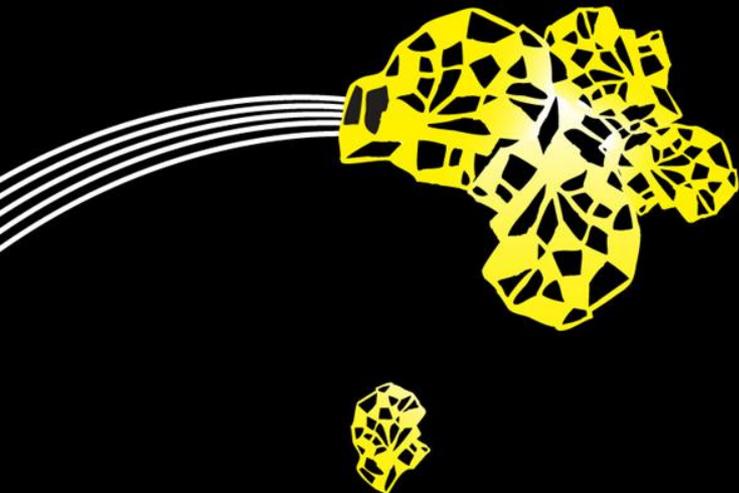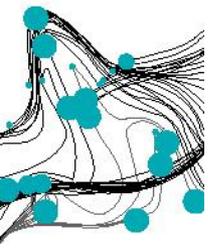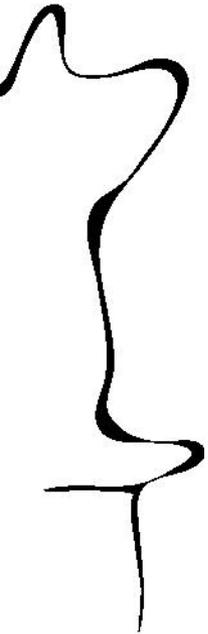# EMERGENCE OF IRT AND CAT IN HEALTH OUTCOMES MEASUREMENT

PETER TEN KLOOSTER

HEALTH OUTCOMES MEASUREMENT

- Outcome measures provide an indication of the severity of the condition and the effects of treatment

  → Increasingly important to inform practice and policy decisions

- Patient-reported outcomes vs. clinical measures

- CMs: Objective → Disease severity

- PROs: Subjective → Patient-perspective on the impact of the disease (pain, fatigue, disability, etc.)

- Psychometrics

  → Measuring (quantifying) "unobservable" characteristics of people

  → Construction and validation of measurement instruments, such as questionnaires and tests

IRT AND CAT IN HEALTH OUTCOMES MEASUREMENT
A NEW PARADIGM?

Quality of Life Research, **6**, pp. 595–600

Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing

Item Response Theory and Health Outcomes Measurement in the 21st Century

Quality Measurement and Improvement

Generic Health Measurement: Past Accomplishments and a Measurement Paradigm for the 21st Century

ITEM RESPONSE THEORY
WHAT IT'S NOT

- Not new

  → Originated in 1960s in educational testing (Lord & Novick, Rasch)

- Not a single technique

  → A framework consisting of several mathematical models for examining measurement properties

- Not a replacement for all existing psychometric techniques

  → Supplements classical test theory methods (e.g., factor analysis, Cronbach's α)
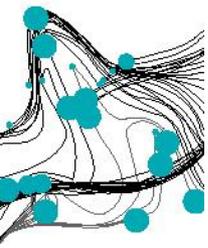
ITEM RESPONSE THEORY
WHAT IT IS

- Aka: Latent trait theory, strong true score theory, modern test theory
- Focus on item-level information
  - → Models that relate an individual's response to an item to the underlying attribute (latent variable, θ) being measured by the scale
  - → Estimating the probability that a person's response to an item will lie in a particular category
  - → Separates person parameters (ability) and item parameters (e.g. difficulty)
  - → Allows for sample and item invariant measurement

- ..

ITEM RESPONSE THEORY
WHAT IT IS
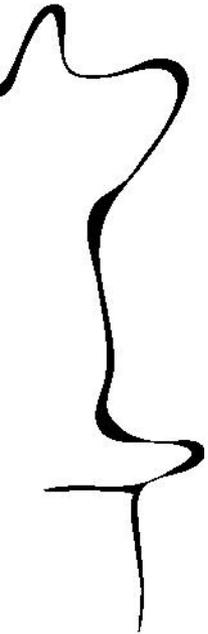
- Family of different models for different number of item parameters and types of response options
- Number of parameters
  → One-parameter (or Rasch) models: Item difficulty
  → Two-parameter models: + Item discrimination
  → Three-parameter models: + Pseudo guessing
- Response options
  → Dichotomous (e.g. Yes/No) vs. polytomous (e.g. Likert scales)
  → Ordered vs. unordered response options
- Unidimensional vs. Multidimensional (presentation Nikolaus)
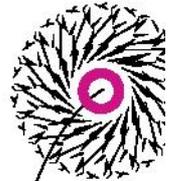
ADVANTAGES OF IRT
WHY USE IT?

- Also applicable to PROs (and clinical measures? presentation Siemons)
- Several distinct advantages
  → Interval-level scoring of a PRO
  → Better measurement precision at the extreme ends of the scale
  → Analysis of unidimensionality
  → Examining the appropriateness of response options
  → Examining the hierarchical nature and measurement range of a PRO
  → Identifying redundant items and 'gaps' in measurement
  → Examining measurement precision (reliability) across the underlying scale
  → Examining differential item functioning (DIF)

ADVANTAGES OF IRT
WHY USE IT?

- Especially useful for

  → Developing new PROs

  → Developing short forms or alternate forms targeted at specific populations

  → Evaluating and comparing existing PROs

  → Linking scores on different PROs (presentation Ten Klooster)

  → Cross-cultural validation of PROs

  → Developing computerized adaptive tests (Presentations Van der Linden, Glas, Krishnan, Rose)
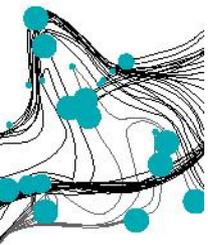
DISADVANTAGES OF IRT
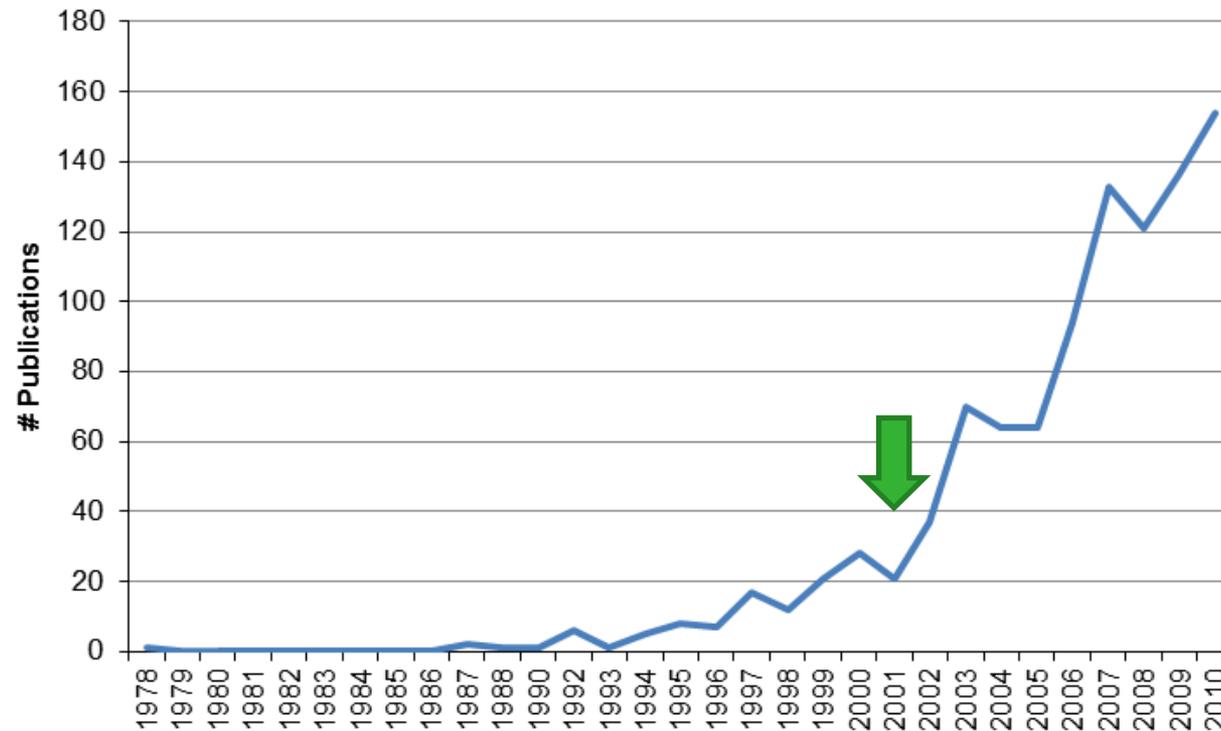WHY DOESN'T EVERYBODY USE IT?

- Strong model assumptions

- Large sample sizes required for more elaborate models

- Advanced statistical skills

- Not available in mainstream, user-friendly software packages

- Scopus: "item response theor*" OR "item response model*" OR rasch AND health

WHAT IS IT ACTUALLY USED FOR?
STATE OF THE ART IN RHEUMATOLOGY

- Rheumatology relatively advanced in multidimensional measurement of disease outcomes
- Systematic review of the application of IRT to patient-reported and clinical outcome measures in rheumatology[1]
- English-language articles that used some form of IRT-based analyses
- Original research articles only
- >50% rheumatic condition or separately analyzed
- Checklist for rating general study information and IRT-specific information

[1] Siemons et al, *submitted*
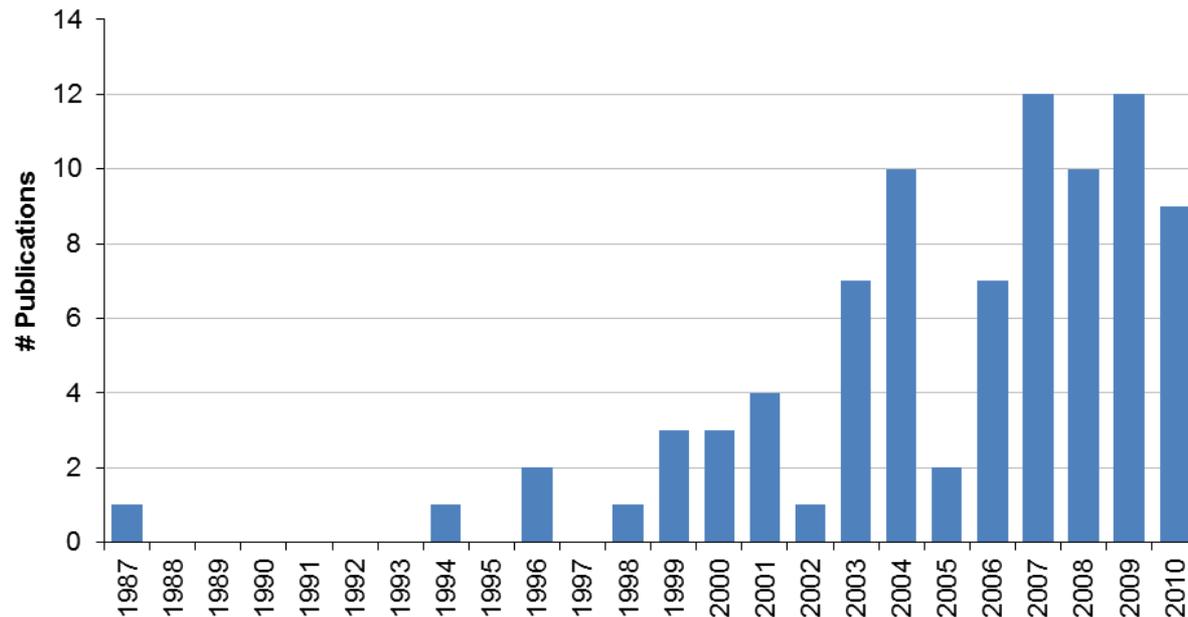
USE OF IRT IN RHEUMATOLOGY
RESULTS

- 85 eligible studies

  → 73 PRO studies

- Disease conditions

  → RA: 51%

  → OA: 34%

- Measurement domain

  → Overall physical functioning: 40%

  → Quality of life: 33%

- Sample size

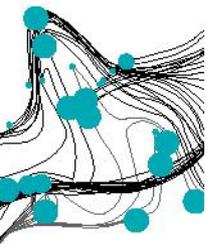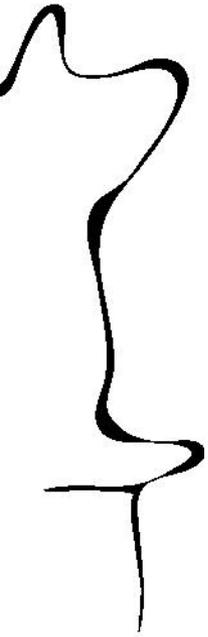 → n=18 to n=16,519

  → n>50 in 92% of studies

- Number of publications over the years

- PubMed search: Development / evaluation of PROs in 2010

    → 45 studies, 7 studies using IRT (~16%)

- Main goal of the studies

  → Development / evaluation new PRO: 32%

  → Evaluation existing PRO: 32%

  → Development alternate / short form: 14%

  → Cross-cultural validation: 8%

  → Development item bank / CAT: 6%

  → Equating or cross-calibrating different PROs: none

USE OF IRT IN RHEUMATOLOGY
RESULTS

- IRT model

  → Rasch model: 84%

  → Two-parameter model: 16%

  → Model choice explained: 24%

- Software

  → Bigsteps / Winsteps: 32%

  → RUMM: 33%

- IRT assumptions tested

  → Model fit: 92%

  → Unidimensionality: 77%

  → Local independence: 15%

UNIVERSITY OF TWENTE.

USE OF IRT IN RHEUMATOLOGY
RESULTS

- Specific IRT analyses

  → Hierarchical distribution of items / persons: 66%

  → Person / item separation or reliability: 59%

  → Differential item functioning: 58% (gender, age, disease duration, time)

  → Appropriateness of response options: 33%

  → Measurement precision across the scale: 12%

UNIVERSITY OF TWENTE.

IRT AND CAT IN HEALTH OUTCOMES MEASUREMENT
IS IT REALLY THE NEW PARADIGM?

- Marked increase in the use of IRT in the past decade

  → Classical test theory methods still dominant to date

- Mostly used for the development and evaluation of traditional (paper and pencil) questionnaires

  → Item banking and CAT still uncommon

- Not all advantages of IRT are used optimally

  → Measurement precision across the scale

  → Equating or cross-calibrating scores on different PROs

  → Cross-cultural validation of PROs

- IRT and CAT appear to be promising and increasingly popular tools to better measure health outcomes, but their full potential has yet to be explored

UNIVERSITY OF TWENTE.