

# Scale construction and evaluation in practice: Factor analysis versus item response theory



Janke C. ten Holt, Marijtje A. J. van Duijn & Anne Boomsma

University of Groningen  
j.c.ten.holt@rug.nl

## Introduction

### FA and IRT for scale analysis

Factor analysis (FA) and item response theory (IRT) are two types of models used to assess to what extent a group of items reliably measures the latent variable(s) researchers are interested in.

#### (Standard) FA

- Continuous item variables
- Linear relation between LV and items
- Model examples: CCFA, ECFA, OMG, PCA

#### IRT

- Categorical item variables
- Nonlinear relation between LV and items
- Model examples: Rasch, 2PLM, GRM, Mokken

Although certain kinds of FA and IRT models are completely equivalent, the more common variants have typical differences.

### Past research comparing FA and IRT

- **Mathematically:** Mehta & Taylor, 2006; Takane & De Leeuw, 1987; see also Kamata & Bauer, 2008
- **Simulated data:** Knol & Berger, 1991; Wirth & Edwards, 2007
- **Empirical data:** Glöckner-Rist & Hoijtink, 2003; Moustaki, Jöreskog, & Mavridis, 2004
- **Simulated and empirical data:** Jöreskog & Moustaki, 2001
- **With regard to measurement equivalence:** Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman & Pugh, 1993

### Central question: What is used in practice and why?

#### Method

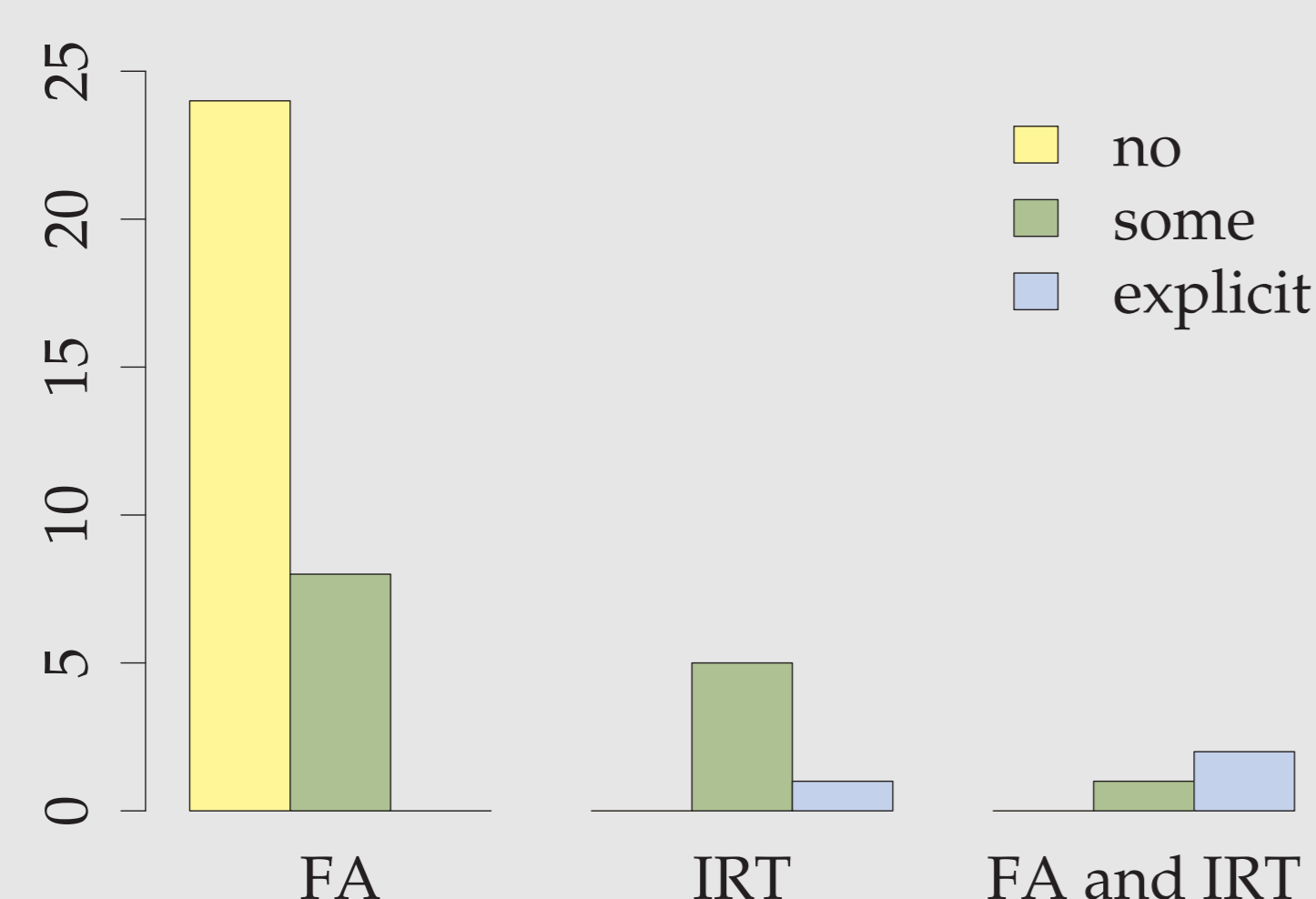
Review of 41 studies

- Concerning scale construction/evaluation
- Published in 2005 in
- *Psychological Assessment* ( $n = 13$ )
- *European Journal of Psychological Assessment* ( $n = 13$ )
- *Educational and Psychological Measurement* ( $n = 15$ )

First, researchers' explicit motivation for applying FA or IRT is examined. Second, data characteristics are examined to try and extract some implicit motivation.

## Results

### Motivation for applying FA or IRT



- No motives are given.
- Some motives are given. E.g.: Benefits of IRT over CTT; EFA compared to CFA.
- Motives are given, *explicitly* mentioning both FA and IRT. E.g.: IRT instead of FA, because of skewed item distributions, or dichotomous items.

Note the little application of IRT.

### Characteristics of the data

	Type of applied analysis		
	FA ( $n = 32$ )	IRT ( $n = 6$ )	FA & IRT ( $n = 3$ )
No. of categories	2	4	1
>2	28	5	2
No. of dimensions	1	1	5
2	4	1	1
3	8	1	1
>3	13	1	1
Exploratory vs. expl. confirmatory	8	2	1
both	13	4	2

IRT *not* more often used for dichotomous data  
IRT more often used for unidimensional data  
Exploratory models used relatively often, despite presence of clear hypotheses about the structure of the data

### Software use

Software	Type of applied analysis		
	FA ( $n = 32$ )	IRT ( $n = 6$ )	FA & IRT ( $n = 3$ )
LISREL	12	1	1
AMOS, EQS, MPLUS, SCA	9	0	0
Various IRT packages	0	6	5
SAS, SPSS, STATVIEW, SYSTAT	4	1	1
No information	15	2	1

For CFA LISREL is most popular.  
For EFA, either no information is given about software use, or general statistical software is used.

### Investigation of model assumptions

#### In FA studies

- 19 studies: no investigation
- 9 studies: investigated properly
  - Item distributions examined and reported.
  - Adequate methods (robust, if necessary) are applied.
- 4 studies: considered to some extent
  - Item distributions not investigated, but robust estimators used.
  - Both robust and nonrobust analyses, but only reported nonrobust because of similar parameter estimates.

#### In IRT studies

- 4 studies: investigated properly
  - Unidimensionality assumption investigated
  - IRFs examined for monotonicity
  - Empirical IRFs compared to estimated IRFs
- 2 studies: no investigation

### Model fit

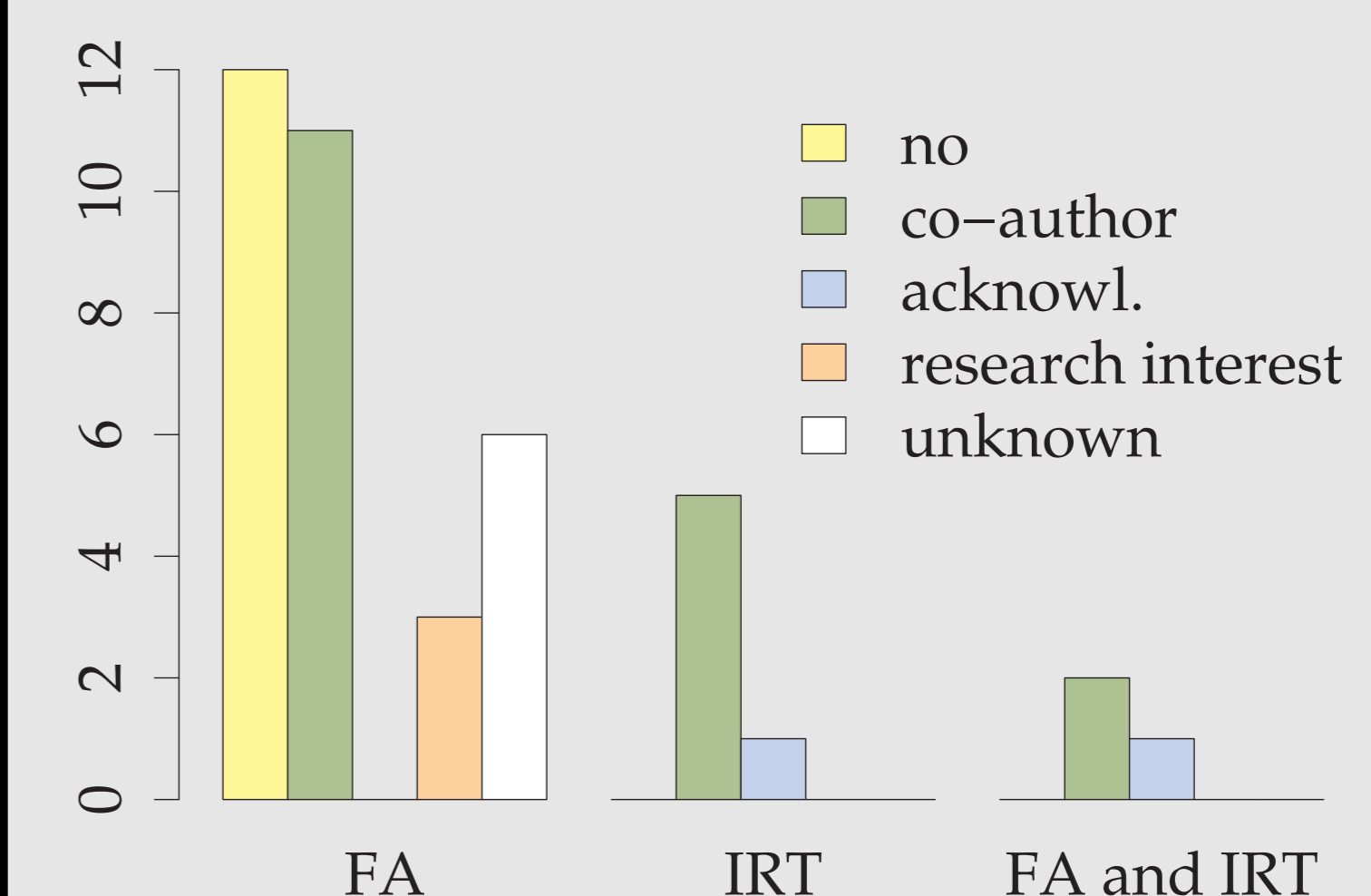
#### In FA studies

- CFA: Model fit tested formally usually with
  - RMSEA, GFI, CFI, TLI (NNFI)
- EFA: No formal test, but criteria to determine #factors and assignment of items to factors:
  - loadings > 0.30 or 0.40
  - # factors determined by screeplot, parallel analysis, eigenvalue > 1
  - in merely 5 (of 21) studies: interpretability as criterion

#### In IRT studies

- No formal tests reported
- Mokken analysis: Loevinger's  $H$  for scale strength
- Unidimensionality tested in 3 studies

### Methodological expert as co-author



Four degrees of involvement:

- No involvement
- A methodological expert *co-authored* the paper
- A methodological expert is *acknowledged* in a note
- One of the authors has a *research interest* in psychometrics or quantitative methods.
- *Unknown* indicates that no information could be retrieved from the website of the authors.

IRT is only applied when one of the authors is a methodological expert or an expert is acknowledged in a note.

## Discussion

### Summary

- FA applied far more often than IRT
- Little explicit motivation in studies
- Possible implicit motives:
  - Expectations about *dimensionality*: IRT applied to unidimensional data; multidimensional IRT software seems to be unknown.
  - FA is more *accessible*; IRT might require a methodological expert.

### Recommendations

- Researchers can take better advantage of their theories:
  - More frequent application of confirmatory techniques. When applying an exploratory model → cross-validate.
  - Add interpretability of factors and content of items to criteria of model evaluation.
- Evaluate model assumptions and report in the paper or on a website.

### Future research

- Both simulated and empirical comparisons of FA and IRT
  - Examine impact of violation of model assumptions
  - Extend past research by including *nonparametric* IRT in the comparison
- Examine differences between latent variable (factor) scores produced by different types of models.
- Examine how to combine exploratory and confirmatory approaches in FA and IRT