Data harmonization tutorial:Teaser for FH2019

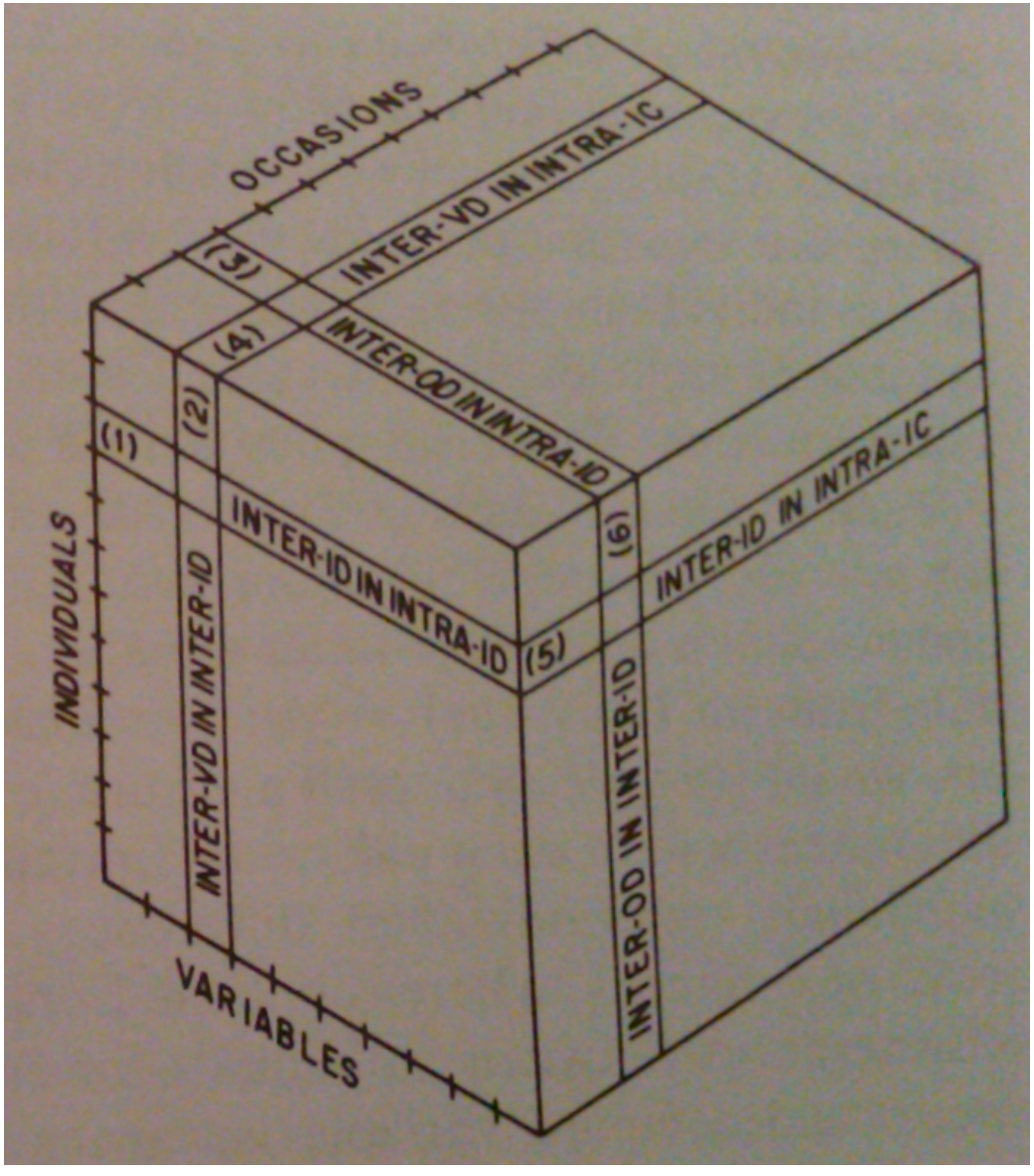
Alden Gross, Johns Hopkins
Rich Jones, Brown University

Friday Harbor Tahoe 22 Aug. 2018

Outline

- What is harmonization?
- Approach
  - Prestatistical harmonization - the accounting job
  - Apply the tool
  - Diagnostics/checks
- Tools
  - Distribution-based
  - IRT-based
  - Missing data

We want to do analyses



- Cross-sectional comparisons at a time point
- Longitudinal change in a variable over time
  - Individual changes
  - Inter-variable differences in intra-individual changes (change regressed on a covariate of interest)

Nesselroade and Baltes, 1979, per Buss, 1974

Statistical harmonization

- Harmonization is broad
  - Qualitative assessments of the comparability of measures
  - Statistical approaches to equate and link measurement scales or tests
- Different studies on the same topic often implement different measures due to:
  - Developmental differences in the target population
  - Investigator proclivities
  - Logistical issues
- Harmonization across data sources can help synthesize information across different sources

Integrative vs coordinated analysis

- Integrative data analysis (IDA): analysis of multiple datasets, together or in parallel, to address substantive research hypotheses
  - Together (pooled): Individual-participant meta-analysis
  - In parallel: Coordinated analysis, in which models are run separately in each data source
- Goals of harmonization of pooled data
  - Larger sample size to afford power for questions that cannot be addressed from individual data sources (e.g., genetics; interactions)
  - Address innovative questions that cannot be answered with one data source (e.g., what does cognitive change look like across the life span? Is education or study membership a stronger predictor of cognitive change?)

Approaches to harmonization (Griffith et al., 2012 AHRQ report)

- Prestatistical harmonization
  - Accounting work
  - Gather available test data
  - Evaluate test responses
  - Describe the sample
- Statistical harmonization approach for test equating
- Diagnostics

Pre-statistical harmonization

University of Southern California
Center for Systems and Software Engineering

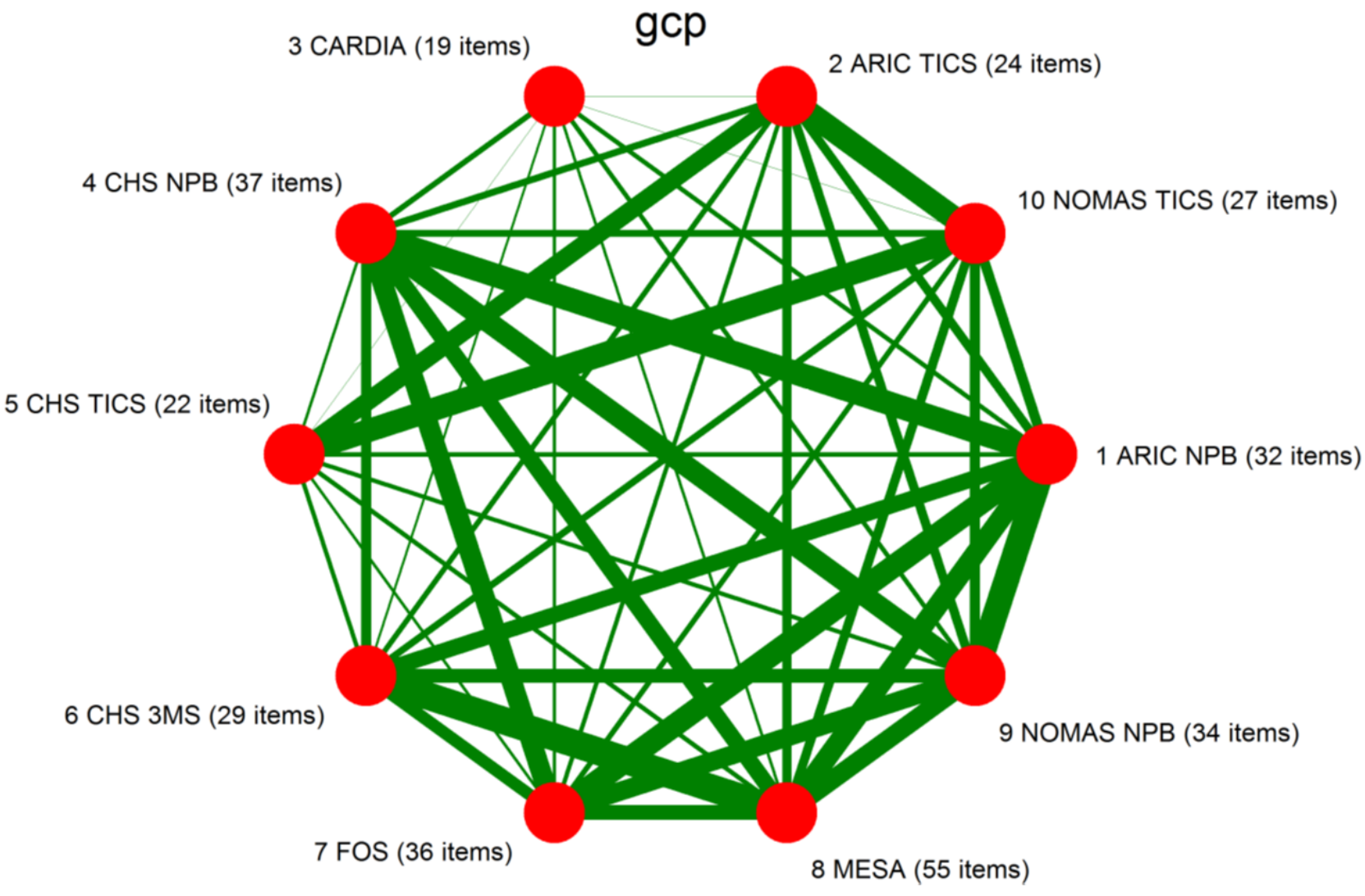
The Procrustean Bed

- **Procrustes: Greek Mythology**
  - **Rogue smith and bandit**
  - **Hostel with one-size-fits-all bed**
  - **Guests too small: stretch them to fit**
  - **Guests too large: lop off the offending parts**

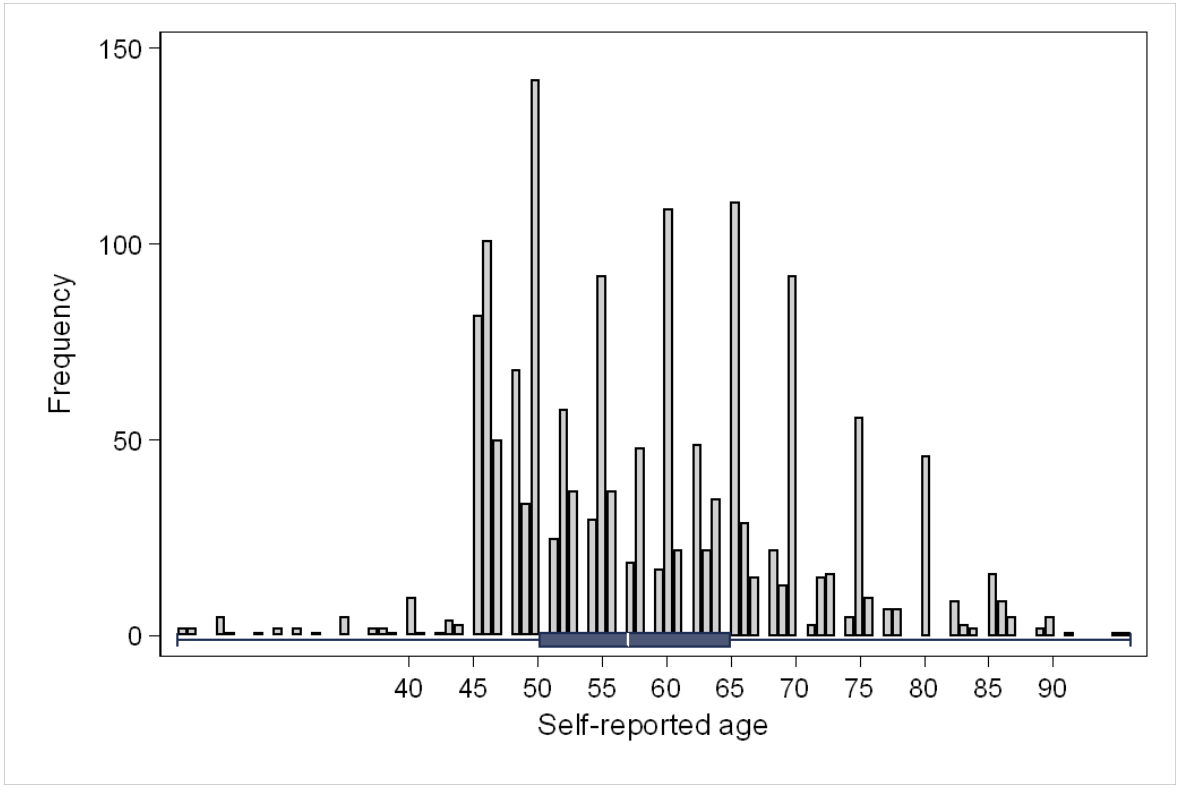
1/13/2014 © USC-CSSE 8

- Identify the items to be used
- Rename variables to common rubric
- Recode missing data codes
- Trim outliers (e.g., winsorize)
- Stretch out variable distributions
- Discretization
- Check for small cells
- Check for anchor items

Visualizing number of anchors

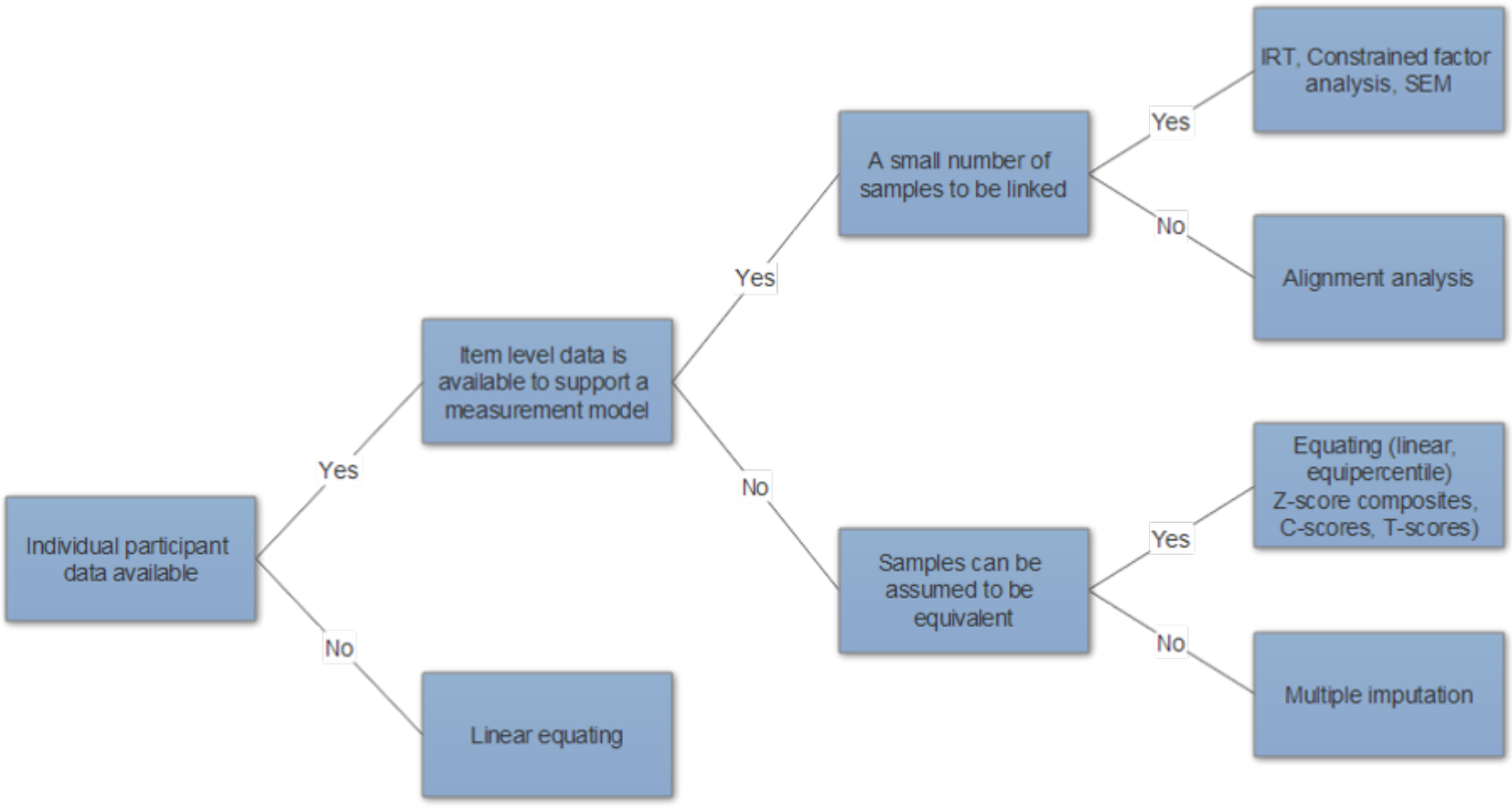


Challenges and lessons in harmonization: The Anna Karenina principle



- From Leo Tolstoy's Anna Karenina:
  - "Happy families are all alike; every unhappy family is unhappy in its own way."
- Our analog
  - Well-designed studies are all alike; every study with design features that present challenges for analysis presents challenges for analysis in its own way

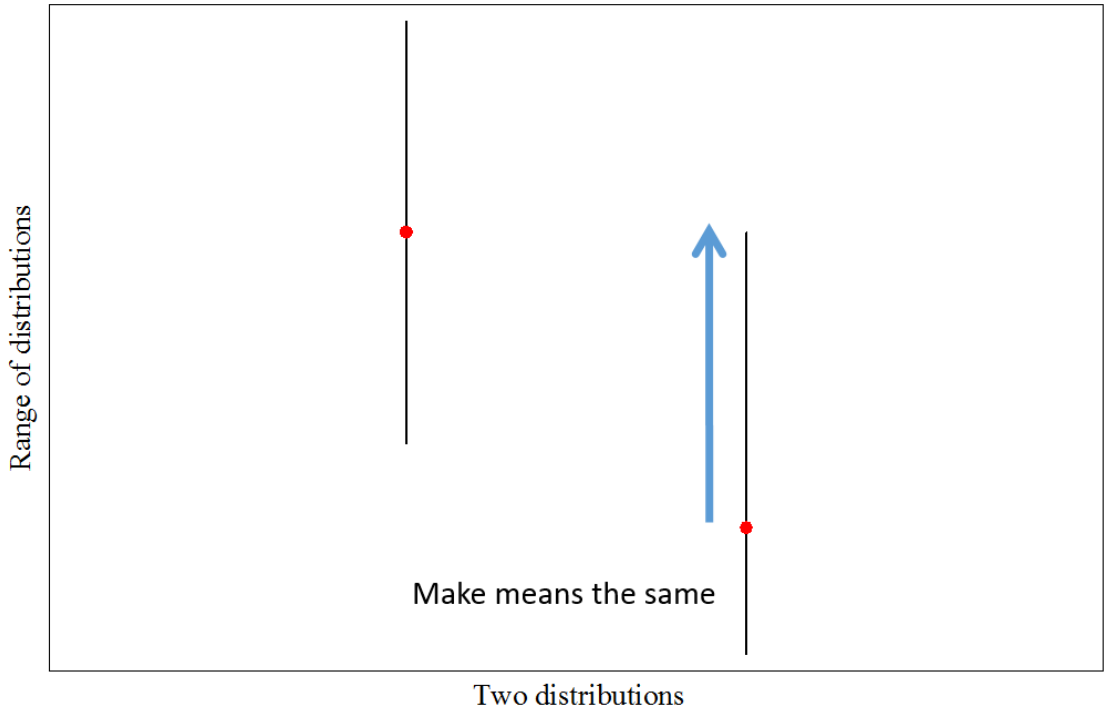
Apply the method



Tools for co-calibration

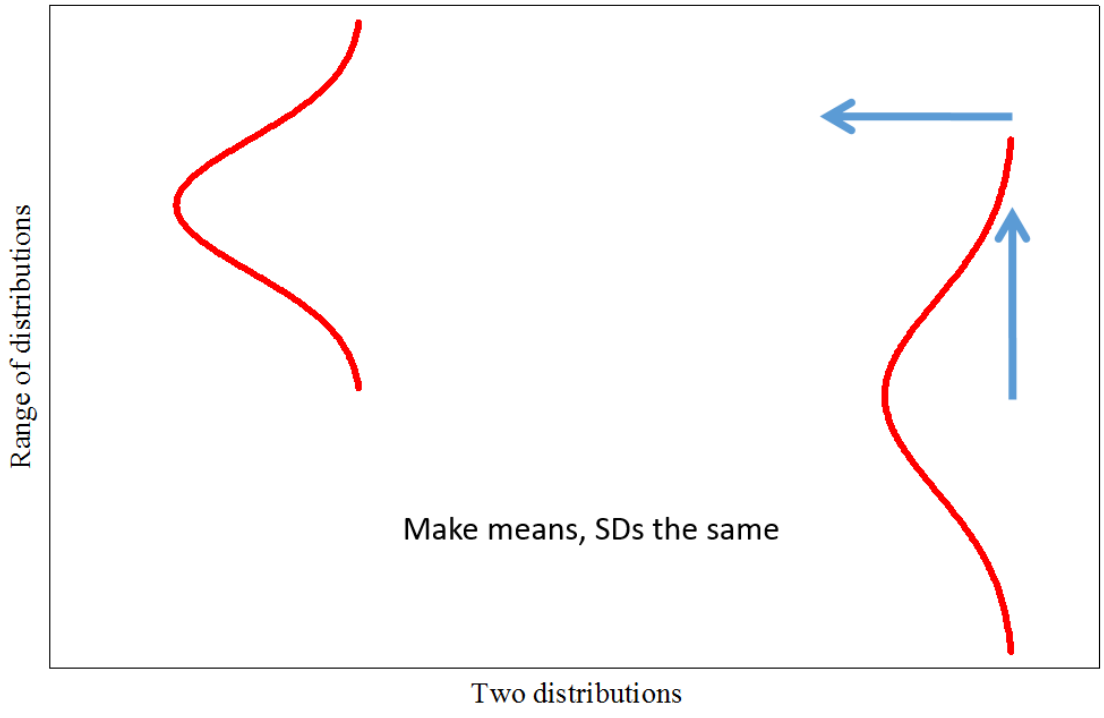
- Distribution-based
  - Goal: define a transformation of a test that returns the same cumulative probability plot as the other variable being compared
  - Mean equating
  - Linear equating
  - Equipercentile equating
- Item-based
  - Goal: define a transformation of a test that places it on the same metric as another
  - Item response theory
- Multiple imputation for missing data

Mean equating



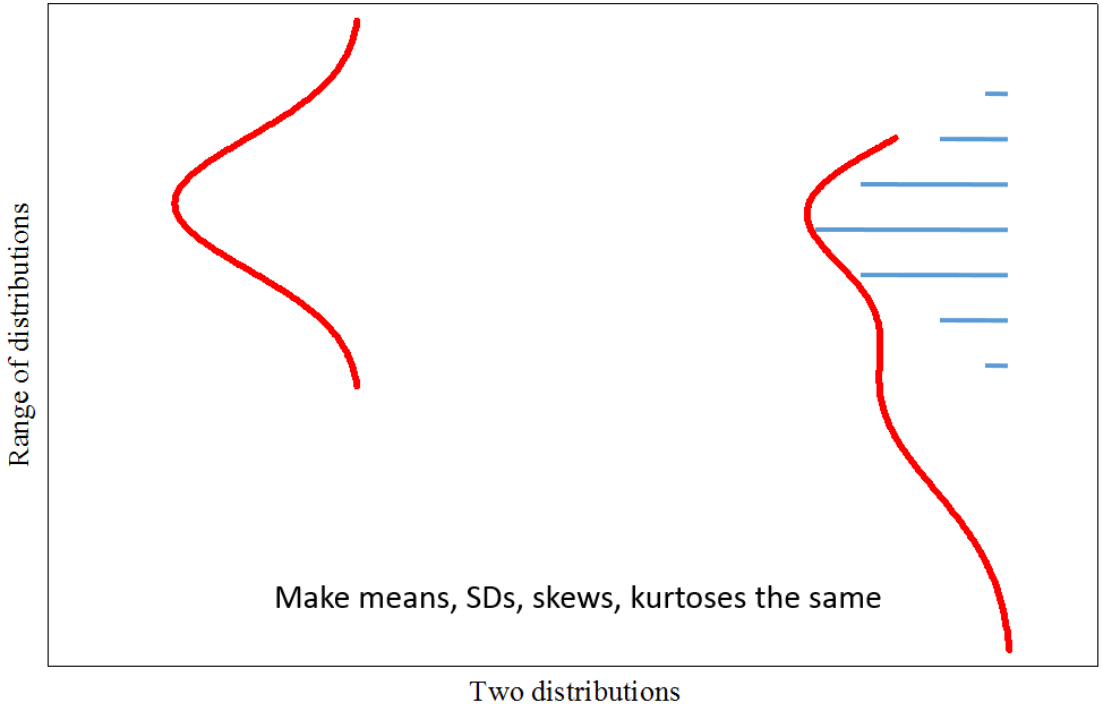
Relative position is defined by the absolute difference from the sample mean of a test, and each individual's score is changed by the same amount to equate the sample mean to that of a reference test

Linear equating



- Relative position is defined in terms of standard deviations from the group mean.
- Linear equating is accomplished by adjusting scores from the new form to be within the same number of standard deviations of the mean of the original form.

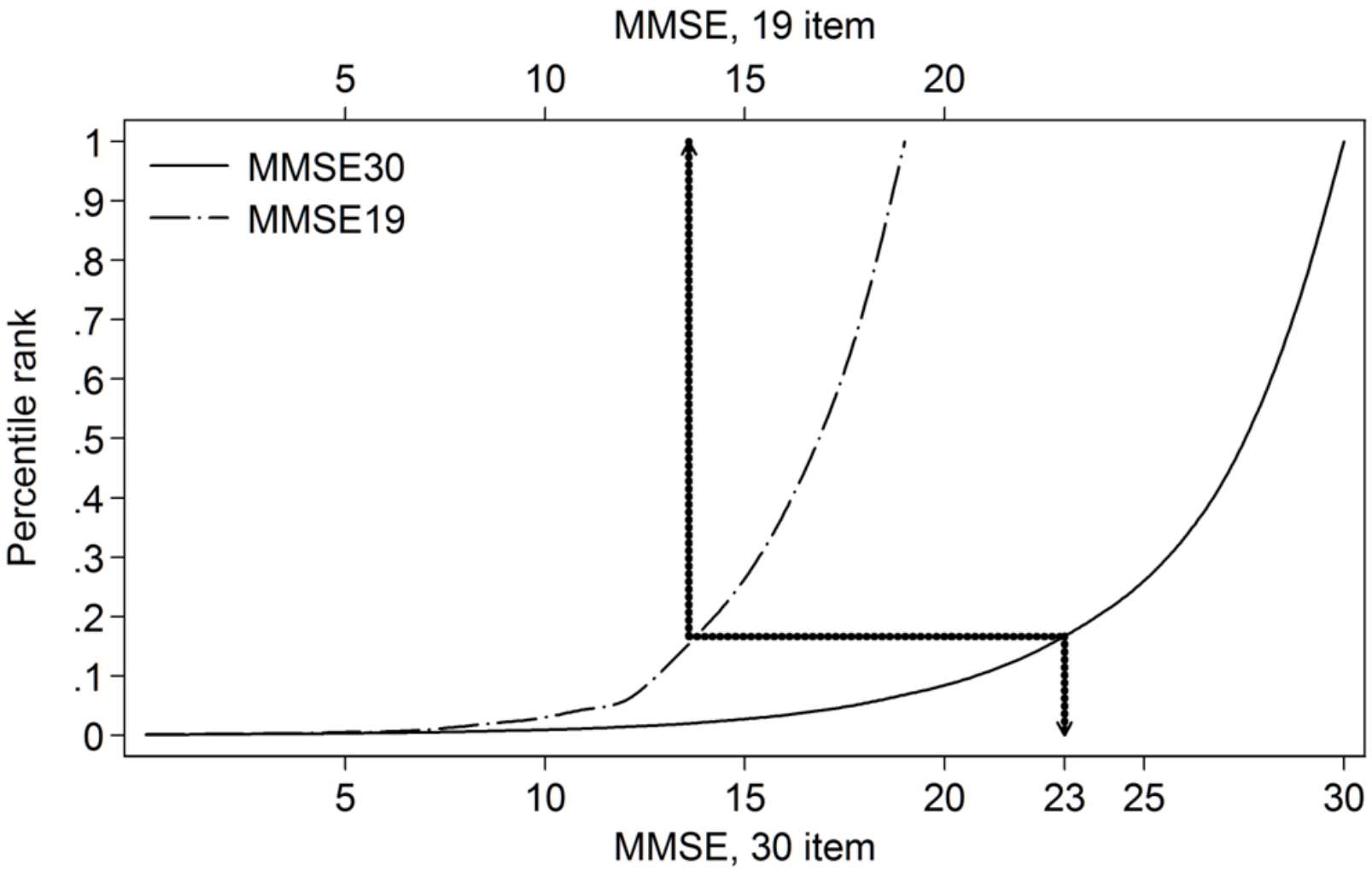
Equipercentile equating



- Defines relative position by a score's percentile rank in the group.
- Accomplished by identifying percentiles for scores on a reference test and transforming each score on a new test to the score on the reference with the same percentile rank

Equipercentile equating

Each score on one test (non-30 point MMSE version) is transformed to the score on the reference test (30-point MMSE) with the same percentile rank



How is it done?

- Mean equating
  - MeanEq_test2 = old_test2-(mean_test2 - mean_test1)
- Linear equating
  - LinEq_test2 = mean_test1 +(old_test2-mean_test2)/(SDtest1/SDtest2)
  - z scores!
- Equipercentile equating
  - R package: equate
  - `https://cran.r-project.org/web/packages/equate/equate.pdf`

Dangers of distribution-based methods

- They equate scales, not metrics
- Blunt force tools. They not only erase measurement differences, they can obliterate age differences and other differences that we wish to preserve
- Think carefully about what you equate
  - Same construct?
  - Is there sufficient variability to support the relative position?

Shoe size and MMSE(Mobilize Boston Study, N=807)

Shoe size	Equipercentiled MMSE (median)
3	6.1
4	11.1
5	16
6	20
7	24
8	26

Think carefully about what you equate!

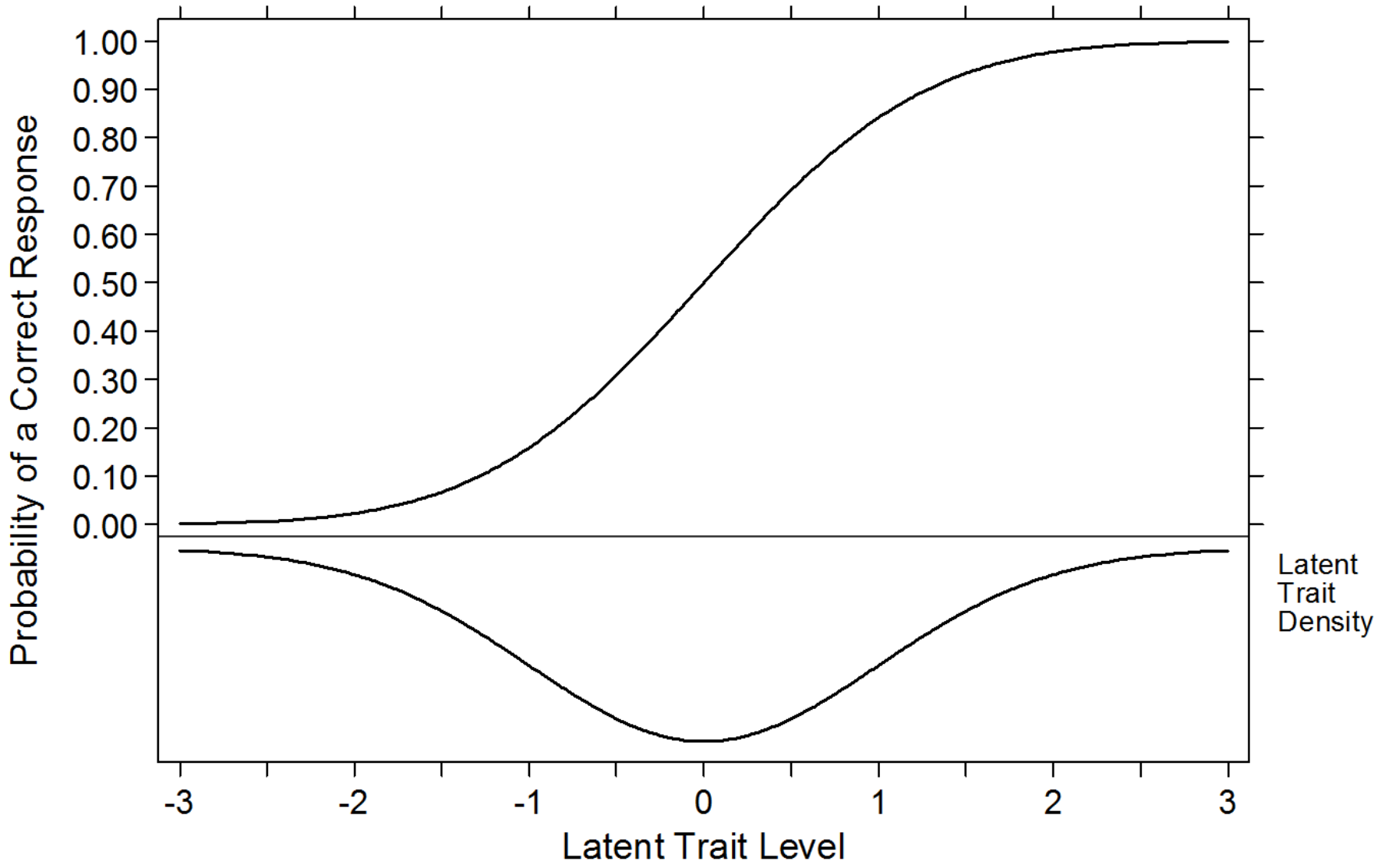
12	30
13	30
14	30
15	30

Among persons with an observed MMSE less than 24, the maximum equated shoe size is 7. So, to screen for dementia using shoe size, flag as possibly demented persons with a shoe size of less than 7

Tools for co-calibration

- Distribution-based
  - Goal: define a transformation of a test that returns the same cumulative probability plot as the other variable being compared
  - Mean equating
  - Linear equating
  - Equipercentile equating
- Item-based
  - Goal: define a transformation of a test that places it on the same metric as another
  - Item response theory
- Multiple imputation for missing data

IRT and the Item Characteristic Curve



Item response theory: An important assumption

- Exchangeability of items
  - Each item conveys information about the latent trait
  - Implication: Missingness on an item might be OK as long as other items are not missing. Less precision

Test name	ADNI	MAP	ROS	NACC	Cache	Tarenfl urbil	Semaga cestat	AddNeu roMed	ACT
MMSE	X	X	X	X	X	X	X	X	X*
Boston Naming Test	30-item	15-item	15-item	30-item	30-item			15-item	15-item
Semantic fluency	A, V	A	A	A, V	A	A		A	A
Digit Span Test	X	X	X	X	X	X			
Logical Memory I & II, Wechsler Memory Scale	X	X	X	X					X
Trail Making Test	X			X	X				X
Word list learning (CERAD battery)		X	X		X			X	
Symbol-Digit Modalities Test		X	X						

Scaling a latent variable in IRT

- There is no natural scale in latent variable space
- By convention, we usually make mean 0, variance 1
- PROMIS scales their constructs to a T-scale (mean 50, SD 10)

How to link items to an internal scale using Mplus

Data are stacked or pooled together

```
MODEL:

  gcp BY u1* u2 – u17 ;

  gcp@1 ;

  [gcp@0] ;
```

Disadvantage: scale has no external meaning

How to link items to an internal scale using Mplus

Data are stacked or pooled together

STEP 1 MODEL:

In dataset1 (or among people aged 65,75) (etc.)

```
gcp BY u1* u2 – u17 ;
gcp@1 ;
[gcp@0] ;
```

STEP 2 MODEL: In the full data:

```
gcp BY u1@0.2 u2@0.8 [...] u17@1.3 ;
gcp* ;
[gcp*] ;
```

Factor is now scaled to a particular dataset or reference group

How to link items to an external scale using Mplus

Data are stacked or pooled together

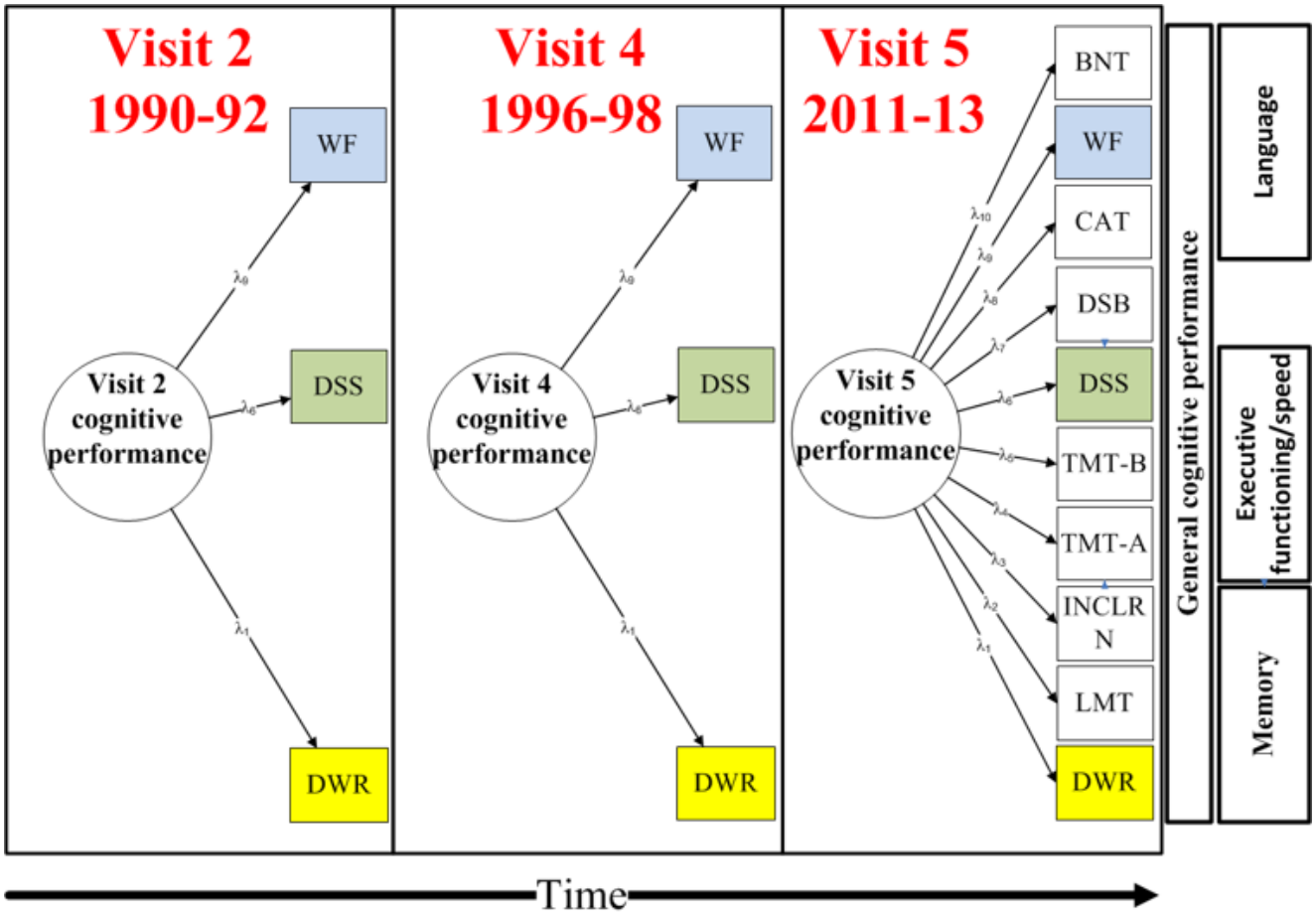
```
MODEL:
 gcp BY u1* u2 – u17 ;
gcp* ;
 [gcp*] ;
gcp BY u14@4.277 ; ! a parameters!!
gcp BY u15@4.196 ;
 [ u14$1@-9.033 u14$2@-3.842
   u15$1@-7.578 u15$2@-2.362] ; ! b parameters
```

Matters to be aware of using IRT for co-calibration

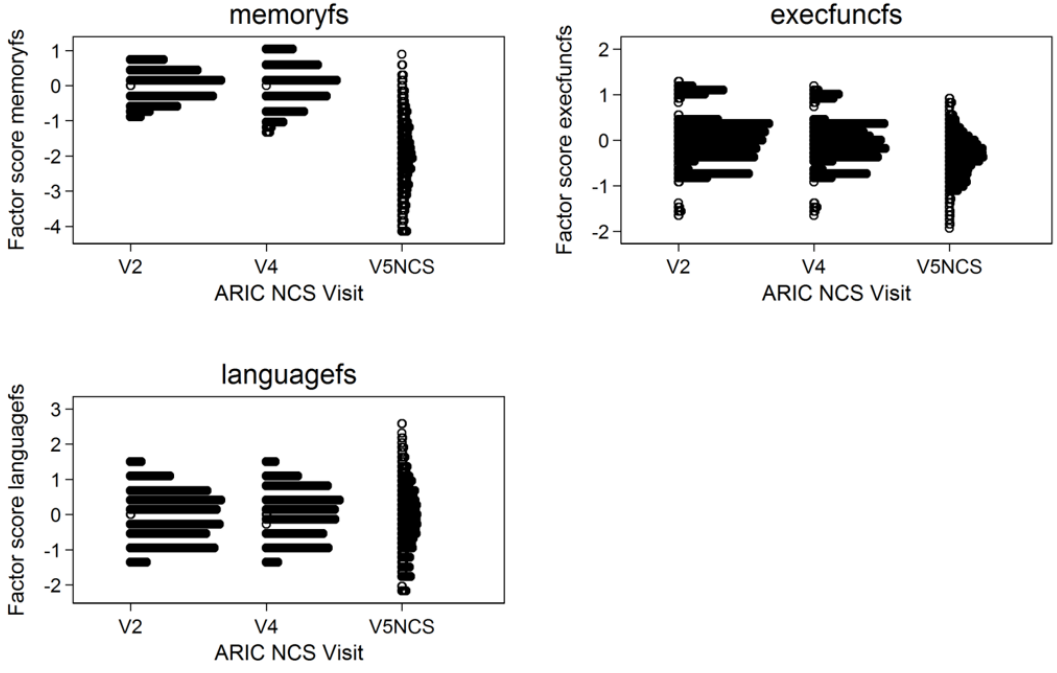
- Is the scale interpretable?
- Check for differential precision / ceiling effects
- Selecting an estimator
- Is the metric the same across studies?
- Is it a problem to estimate an IRT model with repeated measures on people?

Check for differential precision

- In addition to estimating levels for a person at a point in time in a sample, IRT can quantify the precision of the estimated value
- More information, leading to improved precision, isn't always better when it is differential by study visit or data source
- Example: ARIC NCS study



Domain-specific factor scores



- We see differences in floors based on the precision of information available
- We did not think these dropped floors would be differential by a predictor
- In fact they are... Differential precision across visits can bias estimated associations of an exposure with cognitive decline if people with low levels of the exposure have low cognition at baseline

Familiar methods effect

No statistical model can address without more assumptions

Soc Psychiatry Psychiatr Epidemiol (2004) 39: 828–835 DOI 10.1007/s00127-004-0815-8

ORIGINAL PAPER

Richard N. Jones · Stephanie J. Fonda

Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study

Table 2 HRS/CES-D relative response (proportion) among those with complete data at Wave 1, 2 and 3 (N = 5,734)

	Wave 1 (1992)				Wave 2 (1994)		Wave 3 (1996)	
Symptom	None or almost	Some of the time	Most of the time	All or almost all	No	Yes	No	Yes
I felt depressed	0.70*	0.25	0.03	0.02	0.83	0.17	0.85	0.15
Everything an effort	0.64*	0.24	0.08	0.04	0.77	0.23	0.78	0.22
Sleep was restless	0.52*	0.34*	0.08	0.06	0.70	0.30	0.73	0.27
I could not get "going"	0.60*	0.32*	0.05	0.03	0.81	0.19	0.82	0.18
I felt lonely	0.74*	0.20	0.03	0.02	0.86	0.14	0.85	0.15
I enjoyed life	0.03*	0.10	0.34	0.53	0.08	0.92	0.08	0.92
I felt sad	0.65*	0.31*	0.03	0.02	0.84	0.16	0.83	0.17
I was happy	0.04*	0.18*	0.45	0.33	0.12	0.88	0.12	0.88
I had a lot of energy	0.10*	0.28*	0.40	0.23	–	–	0.35	0.65

– denotes item not collected; * indicates response categories collapsed into null symptom group for linking analyses (the remaining levels collapsed into symptom present group). Wave-wise rows may not add up to 1.00 due to rounding

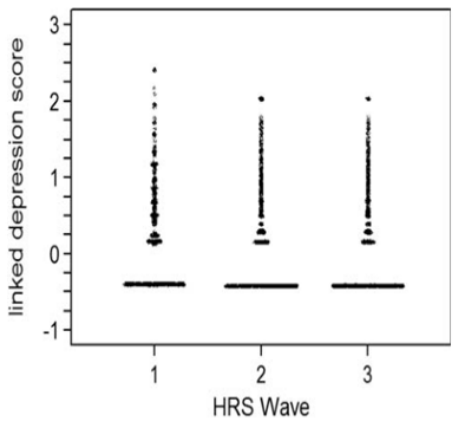


Fig. 2 Dotplot illustrating relative frequency distribution of linked Health and Retirement Study CES-D latent trait scores, using dichotomized Wave 1 responses, for respondents with complete depression information at Waves 1, 2 and 3 (N = 5,734)

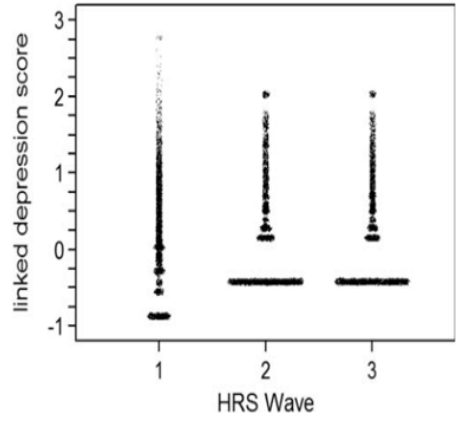


Fig. 3 Dotplot illustrating relative frequency distribution of Health and Retirement Study CES-D latent trait scores, using the four-category response options for Wave 1 respondents, among those with complete depression information at Waves 1, 2 and 3 (N = 5,734)

Matters to be aware of using IRT for co-calibration

- Is the scale interpretable?
- Check for differential precision / ceiling effects
- Selecting an estimator
- Is the metric the same across studies?
- Is it a problem to estimate an IRT model with repeated measures on people?

Selecting an estimator

- Maximum likelihood (with robust estimation) (MLR)
  - All records are used
    - ⋆ Except records with 100
  - Assumes MAR (missing at random)
  - More reasonable in epidemiologic settings
- WLSMV is accurate, efficient (fast), but inappropriate when MCAR (missing completely at random) assumptions are not viable
  - Harmonization across many datasets

Selecting an estimator

- MLR and WLSMV provide "regression-based" factor score estimates
- Bayesian plausible values
  - Based on a mean of k individual plausible values drawn from the posterior distribution
  - As number of draws from the posterior increase, we should reach MLR regression-based factor score estimates
- Regression-based factor scores are fair for high-stakes testing because a record (person) with the same response pattern gets the same values
- If the goal is to estimate population parameters (e.g., epidemiological inference), then plausible values might be desired because they retain imprecision in estimates
- See Asparouhov and Muthen (2010). Plausible values for Latent Variables Using Mplus.

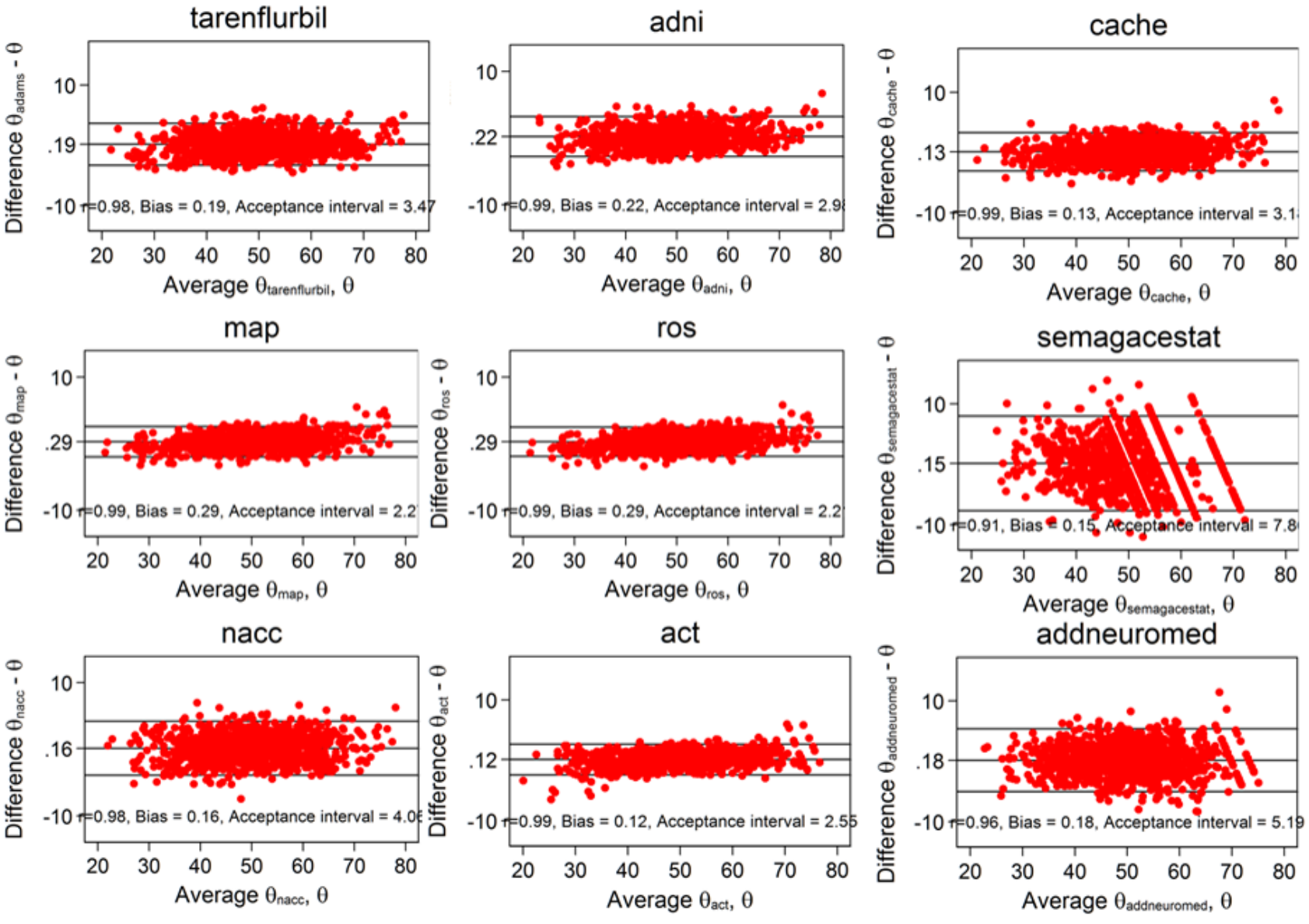
Matters to be aware of using IRT for co-calibration

- Is the scale interpretable?
- Check for differential precision / ceiling effects
- Selecting an estimator
- Is the metric the same across studies?
- Is it a problem to estimate an IRT model with repeated measures on people?

Is the metric the same across studies?

- Differential item functioning (DIF) by dataset
- Simulation
  - Imagine a population in which all indicators were administered to all respondents
  - Derive a "true" theta and a dataset-specific theta based on tests available only in that data
  - Compare

Is the metric the same across studies?



Matters to be aware of using IRT for co-calibration

- Is the scale interpretable?
- Check for differential precision / ceiling effects
- Selecting an estimator
- Is the metric the same across studies?
- Is it a problem to estimate an IRT model with repeated measures on people?

A perennial reviewer critique...

- A possible problem involves the non-independence of the measures. Although robust estimators were used, they can go only so far in eliminating the high degree of non-independence
- Resolution
  - We can run sensitivity analyses by estimating models based on 1 record per person. Do this 10240 times, get bootstrapped estimates
  - This yields the same model parameters (loadings and thresholds) as a model using all records
  - The standard errors of item parameters do become larger when we use just 1 record per person, however we do not typically use those standard errors

Tools for co-calibration

- Distribution-based
  - Goal: define a transformation of a test that returns the same cumulative probability plot as the other variable being compared
  - Mean equating
  - Linear equating
  - Equipercentile equating
- Item-based
  - Goal: define a transformation of a test that places it on the same metric as another
  - Item response theory
- Multiple imputation for missing data

Missing data/multiple imputation

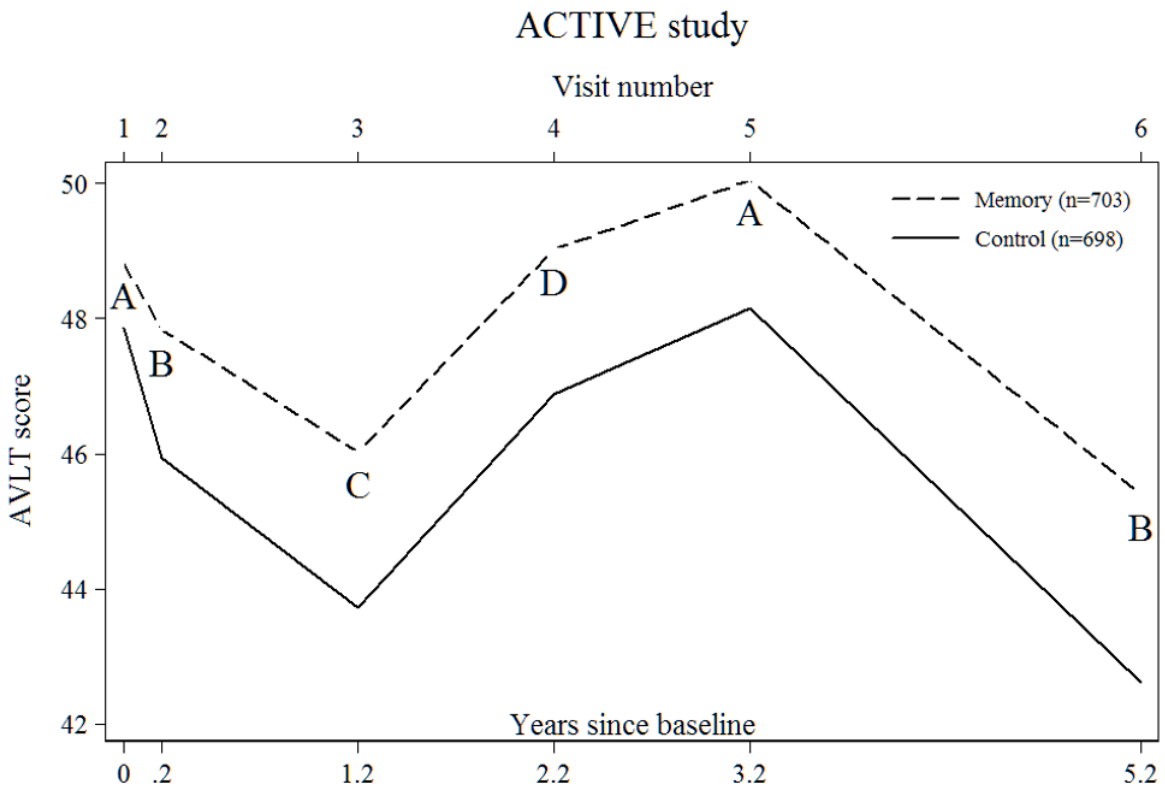
- Assume you have 2 datasets
  - Dataset 1 has hba1c, diabetes diagnosis, glucose
  - Dataset 2 has diabetes diagnosis, glucose
- If I want to use HBA1c in the pooled sample, I could predict it based on diabetes status and fasting glucose using multiple imputation methods in the pooled data
- Multiple imputations reflect uncertainty in unmeasured data
- If both datasets measure the same construct, we can measure it in the pooled data

Stop here

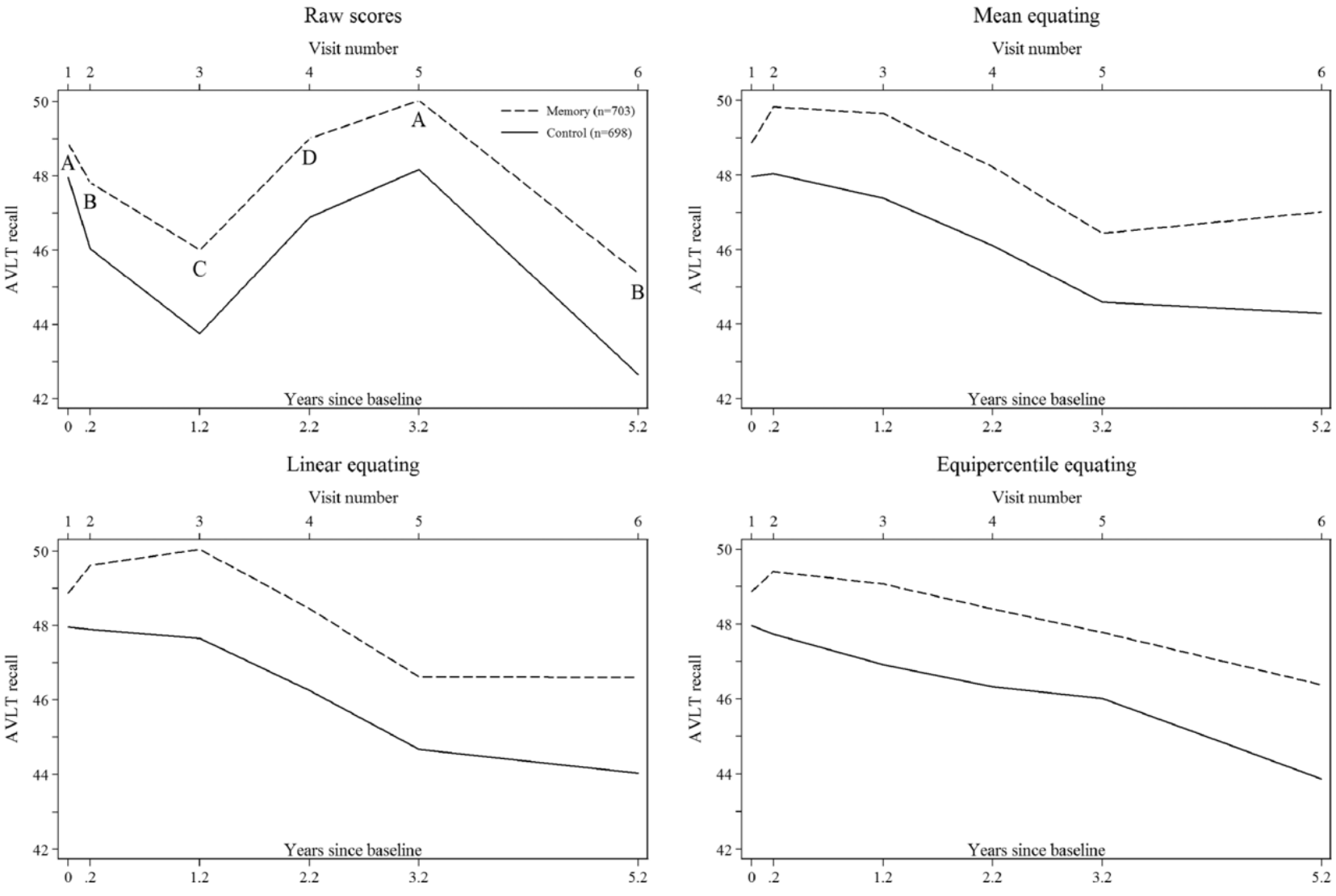
Example challenge - AVLT nonequivalence

Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time

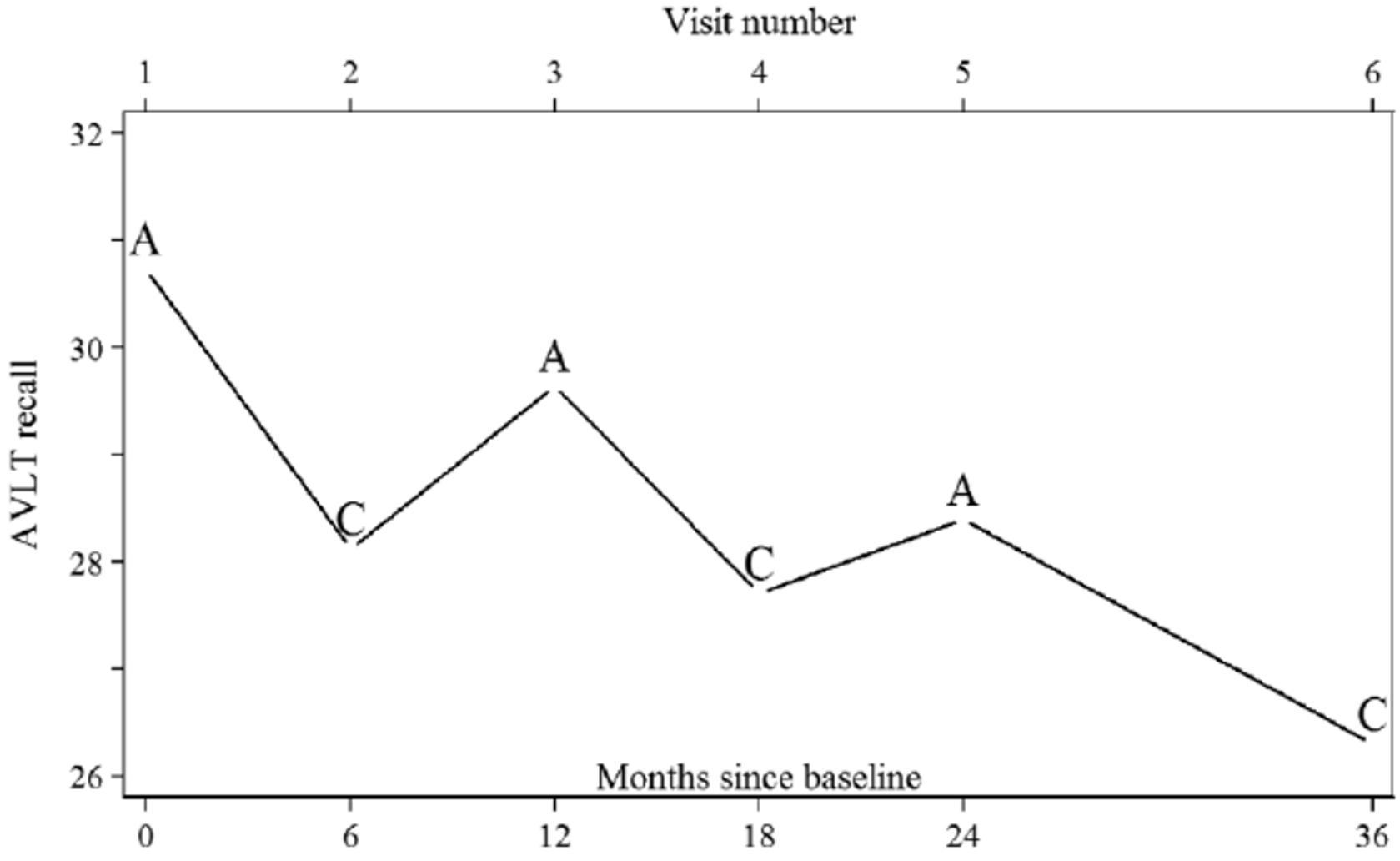
Alden L. Gross[2,3], Sharon K. Inouye[2,3], George W. Rebok[1,5], Jason Brandt[1,5], Paul K. Crane[6], Jeanine M. Parisi[1], Doug Tommet[2], Karen Bandeen-Roche[4], Michelle C. Carlson[1], Richard N. Jones[2,3], for the Alzheimer's Disease Neuroimaging Initiative*



Statistical equating methods: ACTIVE AVLT



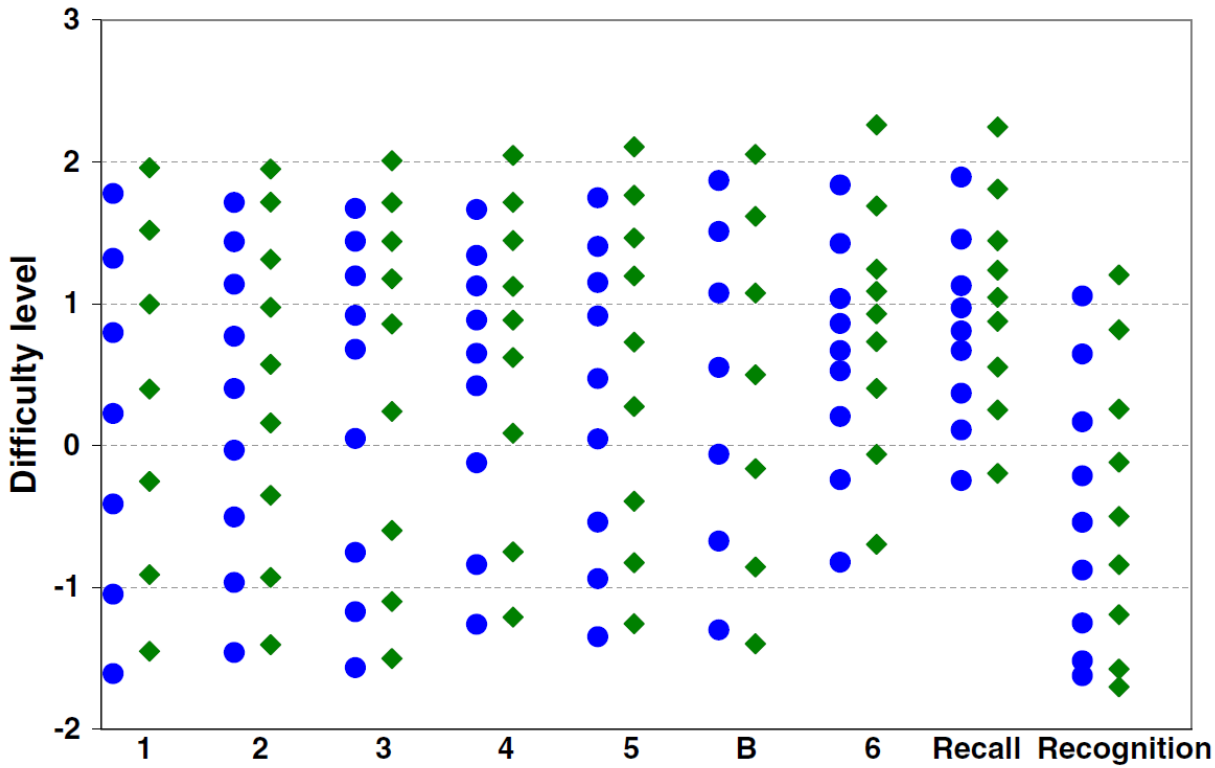
ADNI 1 data (N=825)



An IRT approach in ADNI

Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Paul K. Crane · Adam Carle · Laura E. Gibbons · Philip Insel · R. Scott Mackin · Alden Gross · Richard N. Jones · Shubhabrata Mukherjee · S. McKay Curtis · Danielle Harvey · Michael Weiner · Dan Mungas · for the Alzheimer's Disease Neuroimaging Initiative



Maybe stop here

PSYCHOMETRIKA—VOL. 66, NO. 4, 473–486
DECEMBER 2001

PSYCHOMETRIC ENGINEERING AS ART

DAVID THISSEN

L.L. THURSTONE PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

The Psychometric Society is "devoted to the development of Psychology as a quantitative rational science". Engineering is often set in contradistinction with science; art is sometimes considered different from science. Why, then, juxtapose the words in the title: *psychometric*, *engineering*, and *art*? Because an important aspect of quantitative psychology is problem-solving, and engineering solves problems. And an essential aspect of a good solution is beauty—hence, art. In overview and with examples, this presentation describes activities that are quantitative psychology as engineering and art—that is, as design. Extended illustrations involve systems for scoring tests in realistic contexts. Allusions are made to other examples that extend the conception of quantitative psychology as engineering and art across a wider range of psychometric activities.

Key words: psychometrics, quantitative psychology, design.

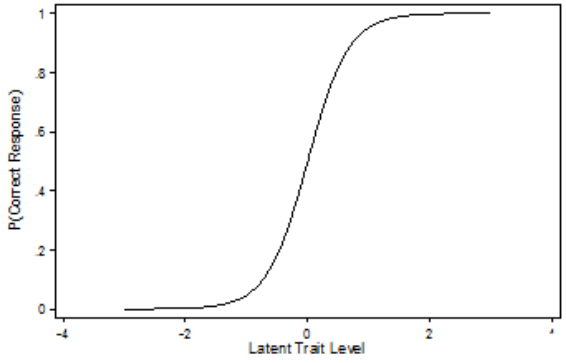
Psychometric engineering as art

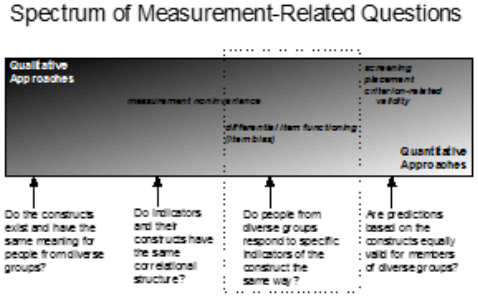
- Psychometrics is a field of study concerned with the theory and technique of psychological measurement
- `https://en.wikipedia.org/wiki/Psychometrics`
- What do psychometricians design?
  - Models
  - Algorithms
  - Statistical procedures

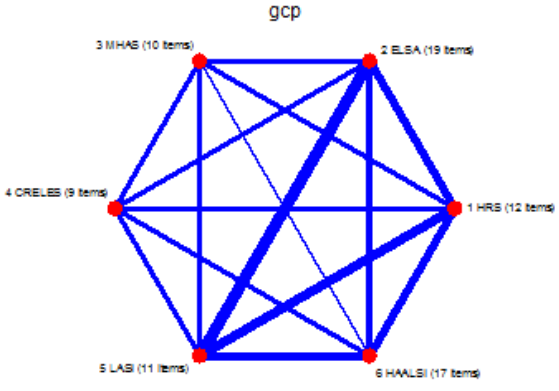
Psychometric engineering as art

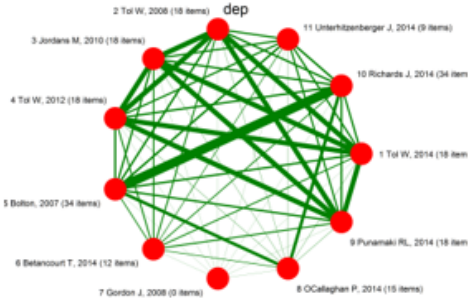
- Engineering is the application of mathematics, empirical evidence and scientific, economic, social, and practical knowledge in order to invent, innovate, design, build, maintain, and improve structures, machines, tools, systems, components, materials, processes and organizations
- Art is about finding beauty in good designs
  - George Box: "Models, of course, are never true, but fortunately it is only necessary that they be useful..." (1979, pg 2)
  - Pablo Picasso: "Art is a lie that enables us to realize the truth"
- Design should integrate art and engineering
- A psychometrician's models may be wrong, but if they are useful (engineering standard) and tell us something about the universe (science), and beautiful (art) then they are valuable

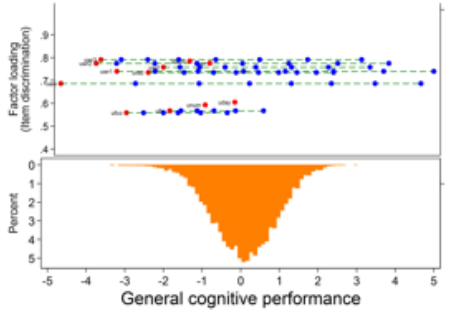
We want to do analyses











- An important aspect of psychometrics involves problem-solving. Engineers solve problems
- The work of a psychometrician is to create a thing of beauty that is useful

Psychometric Engineering (Thissen 2001)

- Goal: make functional, useful, beautiful things
- Harmonization serves as a bridge between 2+ studies. The tests must be useful and functional
- Analogously, a great Battle Station was built in Star Wars to bring harmony to the galaxy, to restore order