Comments on

Developing Tailored Instruments: Item Banking and Computerized Adaptive Assessment

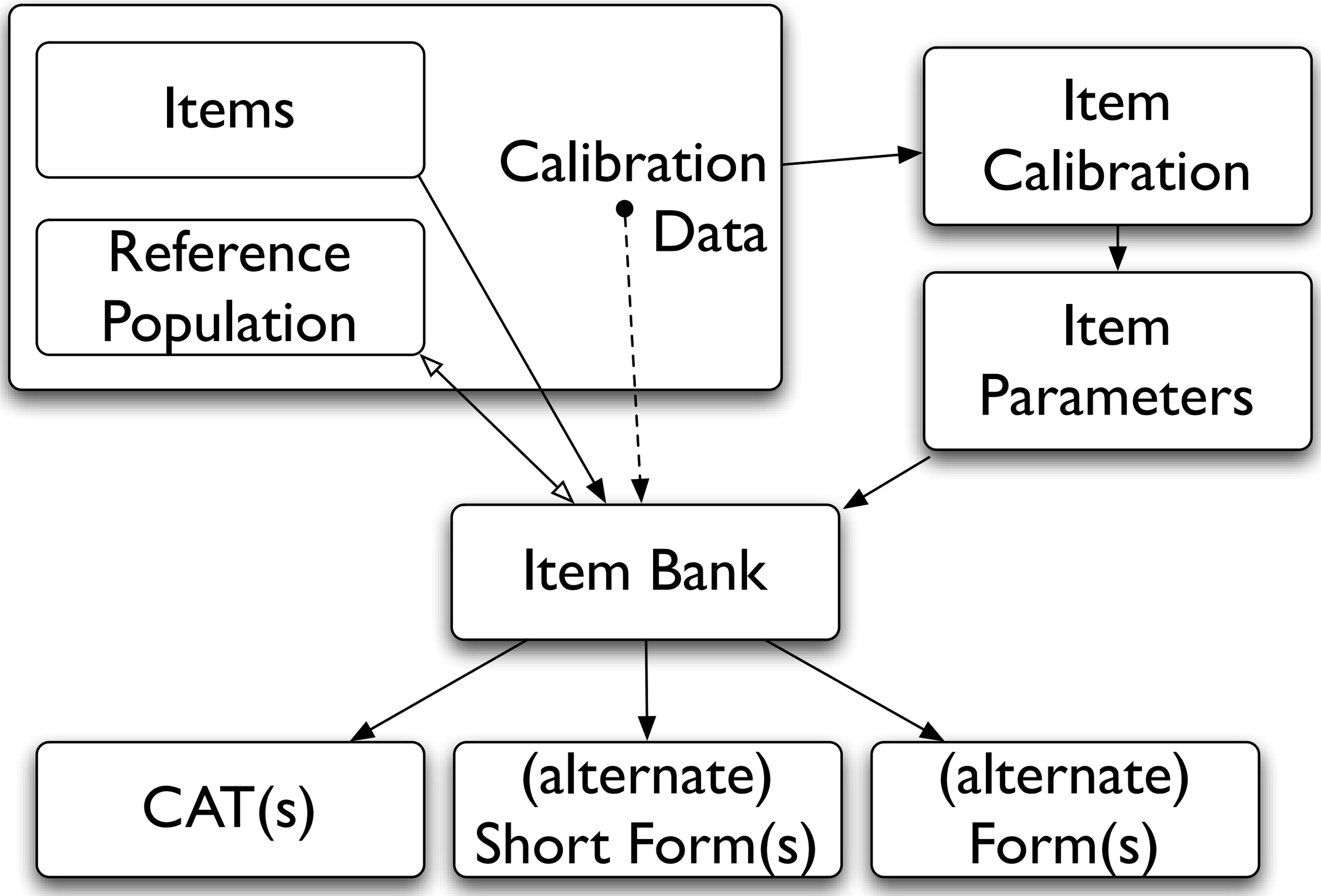
David Thissen

L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Prepared for the conference
"Advances in Health Outcomes Measurement,"
Bethesda, Maryland, June 23-25

Item Response Theory (IRT)

- Uses "trace lines" (or *item characteristic curves* or *item response functions* or …) for **item analysis** and **test scoring**
- In parametric IRT, the trace lines are based on **item parameters** that are estimated using **calibration data**
- Item **information curves** are also useful
- IRT **scale scores** may be computed for responses to fixed tests or **computerized adaptive tests** (CATs)


Items
Reference Population
Calibration Data
Item Calibration
Item Parameters
Item Bank
CAT(s)
(alternate) Short Form(s)
(alternate) Form(s)

Remarks on a Potpourri of Topics

- The Reference Population
- Calibration—Dimensionality
- Dichotomous and Polytomous Response Scales
- Differential Item Functioning (DIF)
    - Conditional Scoring
- Mode Effects
- Local Dependence and Context Effects
- New Ideas

The Reference Population

"The availability of trustworthy item parameter estimates is essential for any measurement application of item banking… Strictly speaking, it is not correct to say that the latent trait models provide invariant item parameter estimates. Only if a common scale… is used from group to group will this be true."

–Robert Wood (1976)

In most applications of IRT this common scale is defined by the mean and standard deviation of some reference population.

The Reference Population

- The general population?
- A well-defined population with a specific disease?

All item parameters linked to the scale of the reference population (whether or not the items were administered there)

Consider the reporting scale?

- *z* scores?
- *T* scores?
- Others?

Calibration—Dimensionality

"The availability of trustworthy item parameter estimates is essential for any measurement application of item banking… "

–Robert Wood (1976)

"Trustworthy" I:

- Sample sizes 500-1000 or more (perhaps 250)
- "Appropriate" respondents:
    - From reference population or linked to it
    - Use all response categories

Calibration—Dimensionality

"Trustworthy" II:

- Unidimensional IRT
- For multidimensional constructs:
    - Data analytic judgment—gray continuum
    - For now: Break into unidimensional scales
    - Summary scores may be combinations

- Except in special cases,

practical MIRT remains "under development"

Dichotomous and Polytomous Response Scales

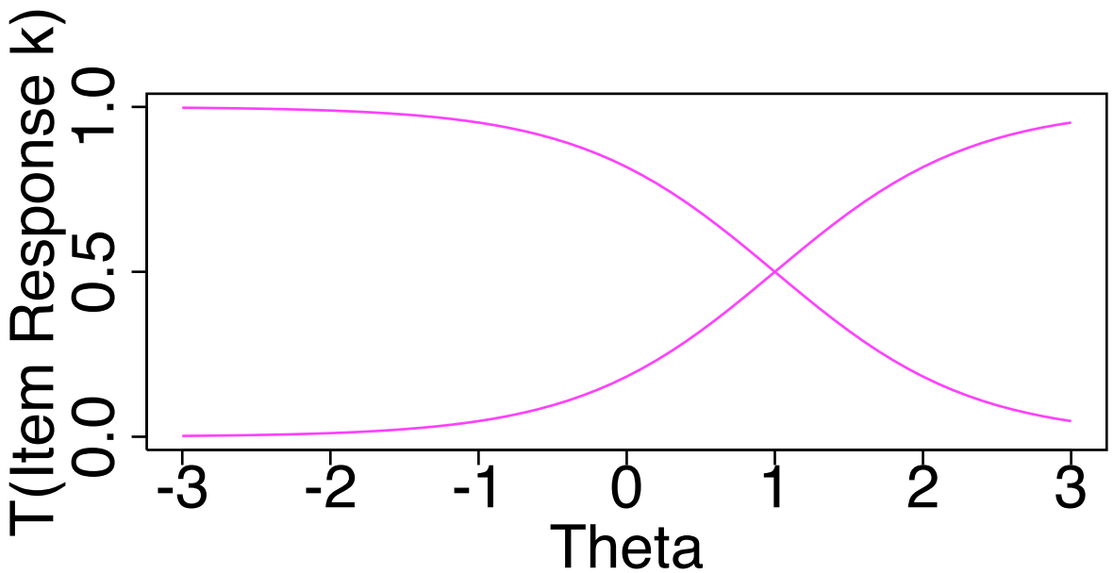
- IRT deals with either, or mixtures

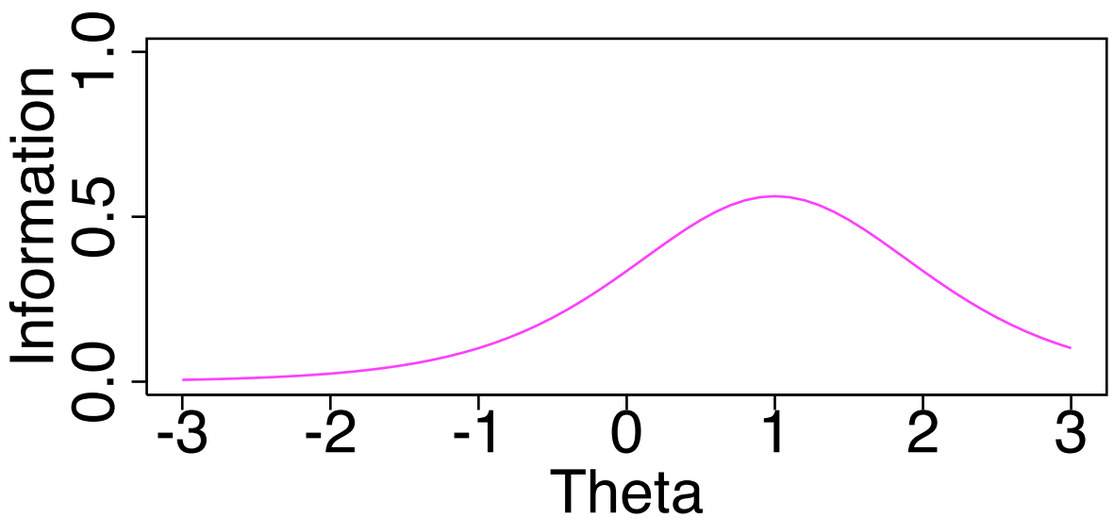
- Adaptive testing most motivated with dichotomous items:
    - Each item provides information over a narrow range
    - Adaptation shortens tests by avoiding useless items

- Polytomous response items "self adapt":
    - Analogous to chained dichotomous items
    - Each item has a relatively flat information curve
    - Short scales may yield precise measurement

Dichotomous and Polytomous Response Scales

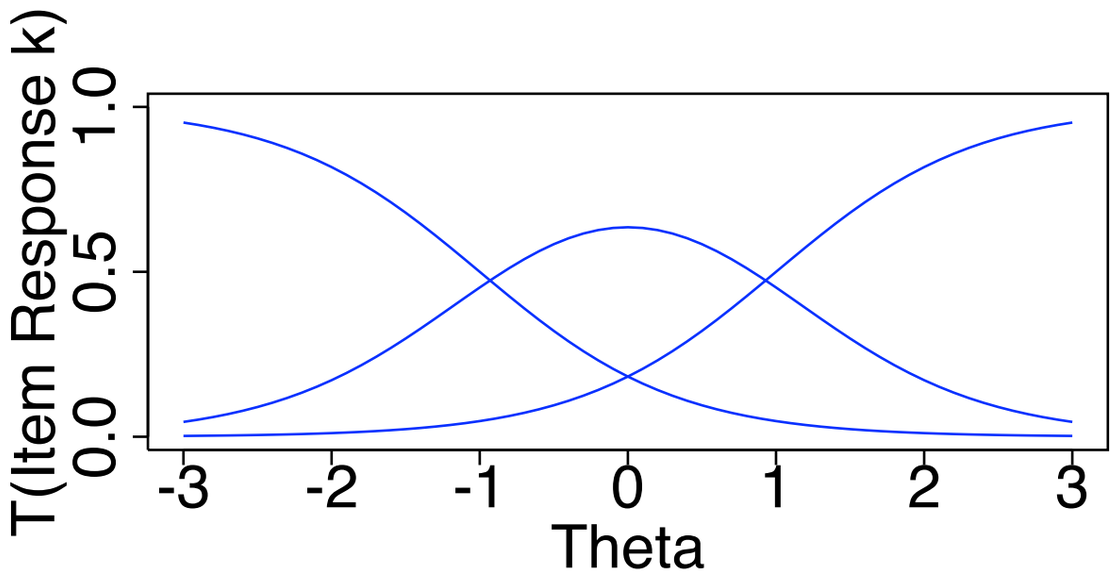
Dichotomous Response

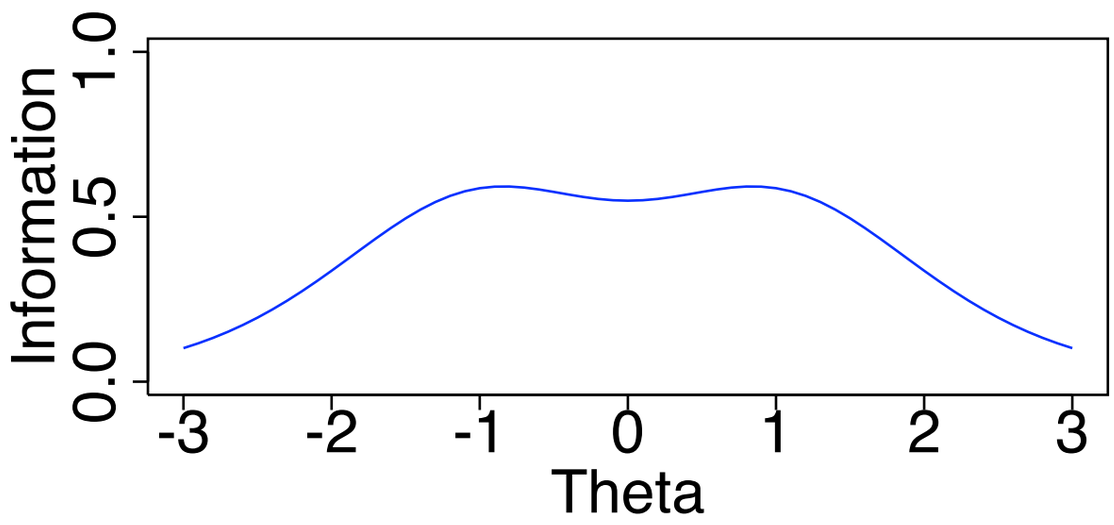




More Peaked Information

Three Response Categories





Relatively Flatter Information

Dichotomous and Polytomous Response Scales

Circumstances under which adaptation may be useful even with polytomous item responses:

- Measurement when few items exist
  but natural response is graded

- Measurement of a construct with a very wide range:
  "I feel irritable because of my headaches" vs.
  "I feel desperate because of my headaches"

  –Bjorner, Kosinksi & Ware (2003)

- Measurement from an item bank for a broad age-range:

More on
Measurement from an item bank for a broad age-range:

- Might dichotomous responses be easier
  for younger or older respondents?

- Perhaps the same "item" can be calibrated
  with dichtomous response choices, and
  with polytomous responses

- Those would be two distinct items
- Note: Any change may change everything

Dichotomous and Polytomous Response Scales

Models and Checking:

- Use IRT procedures for response that are *not* ordered?:
    - Ramsay's TestGraf
    - Bock's Nominal Model

  (Both these may require more data)

- Commonly used models for Likert-type responses:
    - Samejima's Graded Model
    - Masters' Partial Credit

  / Andrich's Rating Scale Models
    - Muraki's Generalized Partial Credit Model

  (aka Yen's Two-Parameter Partial Credit Model)

Differential Item Functioning (DIF)

Differential item functioning (DIF) means and item has different trace lines for different groups:

- Demographic groups
- Diagnostic groups? ("normal" vs. ?)
- Distinct reference populations?

DIF may reflect a lack of validity:

- "Something different" measured between groups
- Omit such items (usually, in education)

Conditional Scoring

For items that exhibit DIF,

if "fairness" is not an issue:

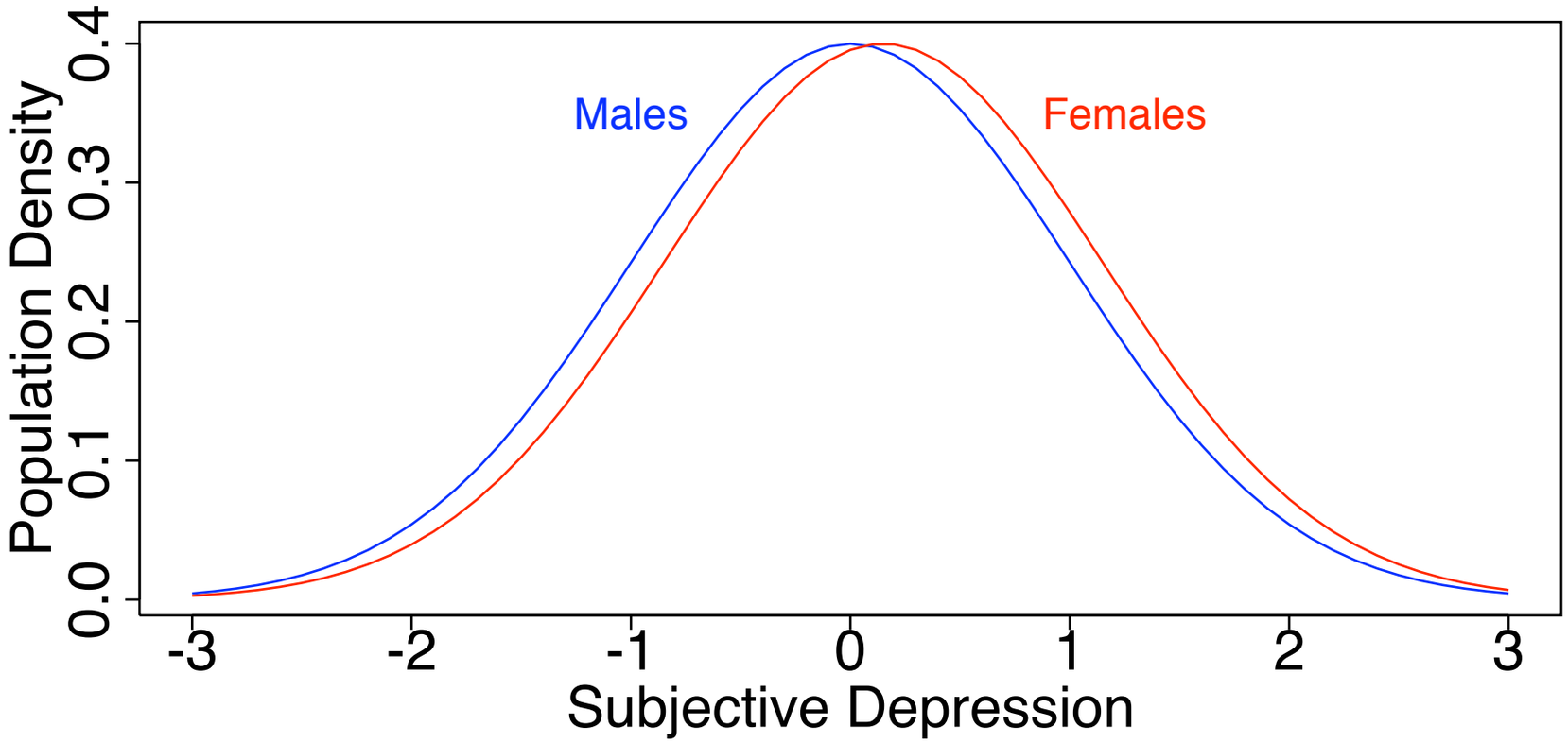
It may be possible to treat the item as two items—

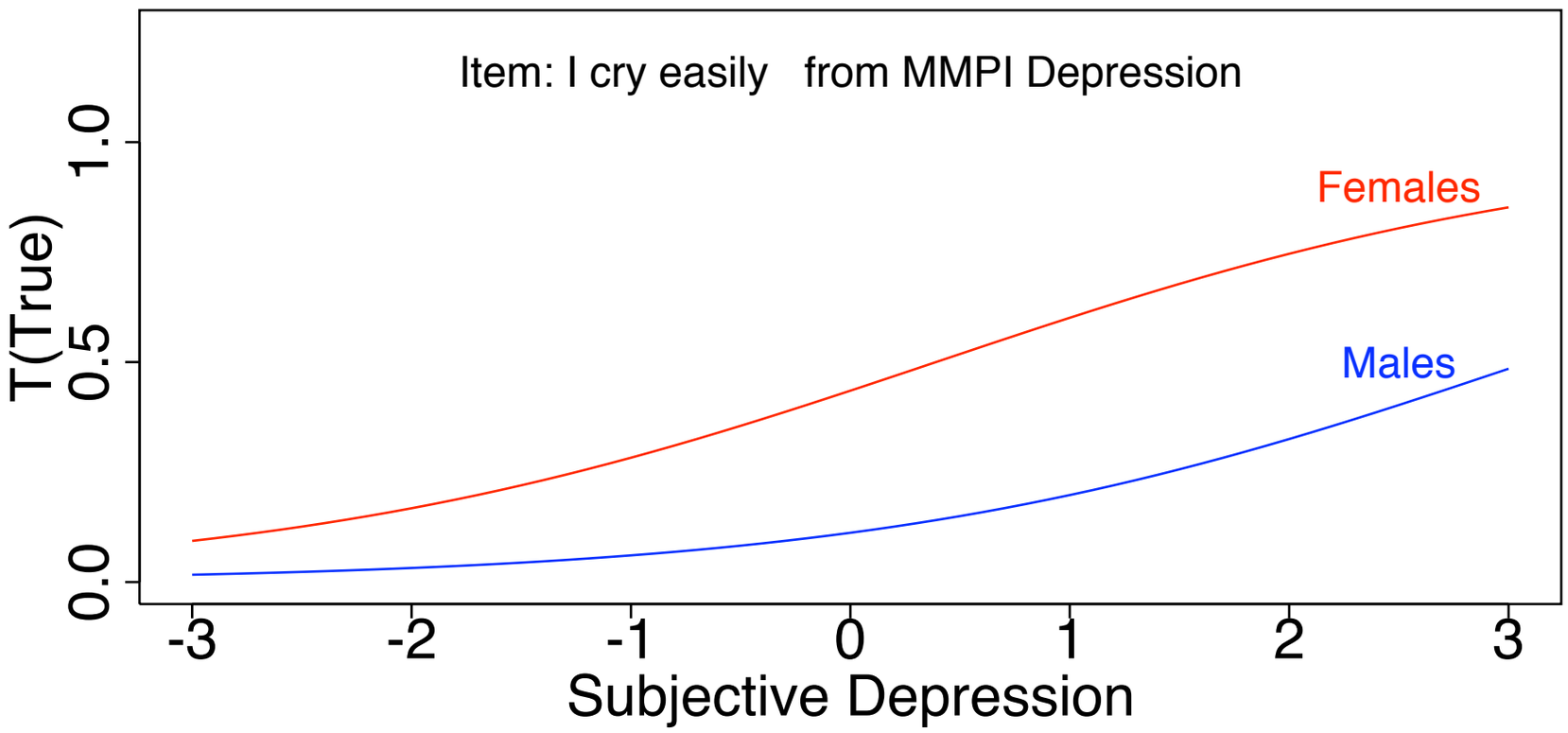
that is, simply "different items" in the two groups

For example:

"crying" items on depression scales

(different for men and women)





Example from Reeve (2000): Reference population is the MMPI-2 restandardization non-clinical sample; scale is Harris-Lingoes Subjective Depression

Mode Effects

The medium of presentation
[paper-and-pencil, computer screen,
(variants within both of those)]
may well affect item responses and parameters

Items are best calibrated as they will be presented

If results from different types of presentation are to be compared, consider multi-trait, multi-method analysis to confirm the comparability of scores?

Local Dependence and Context Effects

CAT assumes items are fungible

> **fungible** \FUN-juh-bul\, adjective :
> **1.** (Law) Freely exchangeable for or replaceable by another of like nature or kind in the satisfaction of an obligation.
> **2.** Interchangeable.

Order- or other context-effects may exist

Items may exhibit local dependence (LD) —

excess covariation between item responses, over and above that accounted for by variation in the construct being measured

Local Dependence and Context Effects

Multi-stage or "testlet" CAT designs
may be used to control context effects

Items remain in fixed positions
relative to a fixed context

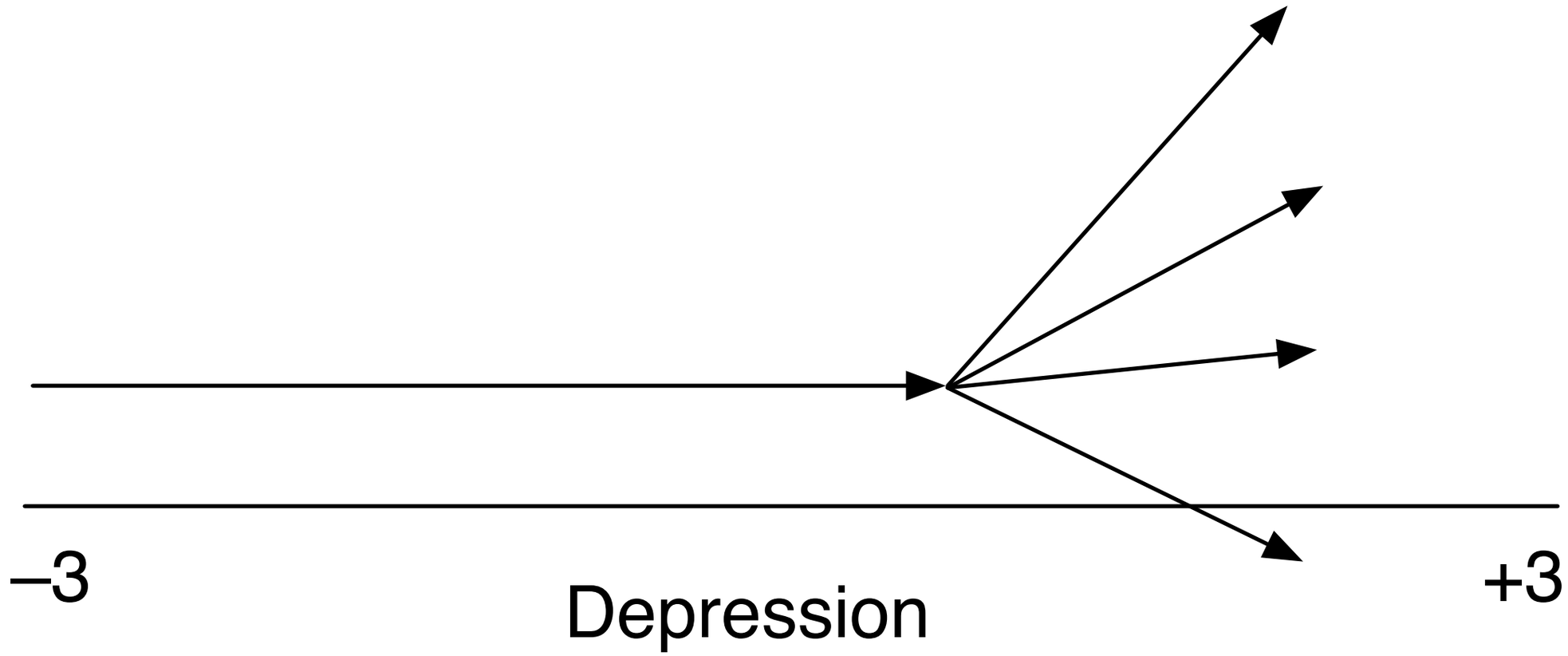
Such designs are for relatively longer CATs

New Ideas

in CATs for the Measurement of Health Outcomes

- Doubly-adaptive CATs (CAATs?):
  - "skip-patterns"
  - and
  - IRT-based item selection
- Variable-criterion variable-length stopping rules

The Weiss Possibility:



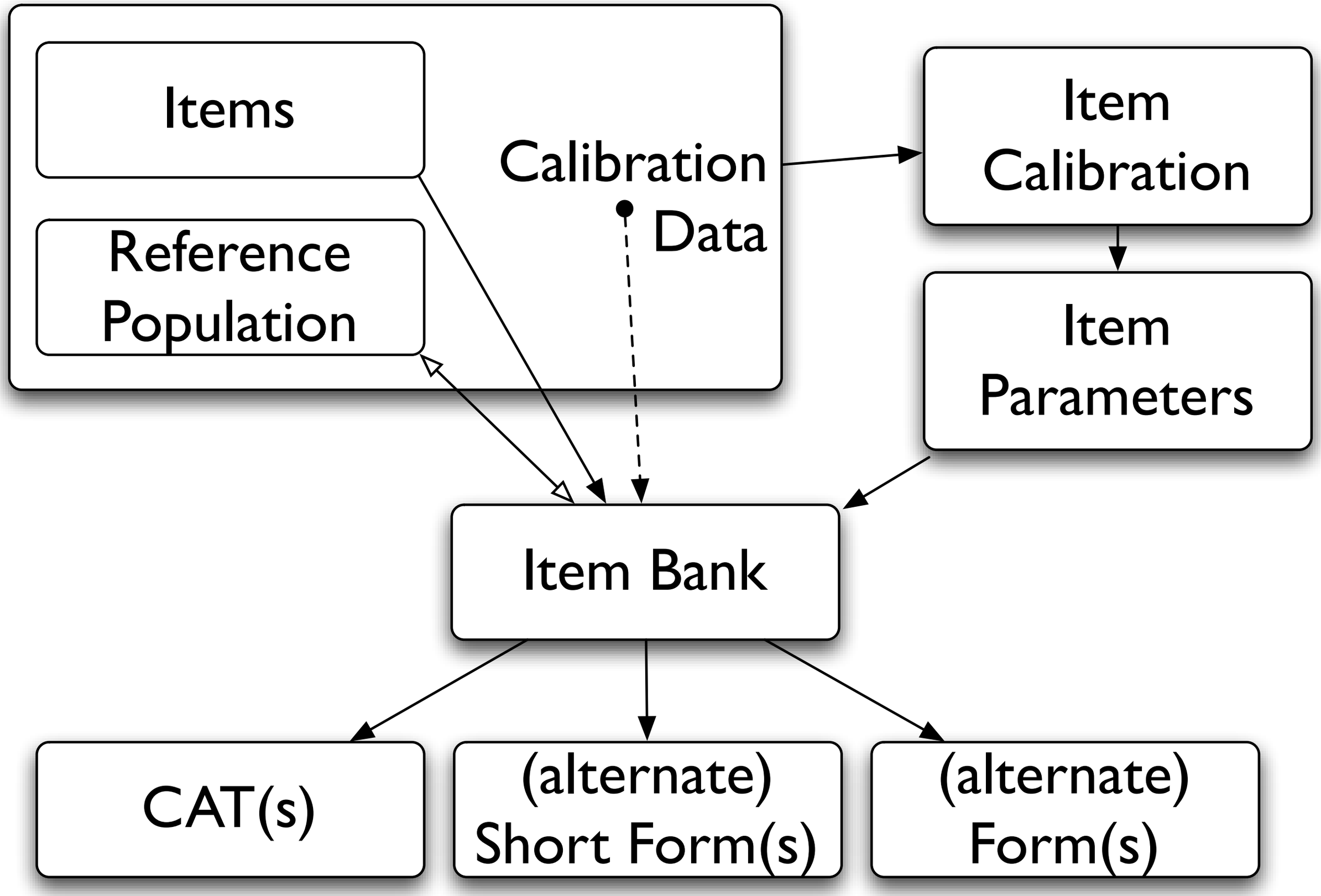
(A specific kind of diagnostic-category DIF)

New (Old) Challenges

Item bank maintenance, and item parameter drift:

- Items will need to be added to banks
- Items' parameters may "drift"

For long-term, widespread use of banked items, it will be best if the calibration data themselves are "deposited in the bank" as well, to facilitate linkage with data collected subsequently with the old items, or with new items—


Items
Reference Population
Calibration Data
Item Calibration
Item Parameters
Item Bank
CAT(s)
(alternate) Short Form(s)
(alternate) Form(s)