Migrating from UDS 2.0 to UDS 3.0: Modern psychometric methods

Paul K. Crane, MD MPH
Department of Medicine
University of Washington

ADC Directors Meeting, San Diego, CA, September 2011

General introduction

- “Modern psychometrics” refers to advances in educational psychology made since the 1960s
  - Allan Birnbaum’s section of Lord and Novick’s 1968 textbook, *Statistical Theories of Modern Test Scores*
- No one with graduate training in test theory since 1968 would use total scores unless they had done a lot of work up front

Physician with an outsider's perspective

- It is weird that one branch of psychology (neuropsychology) is relatively uninformed by another branch of psychology (educational psychology / quantitative psychology)
- Not unique to neuropsychology
  - Borsboom D, "Attack of the Psychometricians", *Psychometrika* 2006

Goals of migration

- Move from an old neighborhood to a better one
- What does “better” mean?
  - Public domain
  - Richer assessment
  - Better measurement properties
- But our old neighborhood had a lot of desirable properties too
  - Lots of data
  - Would really like to move in a way such that our old data are still useful

Challenges of migration

- This is (very) different from the “parallel and equivalent forms” problem because we are moving to a *better* neighborhood
  - Even if we can equate scores, they may be qualitatively different from the old vs. the new neighborhood
  - This leads to important challenges in statistical methodology when combining scores from the old and the new neighborhoods
- Domain coverage is different

Recommendation

- Data collection exercise in which the old and the new batteries are administered to a common group of informative people
- Don’t need all domains (see below)
- Granular data coding (“item level”)
  - Can always obtain summaries
  - Challenge to go back through and obtain item-level data
  - It’s 2011

Easiest (not a) problem: either no change or dropping

- Processing speed
  - Old: Trails A, Digit Symbol
  - New: Trails A
- EF: Inhibition/shifting
  - Old: Trails B
  - New: Trails B
- Language: Semantic memory
  - Old: Animals, vegetables
  - New: Animals, vegetables
- New domains also not a problem; nothing to link them to
  - Benson, FNAME

Domains with changes

Domain	Constructs	Old	New
Attention/ working memory	Immediate span holding	Digit span F	New digit span F
	Manipulation	Digit span B	New digit span B
Episodic memory	Acquisition	Logical Memory Immed	New story Immed
	Retention	Logical Memory Delay	New story Delay
	Orientation	Orientation: 10 items	6 of the 10 words
Language	Object naming	Boston Naming	Multilingual naming
General	Severity of dysfunction	MMSE	MoCA

Digit spans, logical memory

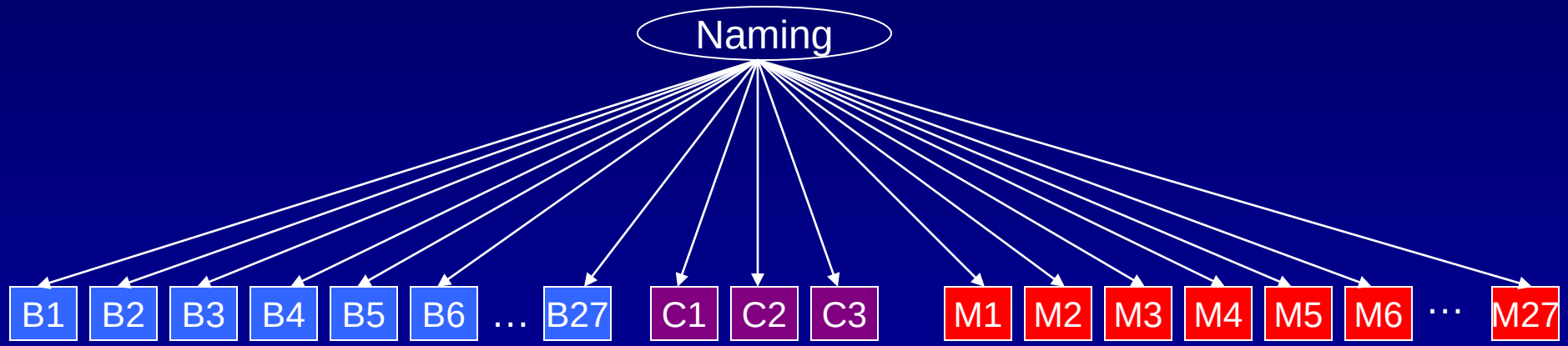
- Administer both sets of digit spans (old and new)
  - Likewise with story recall
- Crosswalk of most likely scores on the other test given the current test

Orientation

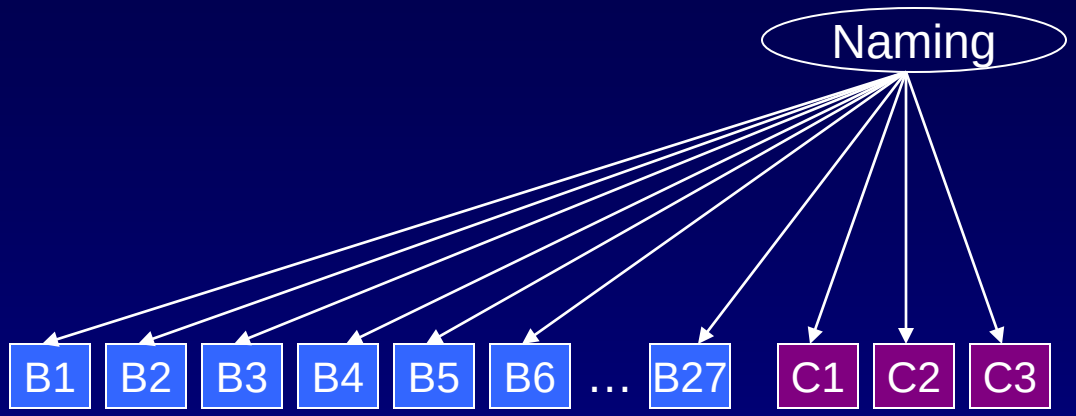
- Subset of 6 of 10 words to be administered
  - Query: why not administer the other 4 as well? Seems like minimal burden, facilitates direct comparability
  - Note that the Blessed has the same orientation items as the MMSE (and Blessed came first) and it is in the public domain
  - 6 orientation items from MoCA, 4 from Blessed that happen to be the same as 4 from MMSE = 10 orientation items
- Sites may (or may not) be able to extract data just for the 6 words
  - I don’t think these are data available at NACC
  - Difficult to extrapolate from 6 to 10

Boston Naming and Multilingual Naming

- Administer both to a large sample
- Missing data formulation (item response theory)

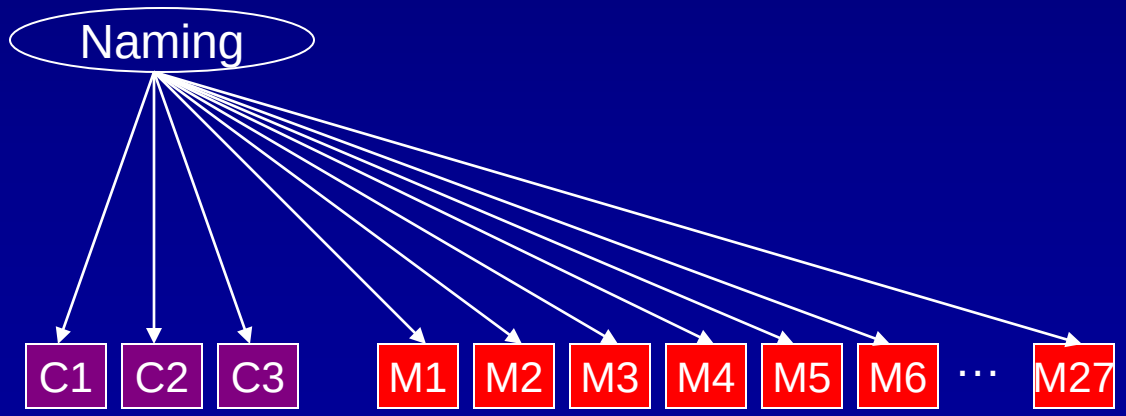


Pictorially



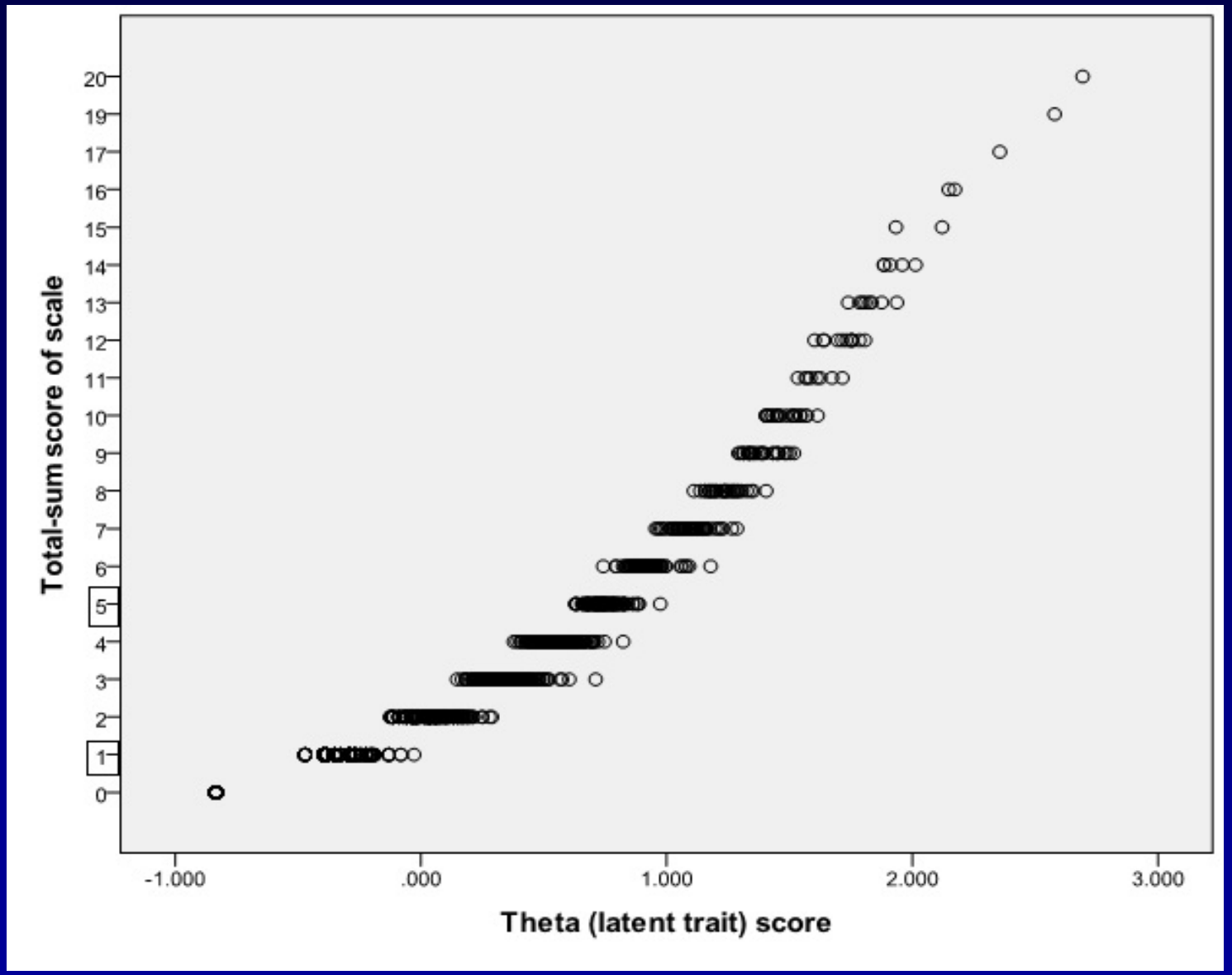
“Boston Naming Test”

“Multilingual Naming Test”



What do you get from co-calibrating?

- Does not assume equal difficulty
- Can get a cross walk
- Even better would be to use scores generated from IRT – and their standard errors!
  - The test characteristic curve shows the most likely scores, but this is not a 1:1 relationship
- Pittsburgh data from Beth Snitz / Mary Ganguli: residual relationship (MS under review)


Total-sum score of scale
20
19
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
-1.000
.000
1.000
2.000
3.000
Theta (latent trait) score

	SCC total score = 1 (*n = 265*)		**SCC total score = 5** (*n = 119*)	
	Standardized β	p-value	Standardized β	p-value
Neuropsychological test				
Global cognition				
MMSE	-0.11	**.08**	-0.08	.36
Memory				
WMS-R Logical Memory II	-0.07	.27	-0.01	.93
WMS-R Visual Reproduction II	-0.13	**.04**	-0.09	.36
FULD-OME	0.01	.84	-0.07	.44
Executive functions				
Trail Making B (s.) *	0.03	.66	0.18	**.07**
Clock drawing	0.06	.38	-0.19	**.05**
Language				
Animal fluency	-0.06	.35	-0.07	.48
Letter fluency	-0.06	.38	-0.17	**.07**
Visuospatial construction				
WAIS-III Block Design	-0.01	.87	0.06	.55
Psychomotor speed				
Trail Making A (s.) *	0.12	**.05**	0.16	**.09**

Abbreviations: IRT = Item Response Theory; SCC = Subjective Cognitive Complaints; MMSE = Mini Mental State Examination; WMS-R = Wechsler Memory Scale – Revised; FULD-OME = Fuld Object Memory Evaluation; WAIS-III = Wechsler Adult Intelligence Scale 3rd Edition.

* Higher values indicate worse performance.

Implications of this

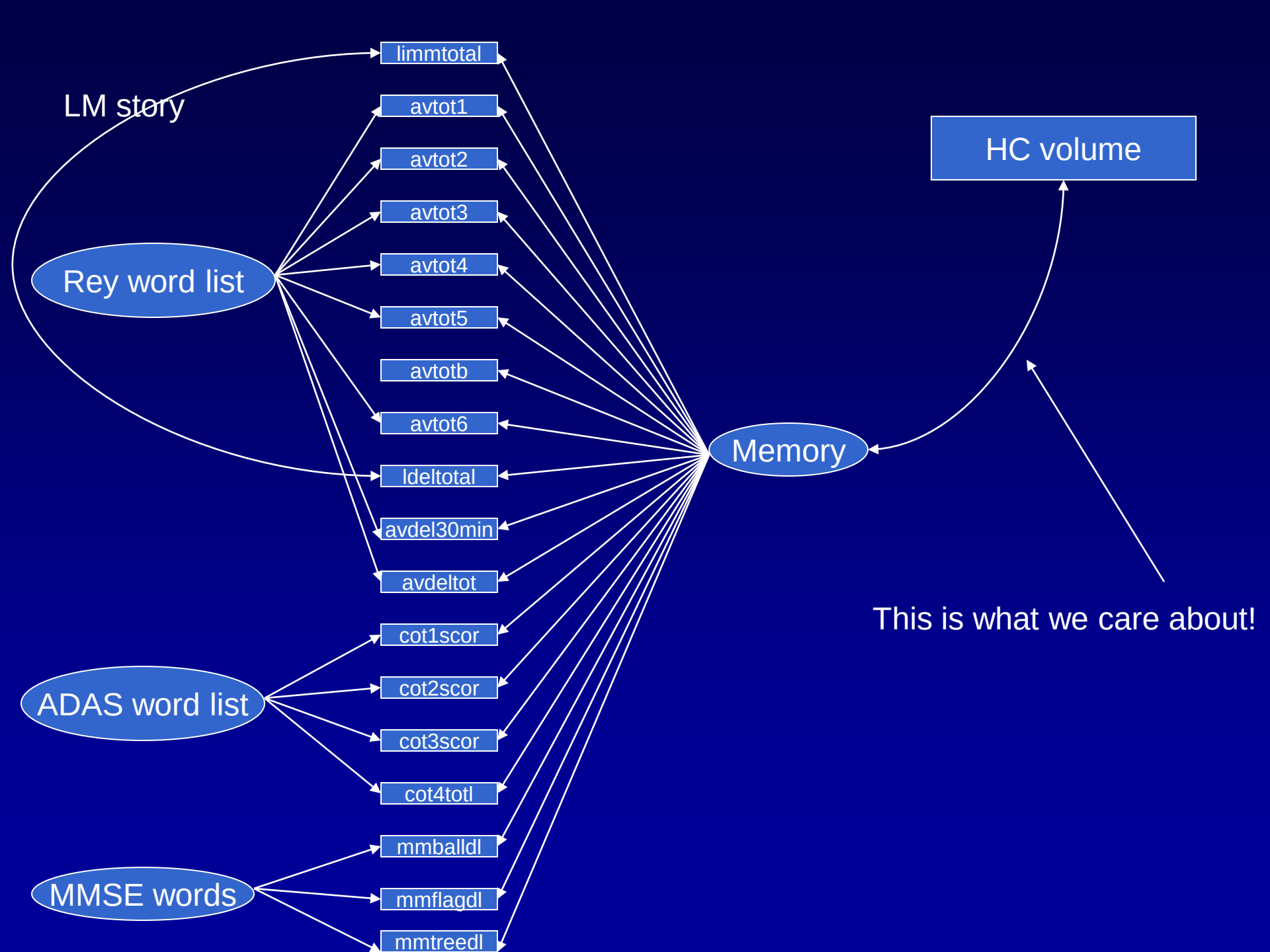
- If standard scores are “correct” there would be no systematic relationship between IRT scores and other tests among people with the same standard score
- 20 comparisons (10 each at 1 and 5 standard score points)
  - Roughly 1 should have $p<0.05$, 2 with $p<0.10$
  - Observed: 3 with <0.05, 7 with <0.10
- Should not have a systematic direction
  - Yet ALL were in predicted direction, with better IRT scores associated with better performance on other tests

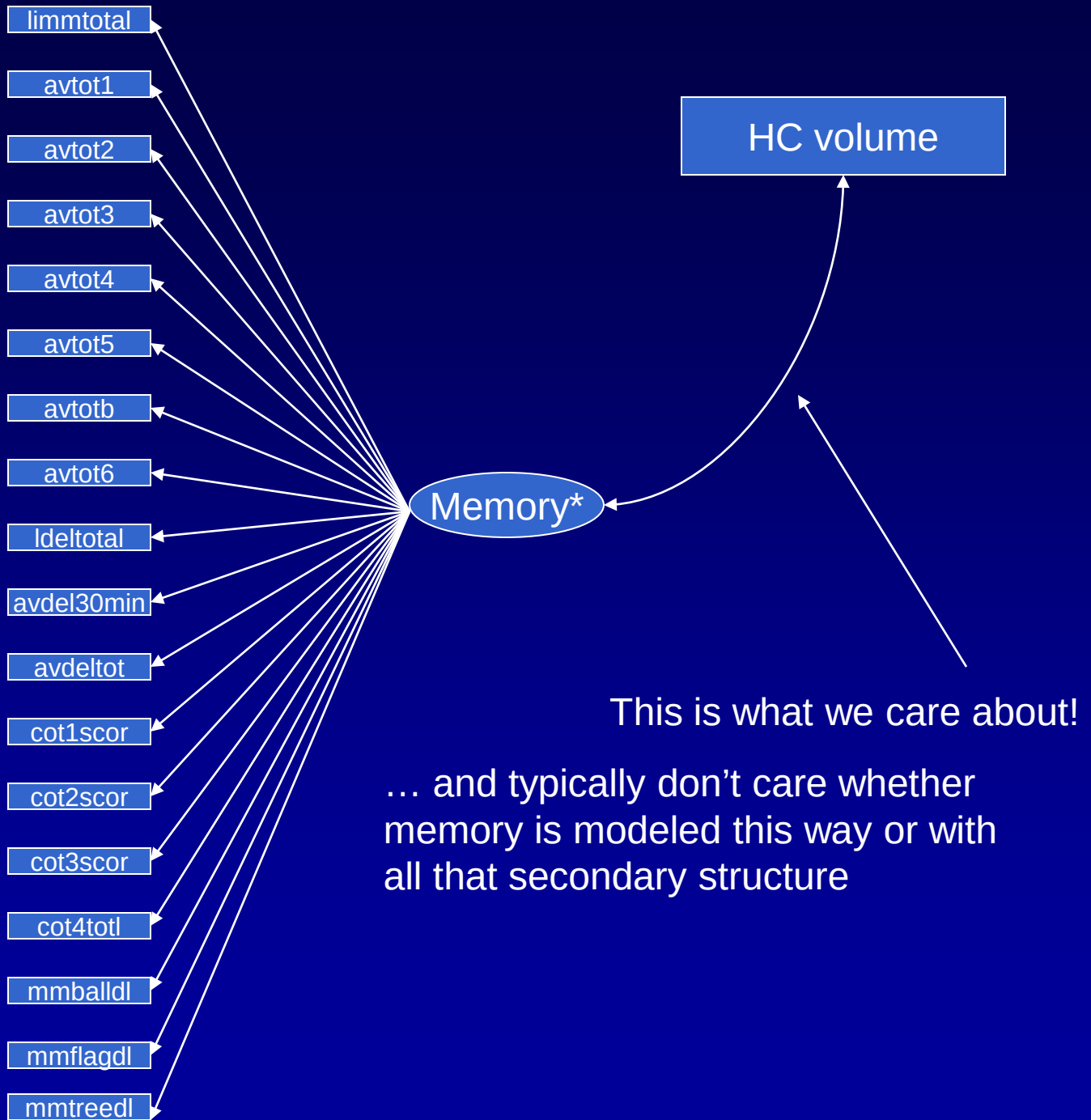
Curvilinearity

- Standard scores are on an arbitrary scale that may not be linear with respect to the underlying ability measured by the test
  - Distinct implications for change over time
  - Distinct implications for any regression analysis with the score as the dependent variable
- Latent trait / IRT / structural equation modeling scores are on an arbitrary scale that is linear with respect to the underlying ability measured by the test
- Crane et al. (2008) J Clin Epidemiol co-calibration paper
- Ehlenbach et al. (2010) JAMA paper on critical illness hospitalization

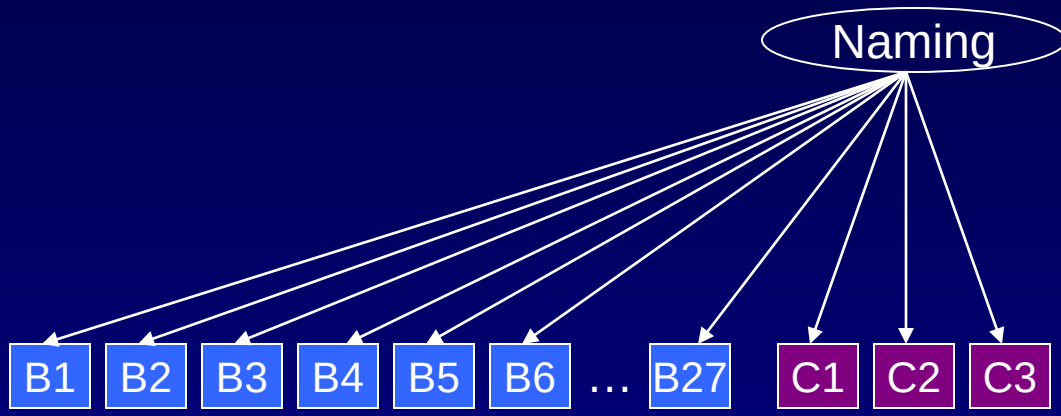
Measurement error

- Tests have uneven distributions of item difficulty
- Measurement precision / measurement error can vary quite a bit
- IRT software produce both scores and standard errors of measurement for the scores
  - Can be used in a plausible values framework
  - Or embedded in a hierarchical IRT framework
  - Or embedded in a SEM framework
- All of these approaches propagate measurement error to other parts of the model, including the inference part we care about

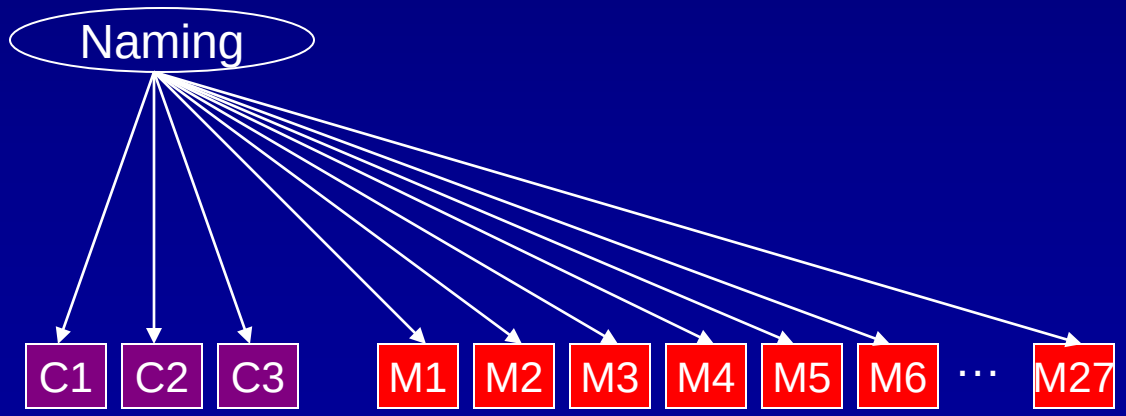

LM story
Rey word list
ADAS word list
MMSE words
limmtotal
avtot1
avtot2
avtot3
avtot4
avtot5
avtotb
avtot6
ldeltotal
avdel30min
avdeltot
cot1scor
cot2scor
cot3scor
cot4totl
mmballdl
mmflagdl
mmtreedl
Memory
HC volume
This is what we care about!


limmtotal
avtot1
avtot2
avtot3
avtot4
avtot5
avtotb
avtot6
ldeltotal
avdel30min
avdeltot
cot1scor
cot2scor
cot3scor
cot4totl
mmballdl
mmflagdl
mmtreedl
Memory*
HC volume
This is what we care about!
… and typically don't care whether memory is modeled this way or with all that secondary structure

Pictorially



"Boston Naming Test"



"Multilingual Naming Test"

CAT

Other data at the sites, e.g. memory

MMSE and MoCA

	MMSE	MOCA
Common	Naming (pen, watch) Drawing (pentagons) Repeat sentence (no if's) Recall (3 words) Orientation to time (5) Orientation to place (5)	Naming (lion, hippo, camel) Drawing (cube) Repeat sentence (x 2) (different) Recall (5 words) Orientation to time (4) Orientation to place (2)
Unique MMSE	Registration (3 words) WORLD Read sentence Write sentence Three step command	Two trials in MoCA but not scored Serial 7's instead None None None
Unique MoCA	None None None None None None – WORLD, different scoring None None	Mini trails Clock Digits forwards Digits backwards Tapping with each A Serial 7's Fluency (F words) Similarities

Not as certain these are measuring the same thing as the two naming tests

- MUCH more executive flavor to the MOCA than the MMSE
- Would want to play with a lot of item-level data from both tests before I recommended any approach for co-calibration
  - "global" in that coverage of any particular domain is too thin to obtain a meaningful subscore
  - Differential coverage of domains leads to different flavors
- I would imagine total score wise a wide variety of MOCA scores at each MMSE score because of variability in executive functioning (ignored by MMSE)
- 13 points of MMSE not covered by the MOCA
- 13 points of MOCA not covered by the MMSE
  - Any analysis providing "equivalent" MMSE scores for a MOCA (or vice versa) would have to be very cautiously interpreted

Other thoughts on the MOCA

- Adding a clock copy might be helpful
- David Libon work on using a digital pen paradigm for clock draw and clock copy
  - Incredibly granular data
  - Can still be rolled up to the MOCA
- Collect everything that is assessed
  - Learning trials

Summary

- Some domains don’t face a problem
- Some domains it’s straight forward to migrate
- Naming I would recommend IRT
- MOCA and MMSE are too dissimilar for strong inference
  - Would collect a cross-walk anyway
  - No reason not to see whether one could score the MOCA with IRT
  - Collect granular data from the MOCA, e.g. letter fluency score, not just >11 or not
  - Consider adding clock recall, digital pen
- Perhaps other opportunities to extend the value of the NACC database
  - UDS is a common denominator, much of the value of existing data is at parent sites, could be so much more

References

- Borsboom D. The attack of the psychometricians. Psychometrika. 2006; **71**(3): 425-40.
- Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. 2008; **61**(10): 1018-27 e9.
- Ehlenbach WJ, Hough CL, Crane PK, Haneuse SJ, Carson SS, Curtis JR, et al. Association between acute care and critical illness hospitalization and cognitive function in older adults. JAMA. 2010; **303**(8): 763-70.
- Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, et al. Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. Qual Life Res. 2011.
- Lord FM, Novick MR. Statistical theories of mental test scores, with contributions by Allan Birnbaum. Reading, MA: Addison-Wesley; 1968.

- pcrane@uw.edu