Methods of Inference in Clinical Neuropsychology

Pacific Northwest Neuropsychological Society

David J. Schretlen

March 3, 2018

JOHNS HOPKINS
MEDICINE

Contact: dschret@jhmi.edu

Disclosure

- Under an agreement with Psychological Assessment Resources, Inc., Dr. Schretlen is entitled to a share of royalty on sales of a test used in the study described in this presentation. The terms of this arrangement are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

Methods of Inference

1. Pathognomonic sign approach
2. Pattern analysis
3. Level of performance or deficit measurement

Pathognomonic Signs

- Characteristic of particular disease or condition
- High specificity
- Present vs. absent
- Often ignored questions
  - How frequent are they in healthy individuals?
  - How reliable are they?

Should the Babinski sign be part of the routine neurologic examination?

Timothy M. Miller, MD, PhD; and S. Claiborne Johnston, MD, PhD

- 10 physicians (5 neurologists & and 5 others)
- Examined both feet of 10 participants
  - 9 w/ upper motor neuron lesions (8 unilateral; 1 bilateral)
  - 1 w/ no upper motor neuron lesion
- Babinski present in
  - 35 of 100 examinations of foot w/ UMN weakness (sensitivity)
  - 23 of 99 examinations of foot w/o UMN weakness (specificity)

Neurology (2005)

Pathognomonic?

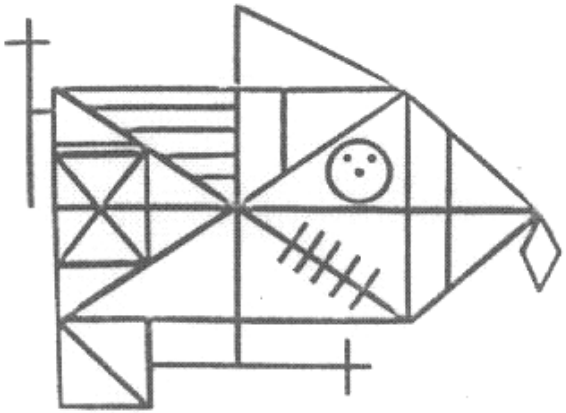


Fig. 4.8 The Complex Figure of Rey (Rey, 1959). Courtesy of Les Editions du Centre de Psychologie Appliquee.

91-year-old Caucasian woman

14 years of educ (AA degree)

Excellent health

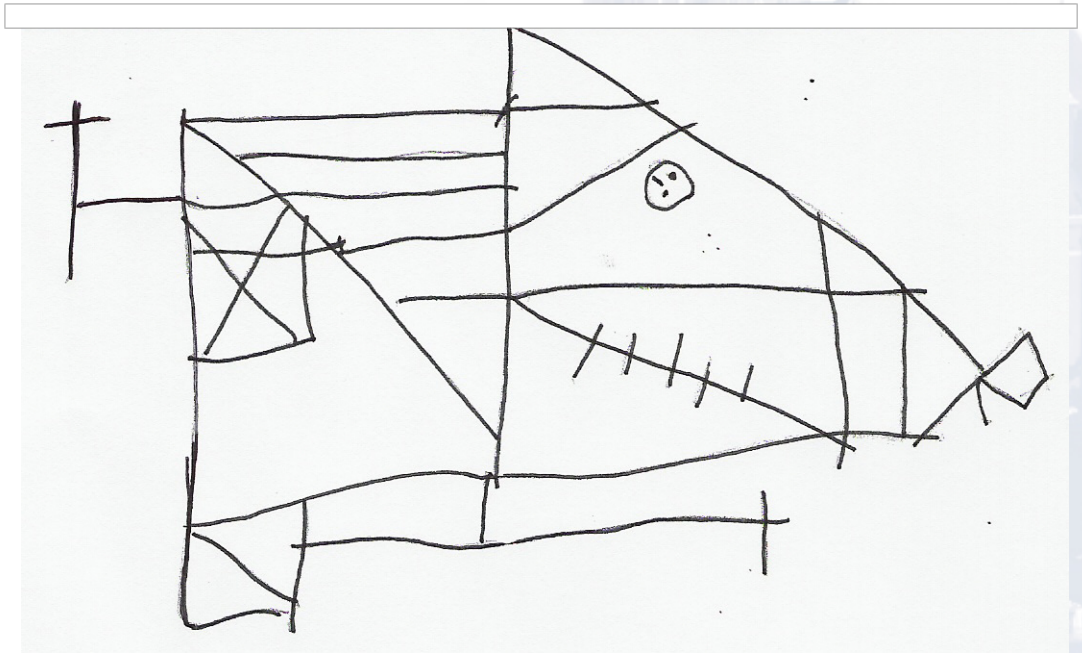
Rx: Floxin, vitamins

MMSE = 27/30

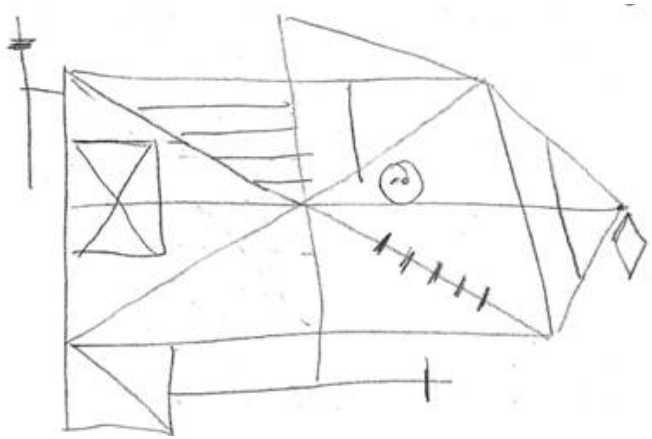
WAIS-R MOANS IQ = 109

Benton FRT = 22/27

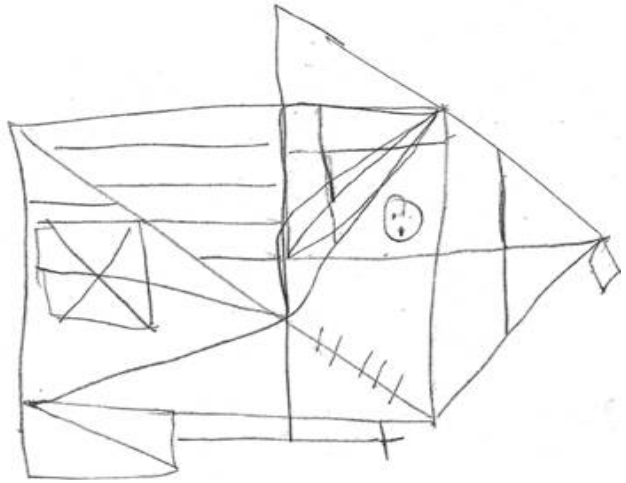
WMS-R VR Immed. SS = 8



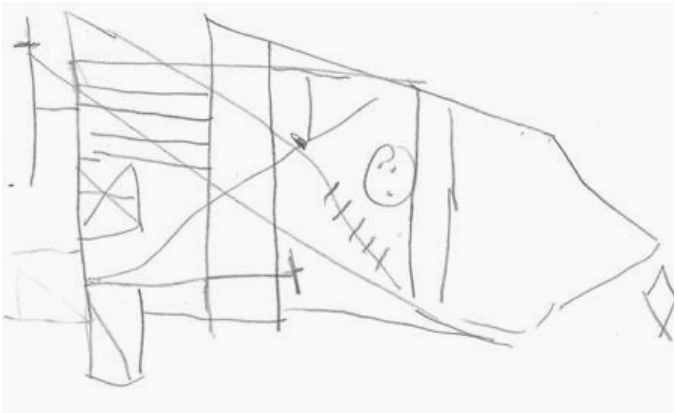
JOHNS HOPKINS MEDICINE



Jan. 2004: 68-year-old retired engineer with reduced arm swing, bradyphrenia & stooped posture. Diagnosed with atypical PD



Apr. 2005: Returns for follow-up testing 2 months after CABG; thinks his memory has declined slightly but PD is no worse



Jan. 2007: Returns & wife reports visual hallucinations, thrashing in sleep, & further memory ↓ but his PD is no worse and he still drives

Pathognomonic Signs: Limitations & Implications

- Are there any in clinical neuropsychology?
  - Unclear if there are any for a specific disease or condition
- Might be more prevalent in normal population than commonly thought
- Reliability is rarely assessed

Methods of Inference

1. Pathognomonic sign approach
2. Pattern analysis
3. Level of performance or deficit measurement

Pattern Analysis

- Many diseases impair specific aspects of cognition
- Yield recognizable gestalts of history, symptoms and cognitive test performance
  - Best for patients with typical presentations of a single disease
  - Vulnerable to errors involving the over-interpretation of normal intra-individual variability (IIV)
- Empirical basis: many studies (e.g., MMSE in AD and HD)
- How much IIV is normal?
  - Kaufman (1976) reported VIQ–PIQ discrepancies or "scatter"
  - Others (e.g., Hultsch et al. 1992; 2002; 2008) use intra-individual standard deviation, both cross-sectionally and longitudinally

Examining the range of normal intraindividual variability in neuropsychological test performance

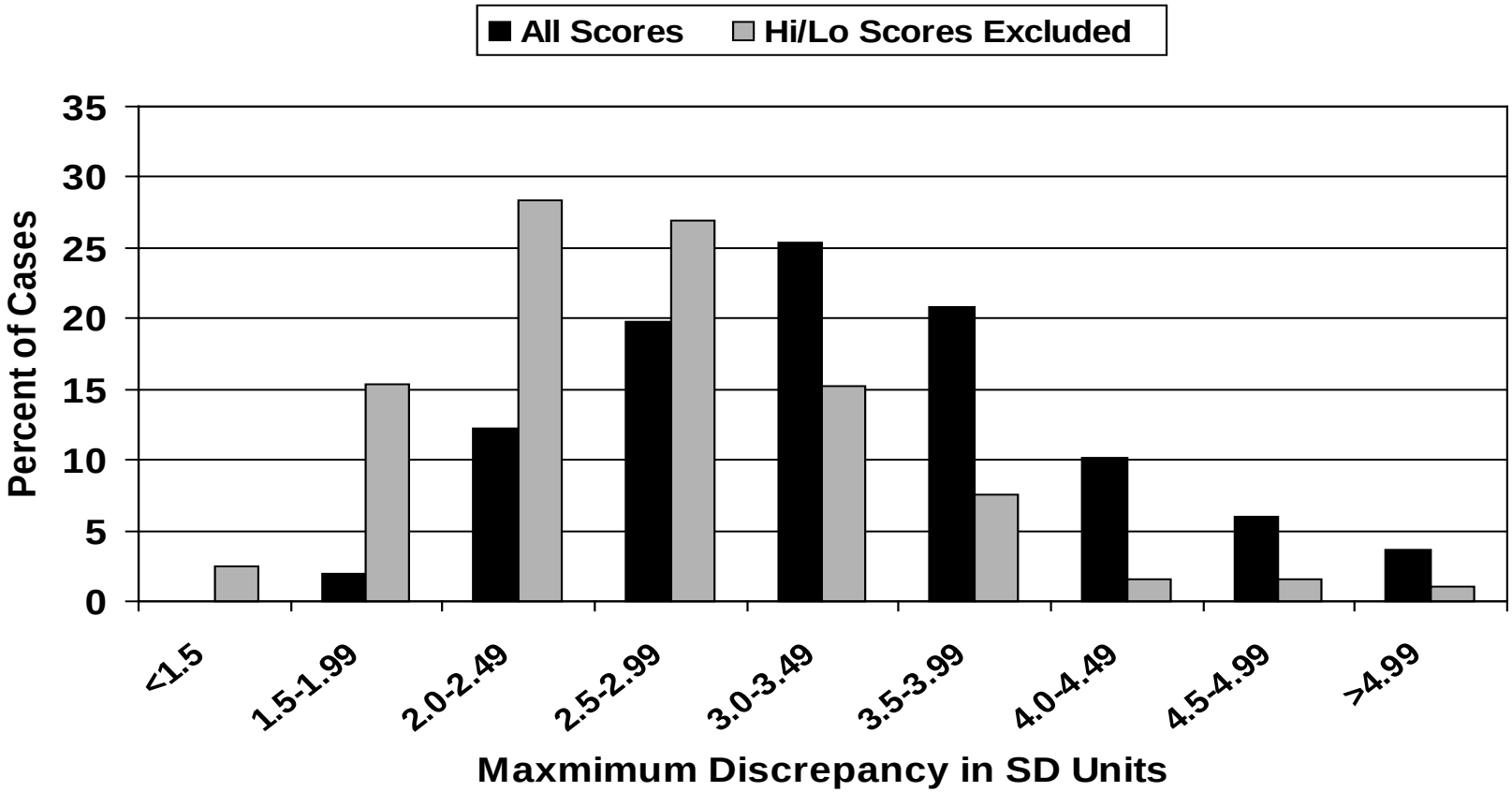
DAVID J. SCHRETLEN,[1] CYNTHIA A. MUNRO,[1] JAMES C. ANTHONY,[1,2] AND GODFREY D. PEARLSON[1,2]

[1]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland
[2]Department of Mental Hygiene, Johns Hopkins University School of Public Health, Baltimore, Maryland

- Derived 32 *z*-transformed test scores for 197 healthy Ss
- Subtracted each person's lowest *z*-score from his or her own highest *z*-score to measure the "Maximum Difference" (MD)
- Resulting MD scores ranged from 1.6 – 6.1 (*M*=3.4)
- 65% produced MD scores $\geq$3.0 and 20% had MDs $\geq$4.0
- Excluding each persons' highest and lowest test scores decreased their MDs, but 27% still produced MD values of $\geq$3.0

JOHNS HOPKINS MEDICINE

Intra-individual variability shown by 197 healthy adults



JOHNS HOPKINS MEDICINE

Pattern Analysis: Limitations

- Applicability varies with typicality of patient
- Non-contingent reinforcement can lead to idiosyncratic clinical beliefs
- Normal variation can be mistaken for meaningful patterns
- What “significant” VCI—PRI discrepancies actually mean
  - That a person’s “true” verbal and nonverbal intellectual abilities are not identical
  - Same thing applies to IQ—Memory discrepancies, etc.

Methods of Inference

1. Pathognomonic sign approach
2. Pattern analysis
3. Level of performance or deficit measurement

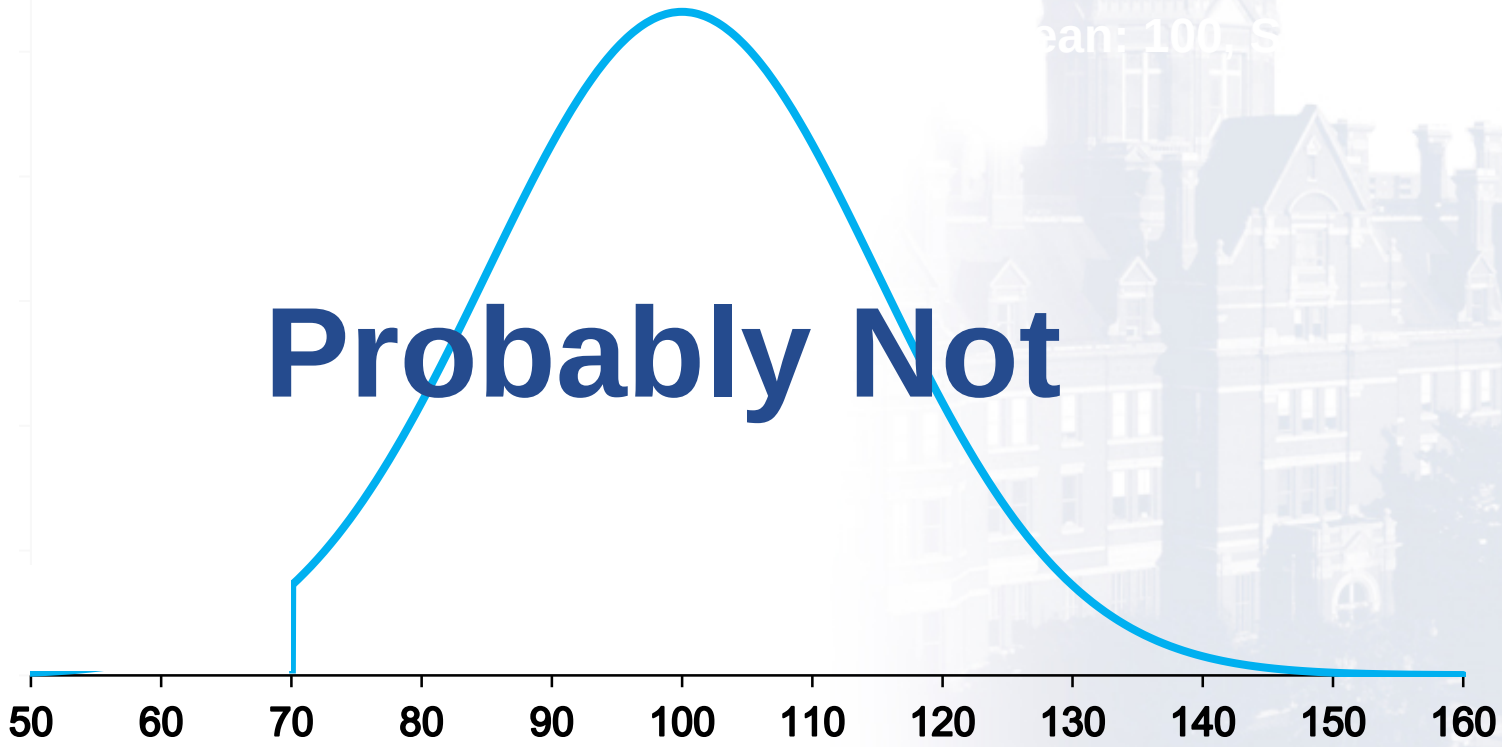
Level of Performance

- Often used to detect impairments or deficits
- But, what is an impairment or deficit?
  - Deficient ability compared to normal peers
  - Decline for individual (but normal for peers)

Level of Performance: Deficit Measurement

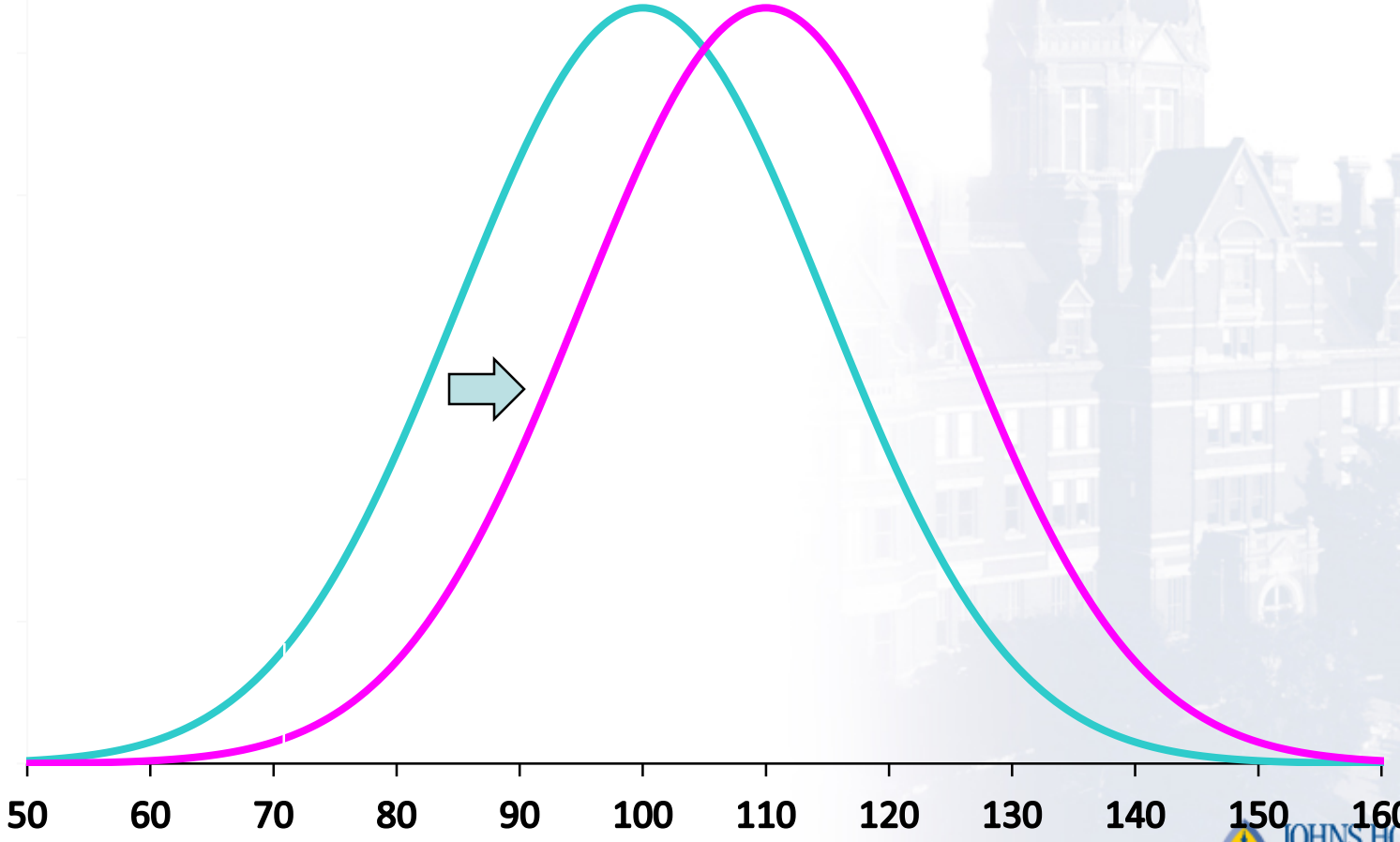
- We infer *ability* from *performance*
  - But factors other than disease (eg, effort) can uncouple them
  - There is no one-to-one relationship between brain dysfunction and abnormal test performance *at any level*
- But even if other factors do not uncouple them, what is an *abnormal* level of performance?
- Thought experiment: Suppose we test the IQs of 1,000,000 perfectly healthy adults

Would the distribution look like this?



JOHNS HOPKINS MEDICINE

More likely, the distribution would be shifted up



JOHNS HOPKINS MEDICINE

Consequently

- If a distribution of one million IQ test scores is shifted up **10 points**, but remains Gaussian, then 4800 people will still score below 70
- How do we understand normal, healthy people with IQs below 70?
  - Chance?
  - Healthy but nonspecifically poor specimens?

Logical Conclusions

- Some of those who perform in the lowest 2% of the distribution are *normal*
- Most of those who perform in the lowest 2% of the distribution are *impaired*
- The probability of impairment increases with distance below the population mean

Cutoff Scores

- Help decide whether performance is abnormal
- Often set at 2 *sd* below mean, but 1.5 and even 1 *sd* below mean have been used
- If test scores are normally distributed, these cutoffs will include 2.3% to 15.9% of normal individuals on any single measure

Multiple Measures

- When a test battery includes multiple measures, the number of normal healthy individuals who produce abnormal scores increases
- So does the number of abnormal scores they produce
- Using multiple measures complicates the interpretation of abnormal performance on test batteries

Frequency and bases of abnormal performance by healthy adults on neuropsychological testing

DAVID J. SCHRETLEN,[1,2] S. MARC TESTA,[1] JESSICA M. WINICKI,[1]
GODFREY D. PEARLSON,[1,3,4] AND BARRY GORDON,[5,6]

[1]Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland
[2]Russell H. Morgan Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, Maryland
[3]Olin Neuropsychiatry Research Center, Hartford Hospital/Institute of Living, Hartford, Connecticut

- Participants
  - 327 neurologically normal adults aged 18–92 years
- Procedure
  - Administered 25 cognitive measures; obtained T-scores
  - Classified T-scores as normal or “abnormal” based on three cutoffs: <40, <35, and <30
  - Tallied number of abnormal scores for each person (CII)
  - Used both unadjusted and demographically adjusted scores

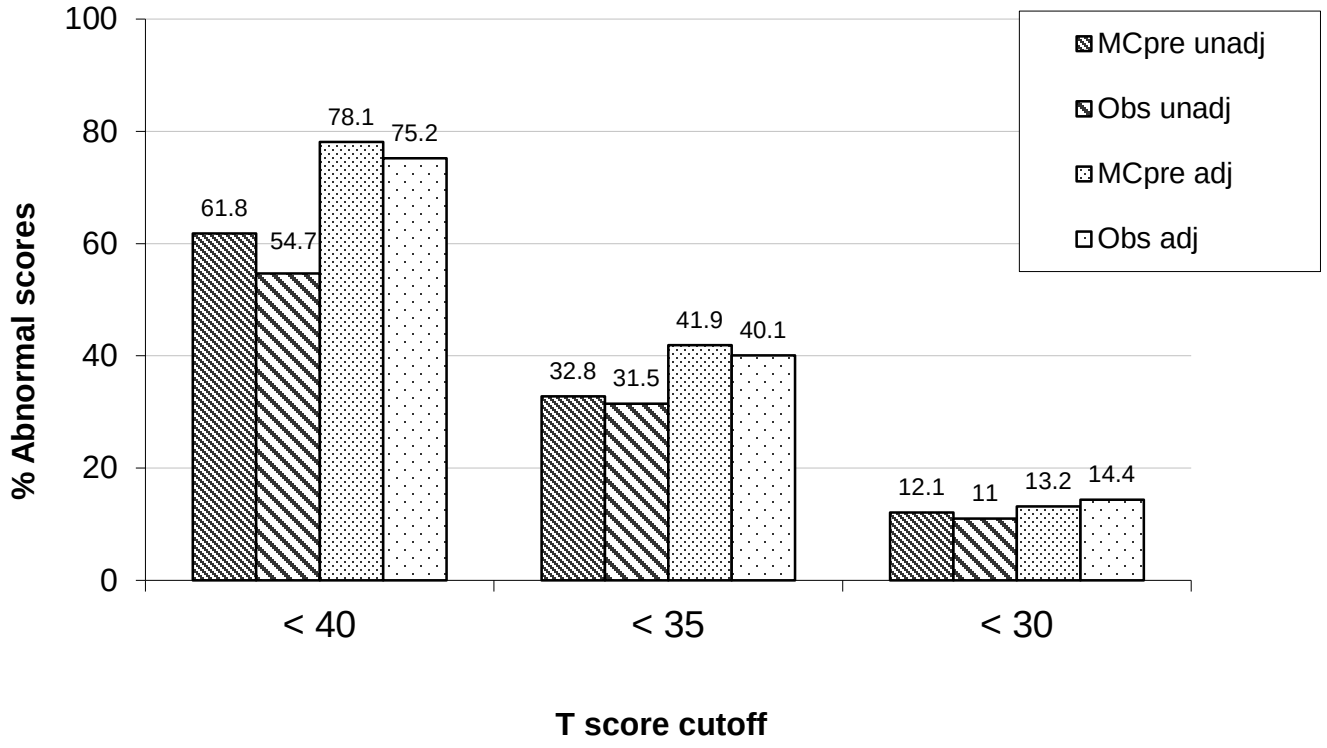
JOHNS HOPKINS MEDICINE

- We estimated how many individuals would produce 2 or more abnormal scores using three T-score cutoffs
  1. Based on binomial distribution (BN)
  2. Based on Monte Carlo simulation (MC) using unadjusted T-scores
  3. Based on Monte Carlo simulation (MC_{adj}) using adjusted T-scores

Test/Measure	**M ± SD**
Mini-Mental State Exam	28.1 ± 1.7
Grooved Pegboard Test	
Dominant hand	80.4 ± 28.1
Non-dominant hand	90.5 ± 34.7
Perceptual Comparison Test	64.5 ± 16.4
Trail Making Test	
Part A	34.9 ± 17.0
Part B	95.0 ± 69.4
Brief Test of Attention	15.4 ± 3.7
Modified WCST	
Category sorts	5.3 ± 1.3
Perseverative errors	2.5 ± 3.9
Verbal Fluency	
Letters cued	28.2 ± 9.2
Category cued	44.8 ± 11.4
Boston Naming Test	28.2 ± 2.6
Benton Facial Recognition	22.4 ± 2.3

Test/Measure	**M ± SD**
Rey Complex Figure	31.3 ± 4.3
Clock Drawing	9.5 ± 0.8
Design Fluency Test	14.2 ± 7.2
Wechsler Memory Scale	
Logical Memory I	26.3 ± 6.9
Logical Memory II	22.4 ± 7.5
Hopkins Verbal Learning Test	
Learning	24.6 ± 4.8
Delayed recall	8.7 ± 2.6
Delayed recognition	10.4 ± 1.6
Brief Visuospatial Memory Test	
Learning	22.2 ± 7.5
Delayed recall	8.7 ± 2.7
Delayed recognition	5.6 ± 0.7
Prospective Memory Test	0.6 ± 0.7

25 Measure Battery



Predicted and observed percentages of participants who produced 2 or more abnormal test scores (y axis) as defined by three different cutoffs (<40, <35, and <30 T-score points)

JOHNS HOPKINS MEDICINE

Spearman correlations between Cog Imp Index scores and age, sex, race, education and estimated premorbid IQ

No. of tests	T-score cutoff	Mean (SD)	Age	Sex	Race	Educ.	NART IQ
25	< 40	3.6 (4.4)	.573**	-.029	.215**	-.327**	-.360**
25	< 35	1.6 (2.7)	.528**	-.039	.186*	-.325**	-.354**
25	< 30	0.5 (1.3)	.409**	-.066	.176	-.312**	-.318**

* = p < 0.001; ** = p < 0.0001

JOHNS HOPKINS MEDICINE

This study shows

- Neurologically normal adults produce abnormal test scores
  - Rate varies with battery length & cutoff used to define abnormal
- This is not due purely to chance
  - Varies with age, education, sex, race and est. premorbid IQ
  - Demographically adjusting scores eliminates the relationship between these characteristics and abnormal performance
- Findings underscore distinction between “abnormal” test performance and “impaired” functioning
  - Test performance can be abnormal for many reasons: impaired functioning is but one

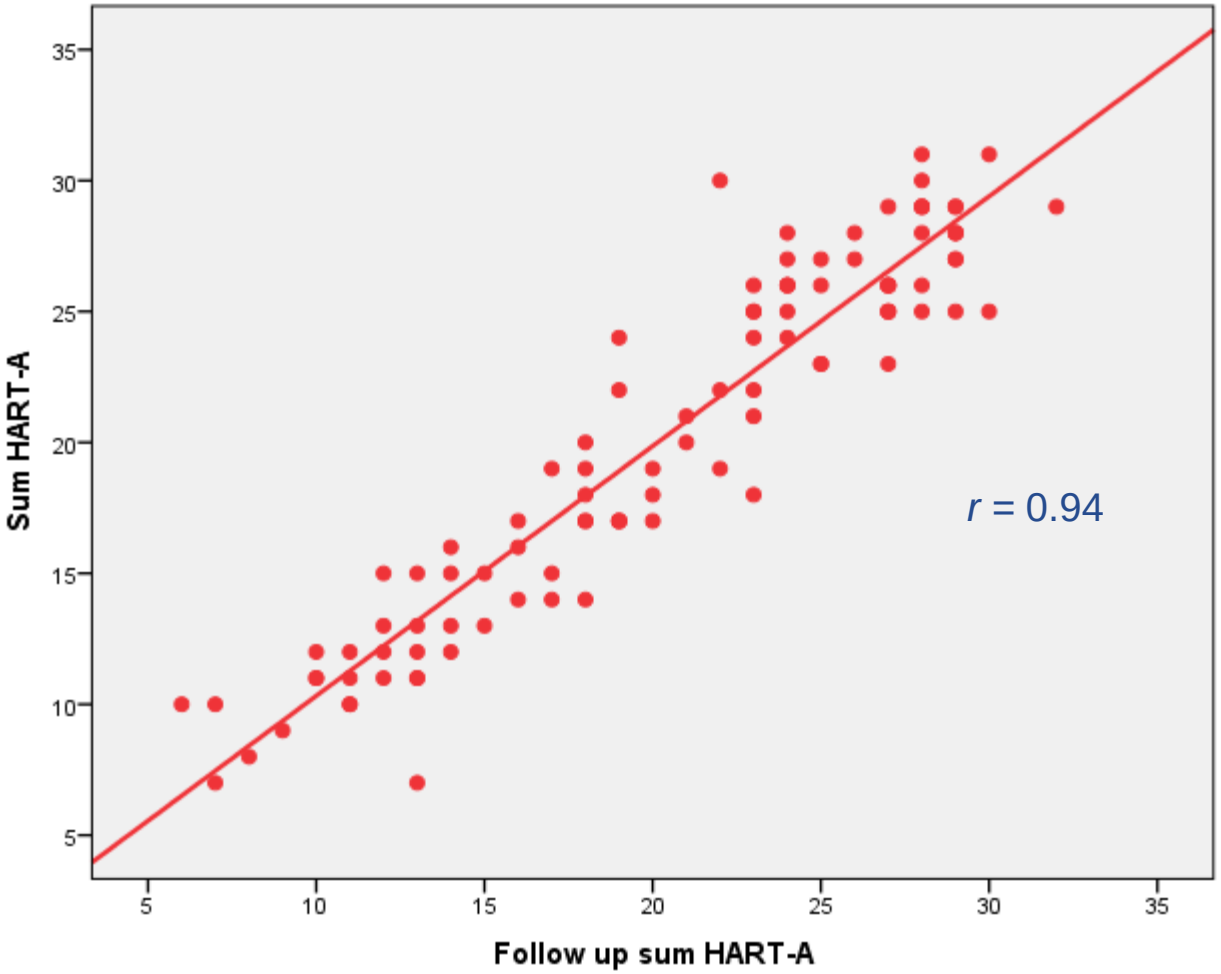
Decline from Premorbid Ability

- If we know a person's "premorbid" ability, then it is relatively simple to determine decline
  - Unfortunately, we rarely know this
  - Therefore, we have to estimate it
  - So how do we do that?
- Research has focused on estimating premorbid IQ

Estimating Premorbid IQ

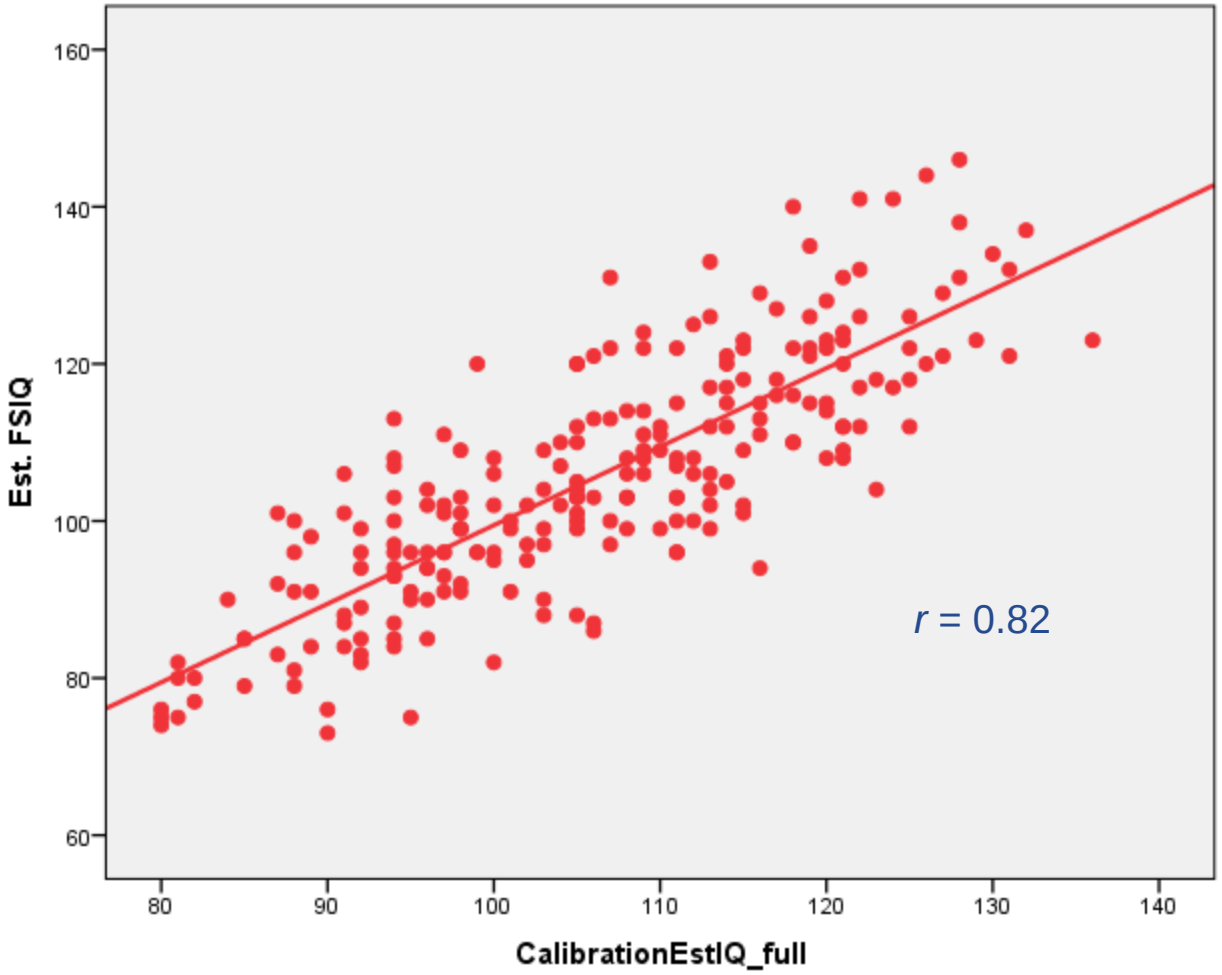
- Demographic prediction
  - Barona formula SE_{est} = 12 points (95% CI = $\pm$24 points)
- Word reading tests are more accurate
  - Except for persons with very limited education
  - And those with aphasia, reading disorders, or severe dementia
  - And persons for whom English is a second language

HART IQ estimates over 5 years



JOHNS HOPKINS MEDICINE

Correlation of HART and WAIS-R



JOHNS HOPKINS MEDICINE

The use of word-reading to estimate "premorbid" ability in cognitive domains other than intelligence

DAVID J. SCHRETLEN,[1,2] ANGELA L.H. BUFFINGTON,[1] STEPHEN M. MEYER,[1] AND GODFREY D. PEARLSON[1,3,4]

[1]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland
[2]Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland
[3]Olin Neuropsychiatry Research Center, Institute of Living/Hartford Hospital, Hartford, Connecticut
[4]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut

But how well does the NART-R predict cognitive abilities other than IQ?

Administered 26 cognitive measures to 322 healthy adults

Regressed each on age, saved the residuals, and correlated these with NART-R scores

Compared the correlation of NART-R and IQ with correlations of the NART-R and other age-adjusted cognitive measures

JOHNS HOPKINS MEDICINE

NART-R correlation with FSIQ = .72

NART-R correlations with other test scores ranged from -.53 to .48

(Every one of the latter was significantly smaller than the correlation with FSIQ)

Table 1. Pearson r (or Spearman ρ) correlation of the NART–R with age-corrected scores on each cognitive test, standard errors of the estimates of NART–R predicted performances on the same measures, and standard scores corresponding to 5th percentile of NART–R predicted minus actual scores for each cognitive test variable

Test/variable	Correlation[1]	$p <$	SE_{Est}	5th %ile[2]
Verbal IQ (prorated)[3]	$r = .755$	.0001	9.4	13.4
Full Scale IQ (prorated)[3]	$r = .724$	.0001	10.1	15.4
GPT Dominant Hand	$\rho = -.286$	.0001	12.9	26.7
GPT Nondominant Hand	$\rho = -.276$	.0001	13.6	24.5
Trail Making Test, Part A	$\rho = -.237$	.0001	14.6	35.3
Trail Making Test, Part B	$\rho = -.528$	.0001	12.1	25.5
Brief Test of Attention	$r = .319$	.0001	14.2	31.5
mWCST Categories	$\rho = .311$	.0001	14.3	37.8
mWCST Perseverative Errors	$\rho = -.259$	.0001	14.5	33.4
Cognitive Estimation Test	$r = -.500$	.0001	13.0	27.1
CPT Hit Reaction Time	$r = .071$	n.s.	15.0	33.1
CPT Discrimination (d')	$r = .061$	n.s.	15.0	39.8
Boston Naming Test	$\rho = .384$	.0001	13.0	28.7
Word Fluency (Letters)	$r = .481$	.0001	13.1	25.7
Word Fluency (Category)	$r = .386$	.0001	13.8	29.0
Design Fluency Test	$r = .403$	.0001	13.7	27.4
Benton Facial Recognition	$r = .284$	.0001	14.4	30.3
Rey CFT (Copy)	$\rho = .328$	.0001	14.2	31.6
HVLT–R Learning	$r = .356$	.0001	14.0	31.6
HVLT–R Delay	$\rho = .349$	.0001	14.2	35.5
HVLT–R Recognition	$\rho = .142$	.05	14.4	33.0
BVMT–R Learning	$r = .318$	.0001	14.2	31.5
BVMT–R Delay	$r = .300$	.0001	14.3	31.1
BVMT–R Recognition	$\rho = .119$	.05	15.0	39.6
WMS–R Logical Memory I	$r = .419$	.0001	13.6	29.7
WMS–R Logical Memory II	$r = .422$	.0001	13.6	28.3
WMS–R Visual Reproduction I	$r = .343$	.0001	14.1	33.5
WMS–R Visual Reproduction II	$r = .258$	.0001	14.5	33.8

[1]Spearman rank order correlations were used for cognitive measures whose distributions were characterized by skewness or kurtosis >1.0; Pearson product-moment correlations were used for all others. [2]Difference between NART–R estimated Full Scale IQ and each standardized test score that included the 5% of participants with the largest discrepancies. [3]Prorated using Ward's (1990) seven-subtest short form of the WAIS–R or WAIS–III.

Estimating Premorbid Abilities

- An *essential* and *unavoidable* aspect of every neuropsychological examination
- If we don't do explicitly, then we do it implicitly
- Even the best methods yield ballpark estimates
- We're better at estimating premorbid IQ than other premorbid abilities

Conclusions

- Deficit measurement limitations and implications
  - No isomorphic relationship between performance and ability
  - Adding tests can increase false positive (type 1) errors
  - Setting stringent cut-offs can increase misses (type 2) errors
- Pathognomonic sign and pattern analysis approaches also have limitations and threats to their validity
- Recognizing these is essential to maximize the usefulness and minimize the dangers of assessment
- Understanding them can guide future research

How Normal is “Normal”?

- Hypothesis
  - Most healthy adults will produce normal (Gaussian) distributions of scores on a battery of tests

Method

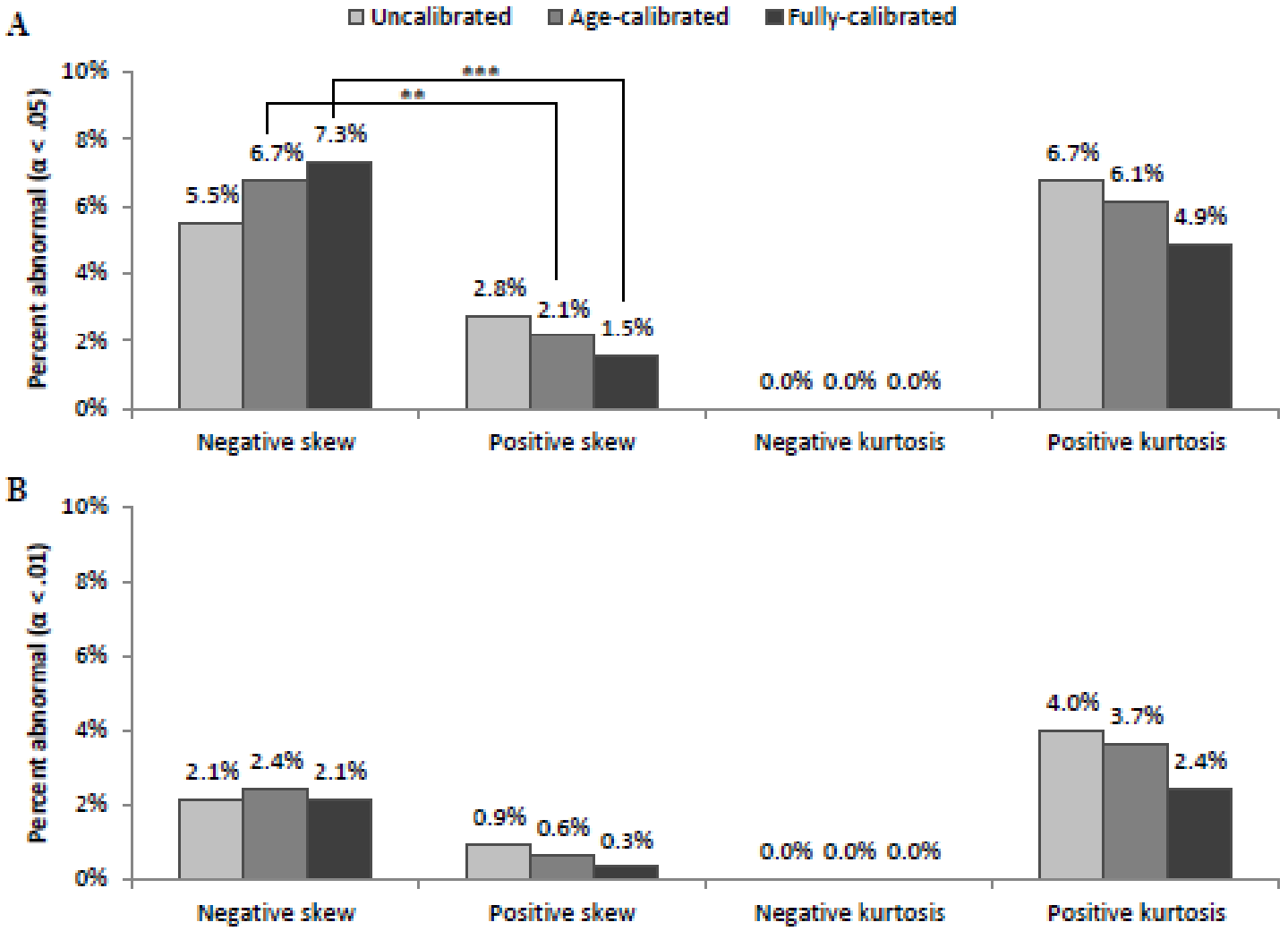
- Participants:
  - 327 neurologically normal adults from the ABC study
  - Constituted the normative sample for the Calibrated Neuropsychological Normative System (CNNS)

Variable	**Mean ± *SD* or *n***	***Range or %***
Age, years	54.8 ± 18.8	18 – 90
Education, years	14.2 ± 3.0	3 – 20
Sex, male / female	142 / 185	44.4 / 56.6
Race, white / black	268 / 59	82.0 / 18.0

Method

- Procedure:
  - Based analyses on 30 measures derived from 19 tests
  - Raw scores transformed to non-calibrated, age-calibrated, and fully-calibrated T-scores
  - We computed each person's overall test battery mean, standard deviation, skew, and kurtosis
  - We converted each person's skew and kurtosis values to z-scores by dividing each by their respective standard error
    - Determined proportion of participants with significant levels of skew and kurtosis at $p < .05$ ($|z| > 1.96$) and $p < .01$ ($|z| > 2.58$)
  - Examined correlates of within-person distribution parameters

Results



JOHNS HOPKINS MEDICINE

Results

Correlations between distributional properties and participant characteristics.

Calibration	Parameter	Age	Education	Sex†	Race†	IQ
Uncalibrated	Mean T-score	-.62**	.40**	.02	-.27**	.54**
	Standard deviation	.04	-.09	-.02	.13*	-.17**
	Z-skew	.17**	.03	-.03	< .01	.03
	Z-kurtosis	-.08	.02	-.03	-.03	.04
Age calibrated	Mean T-score	< -.01	.47**	-.03	-.44**	.77**
	Standard deviation	.13*	-.14*	-.01	.20**	-.25**
	Z-skew	< .01	-.09	.02	.08	-.20**
	Z-kurtosis	-.05	.05	< -.01	-.10	.21**
Fully Calibrated	Mean T-score	-.01	-.01	.01	-.02	.39**
	Standard deviation	.13*	-.12	.01	.16**	-.23**
	Z-skew	-.02	< -.01	.01	< -.01	-.12
	Z-kurtosis	-.06	-.02	< -.01	-.05	.19**

†Point-biserial correlations [Male (Sex) and Caucasian (Race) coded as the smaller value].

* $p < 0.05$; ** $p < 0.01$ (2-tailed)

JOHNS HOPKINS MEDICINE

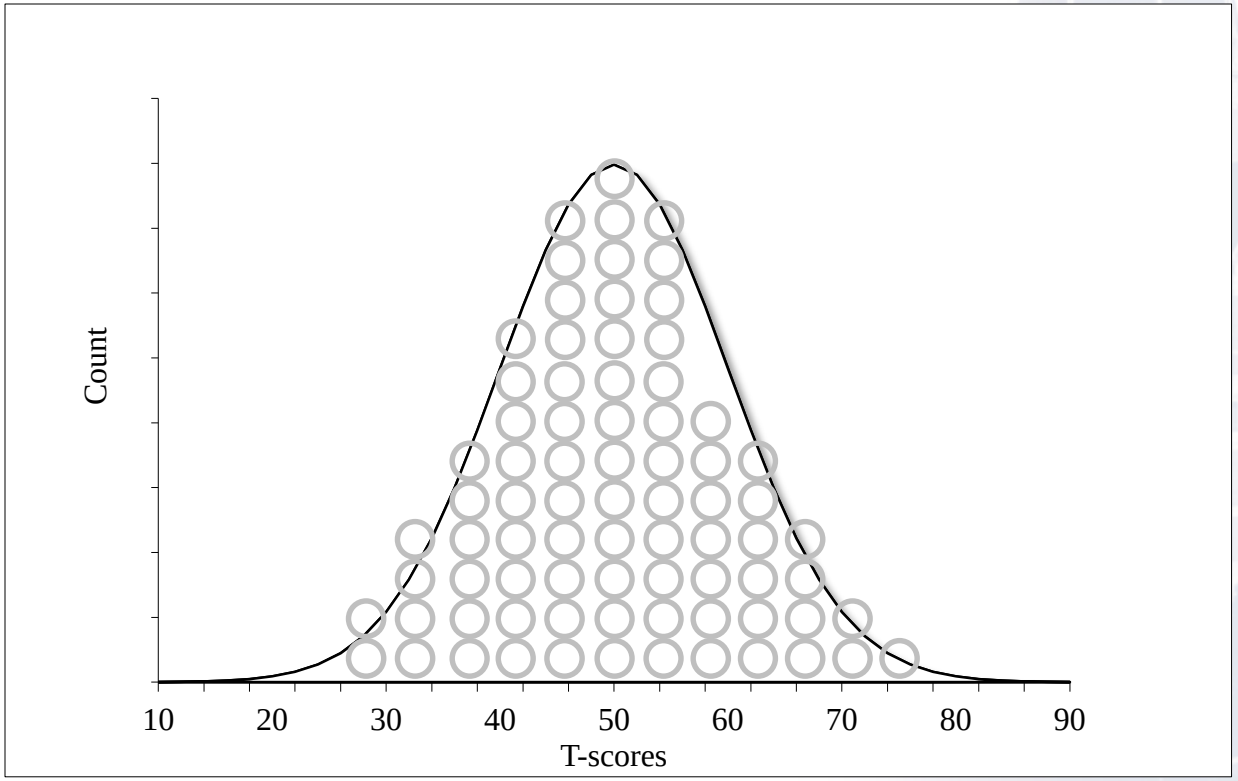
Conclusions

- Most participants produced battery-wide T-score distributions that are normal
- Protocols that differed from this showed
  - Slightly higher prevalence of negative than positive skew
  - When present, kurtosis was always positive
- Using uncalibrated data, advancing age is associated with ↓ in battery-wide mean T-scores and ↑ in skew
- Fully calibrating scores uncouples correlations of mean T-scores with age, sex, race, and education, but had little effect on rates of abnormal skew or kurtosis

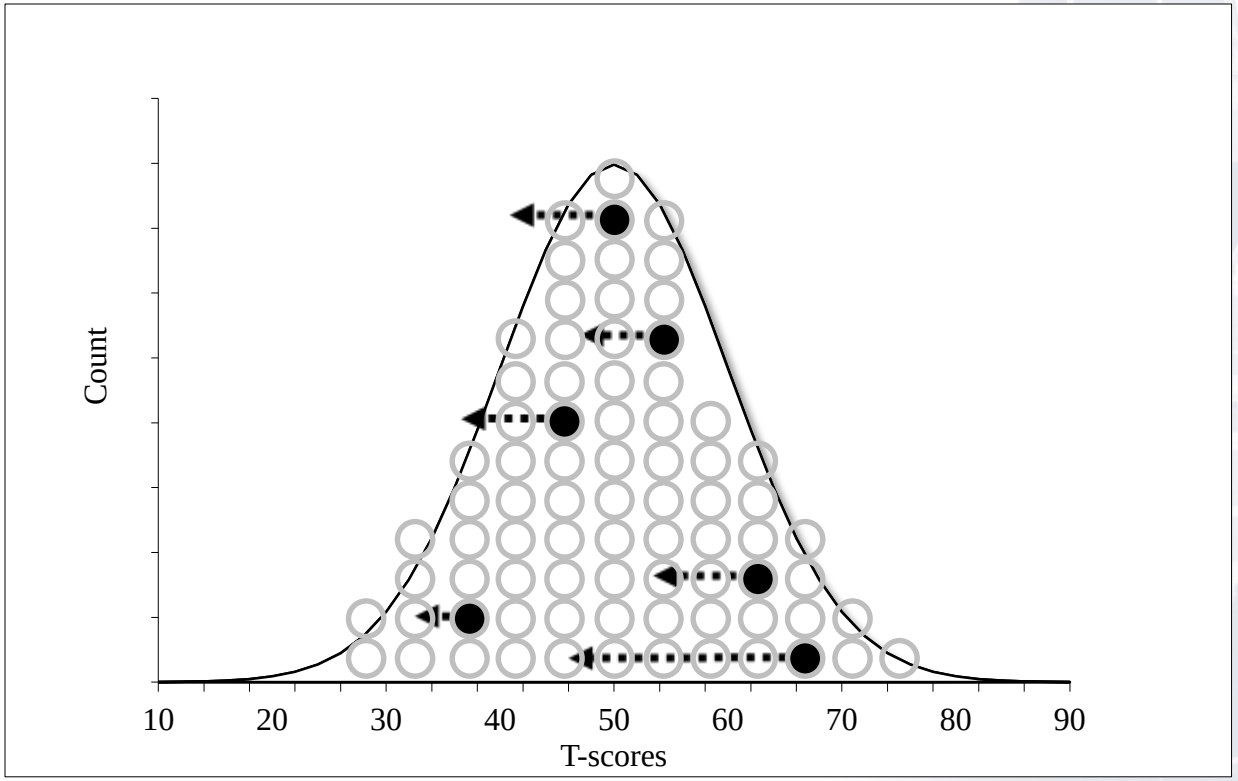
JOHNS HOPKINS MEDICINE

Can intraindividual variability help diagnose cognitive dysfunction?

Hypothetical distribution of 80 test scores shown by a healthy older adult at baseline

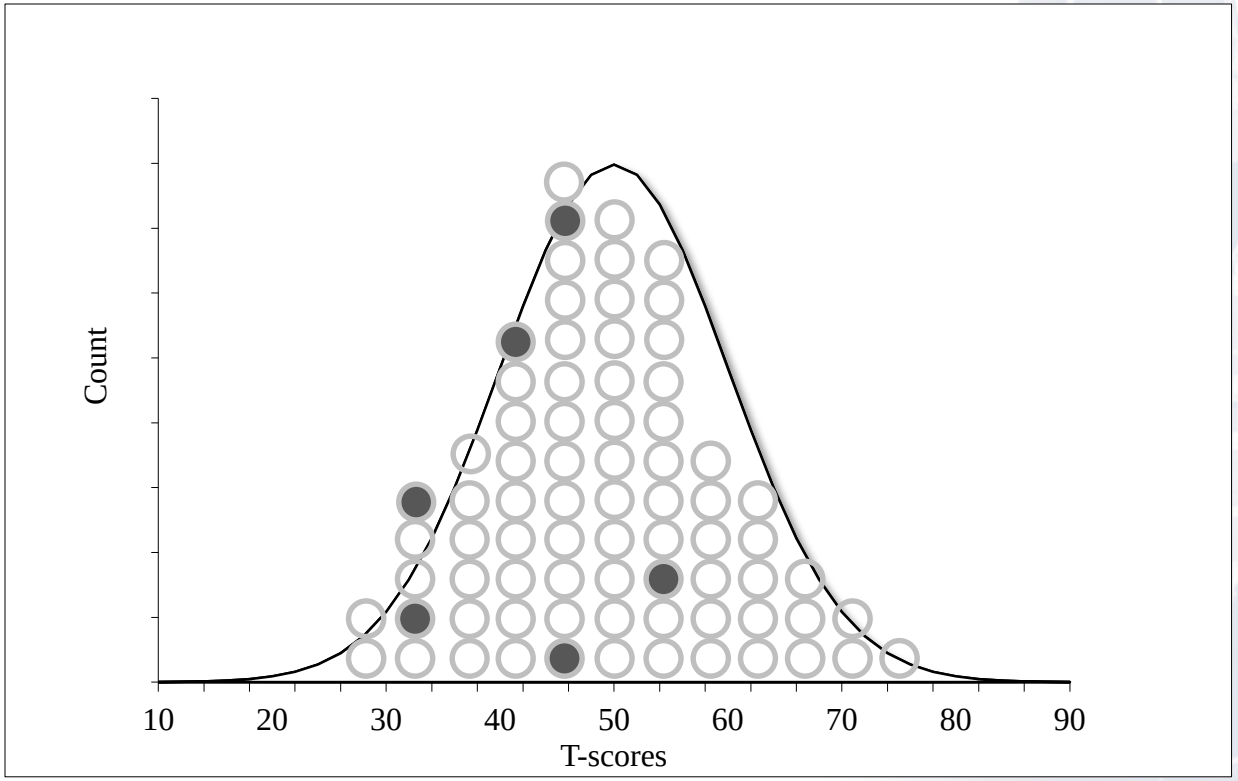


Hypothetical changes on same 80 tests after onset of MCI due to early AD

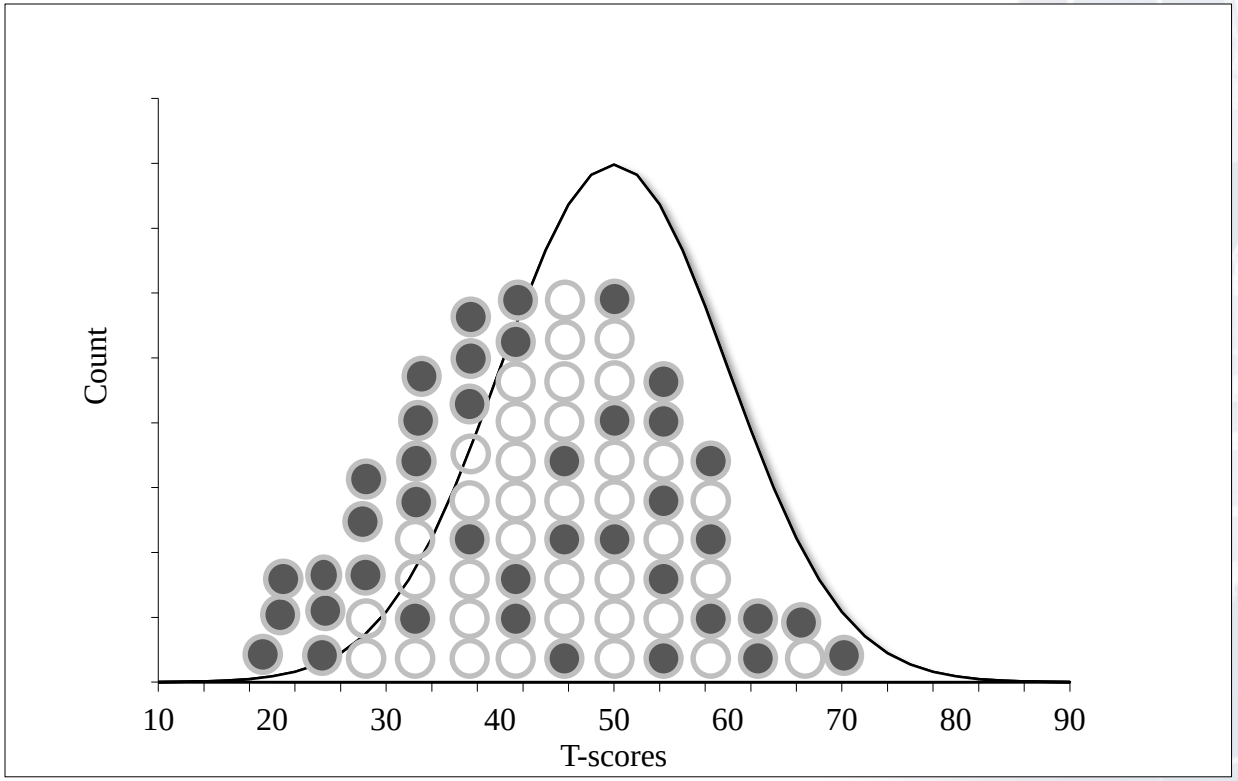


JOHNS HOPKINS MEDICINE

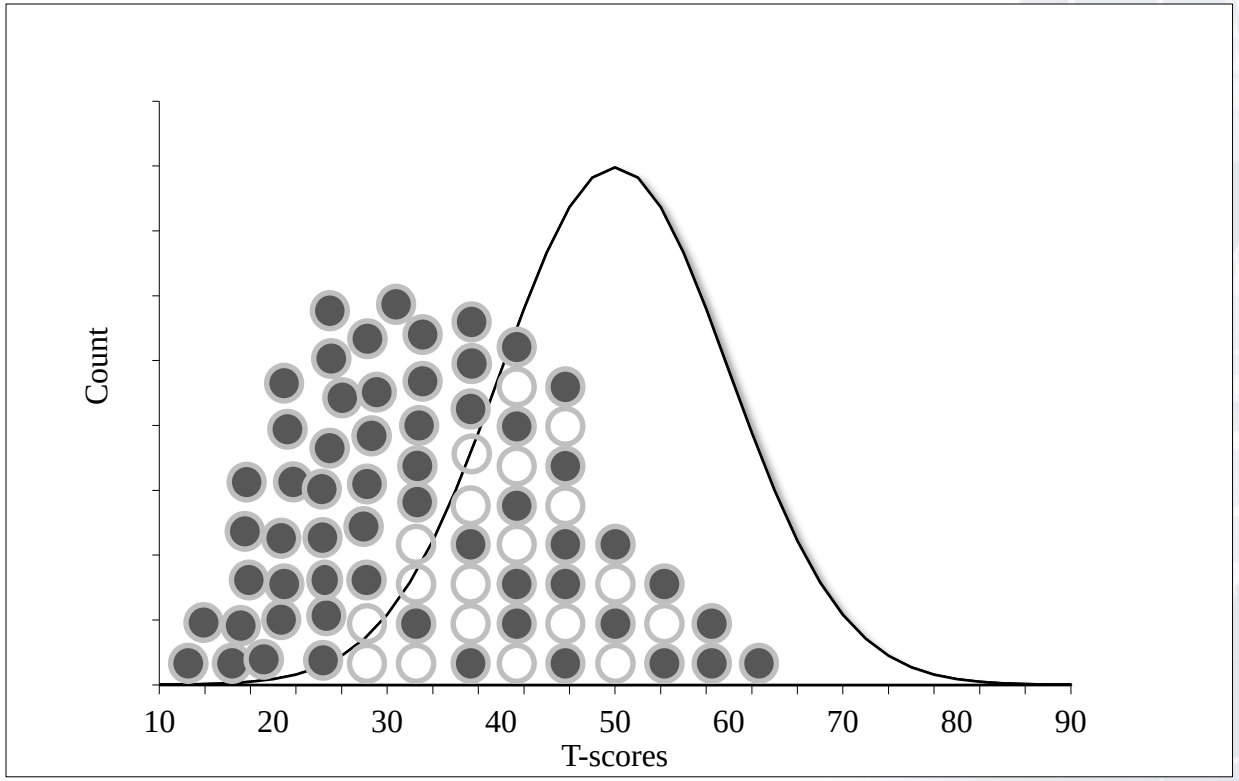
Altered distribution of 80 test scores shown by the same person with MCI at follow-up



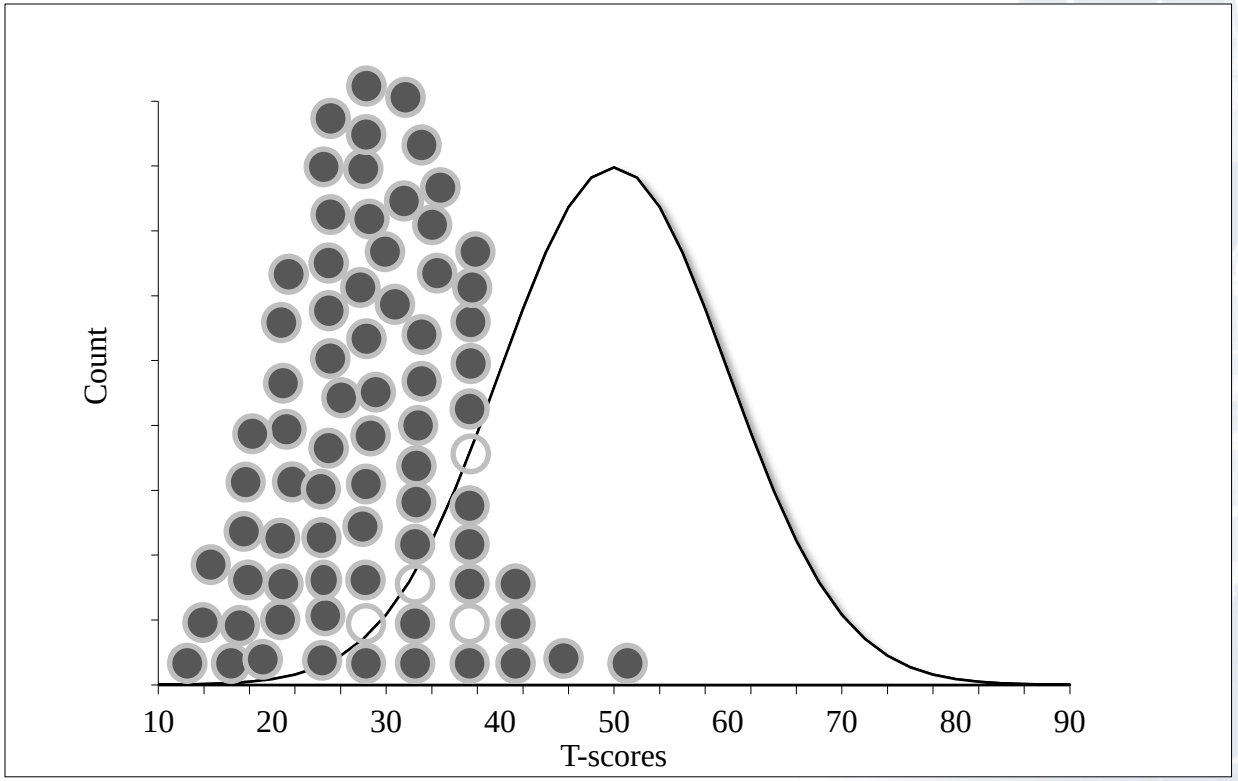
Hypothetical distribution of the 80 test scores shown by the same person, now with mild AD



Hypothetical distribution of the 80 test scores shown by the same person with moderate AD



Hypothetical distribution of the 80 test scores shown by the same person with severe AD



Method

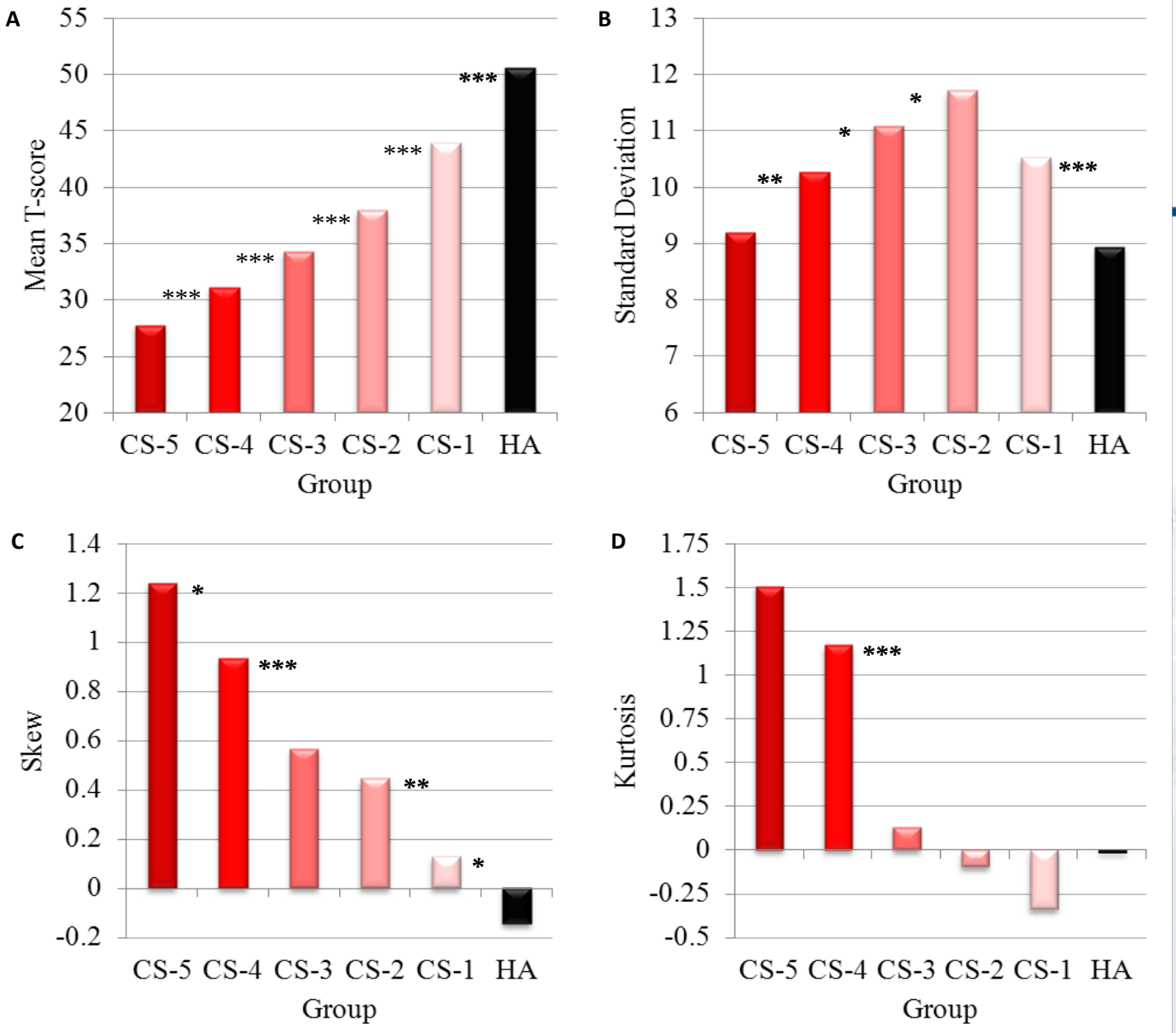
- Participants:
  - 395 patients tested for dementia work-up (MMSE 9-30)
  - 135 healthy adults from the ABC study (MMSE = 24-30)
- Procedure:
  - 13 measures from 6 tests
  - Calibrated raw scores for age, sex, race & education
  - Estimated pre-morbid ability with HART + demographics
  - Stratified patients by MMSE

Variable	Patients (n=395)	HA (n=135)	*p*
Age	76 ± 7	73 ± 8	<0.001
Education	13 ± 4	14 ± 3	0.003
Sex (% male)	38	49	0.02
Race (% white)	83	84	0.91

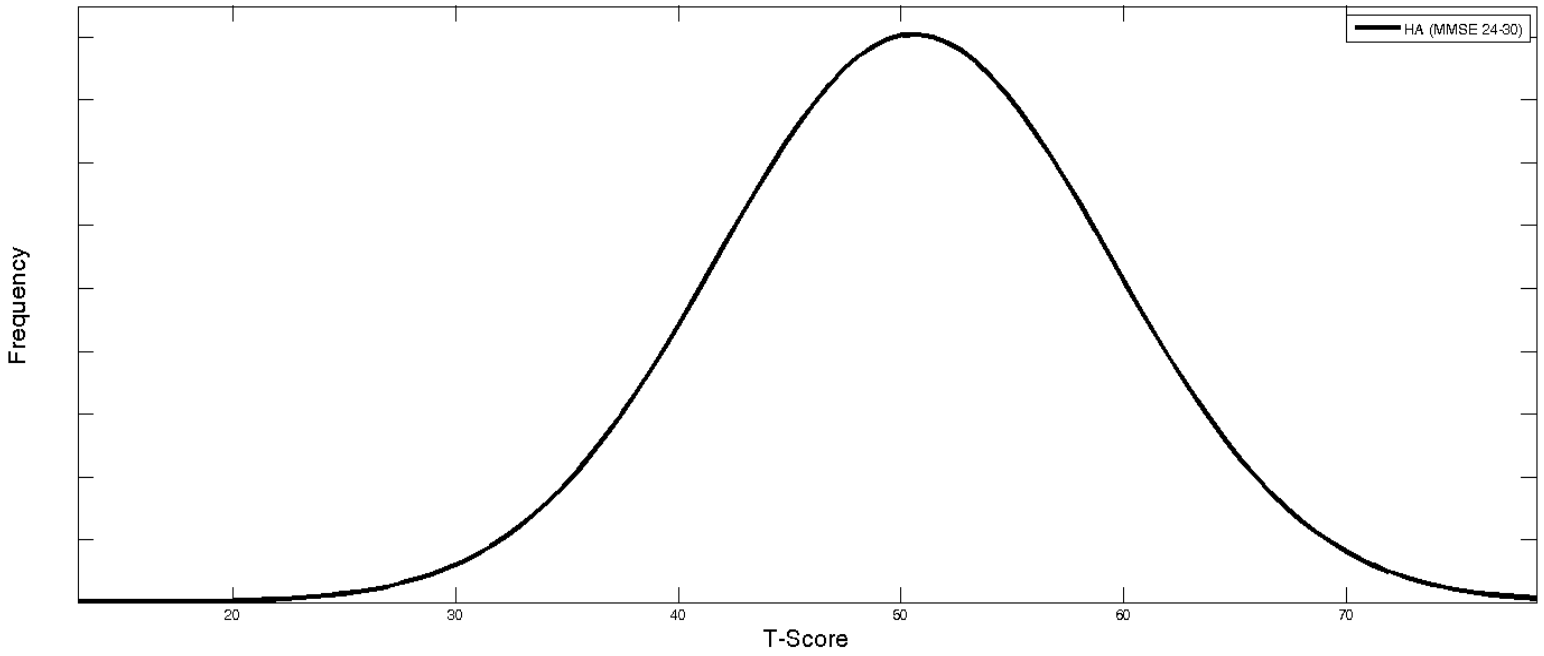
Test	Measures
TMT	Part A, Part B times
CIFA	Letter fluency, Category fluency
BNT-30	Total spontaneously correct
Clock Drawing	Command, copy
HVLT-R	Learning, delay, discrimination
BVMT	Learning, delay, discrimination

Results

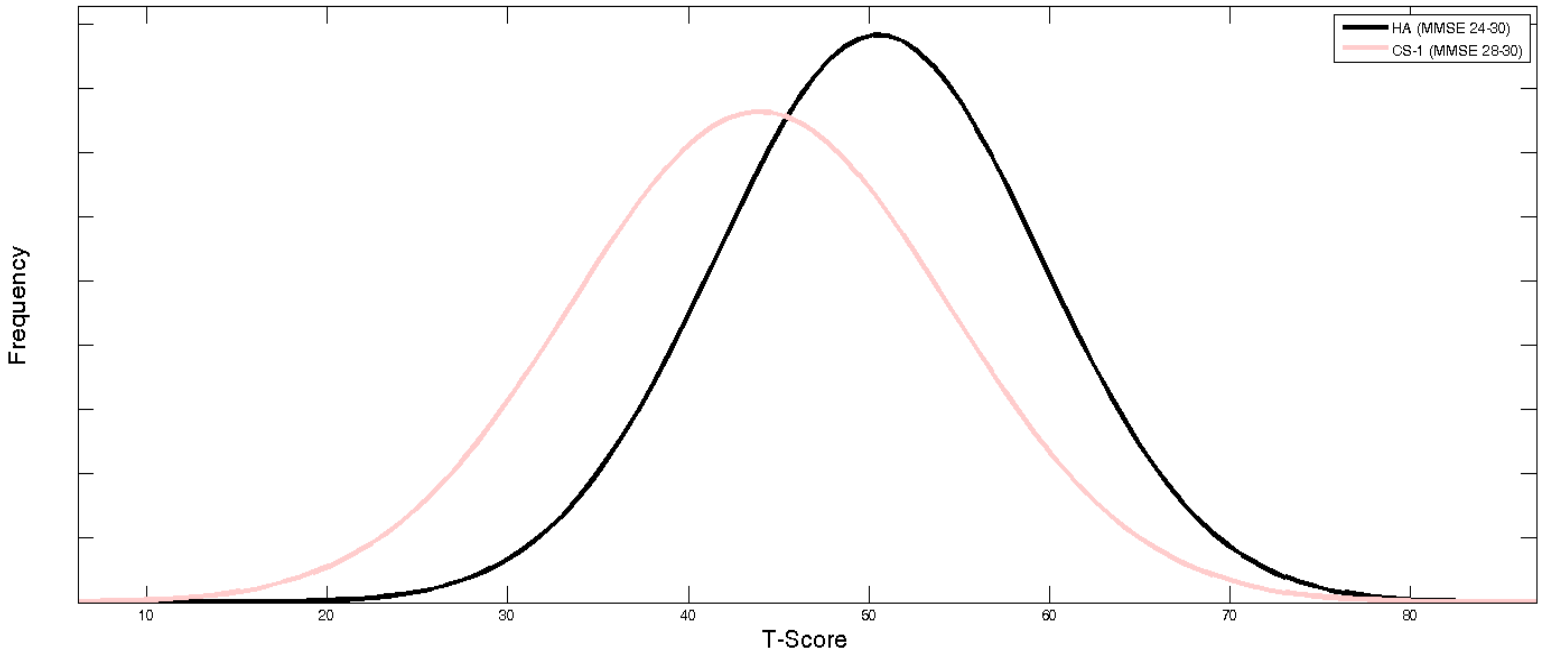
Group	MMSE range	*n*
Healthy adults (HA)	24 – 30	135
Clinical sample (CS-1)	28 – 30	47
Clinical sample (CS-2)	24 – 27	117
Clinical sample (CS-3)	20 – 23	107
Clinical sample (CS-4)	16 – 19	79
Clinical sample (CS-5)	9 – 15	45



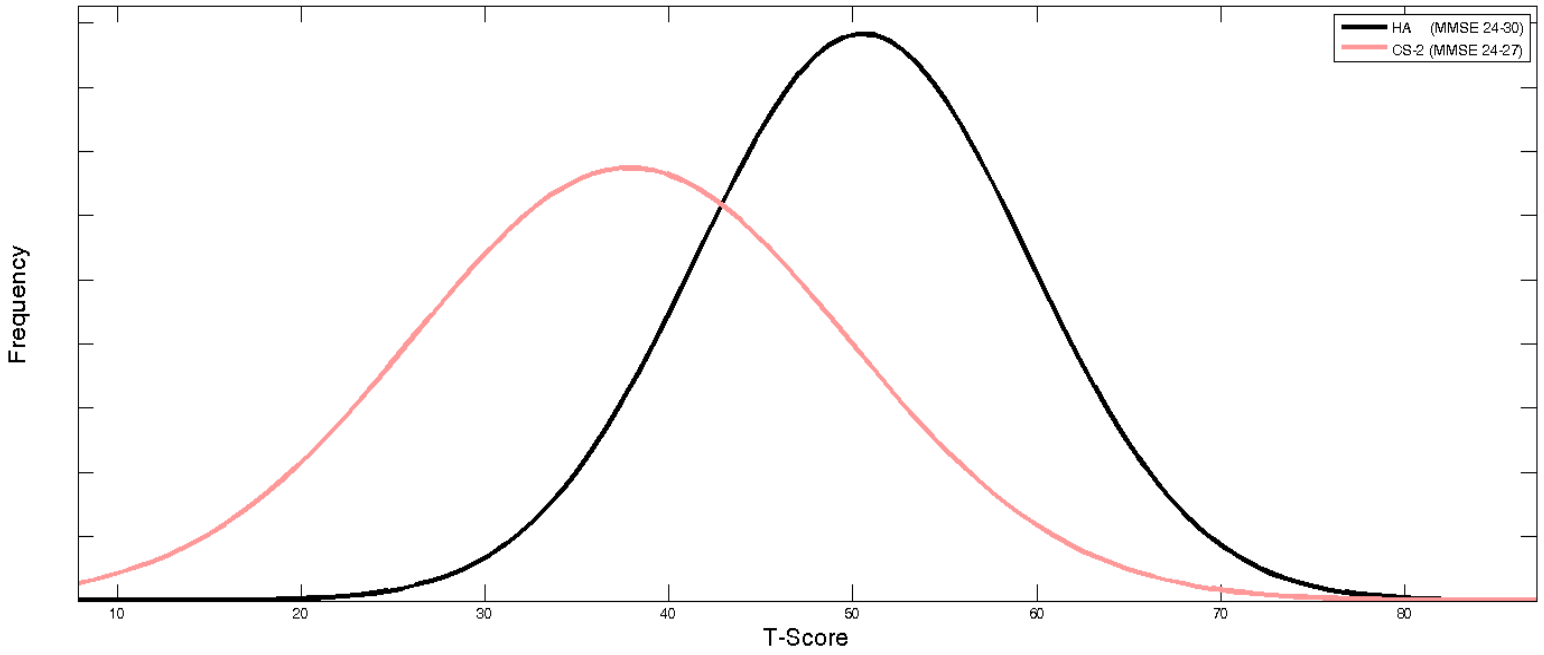
Within-person test score distribution produced by healthy adults



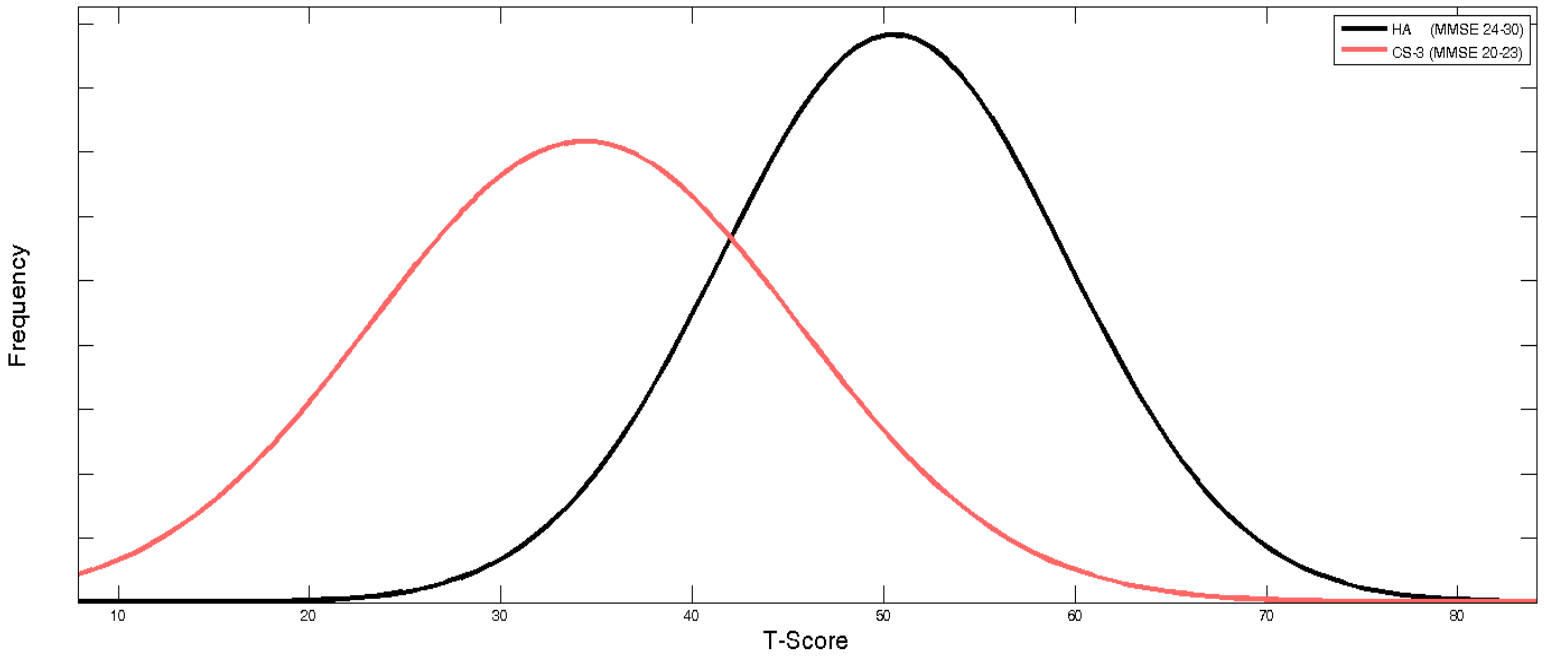
Comparison of within-person test score distributions produced by healthy adults (black line) and patients with MMSE scores of 28–30 (pink line)



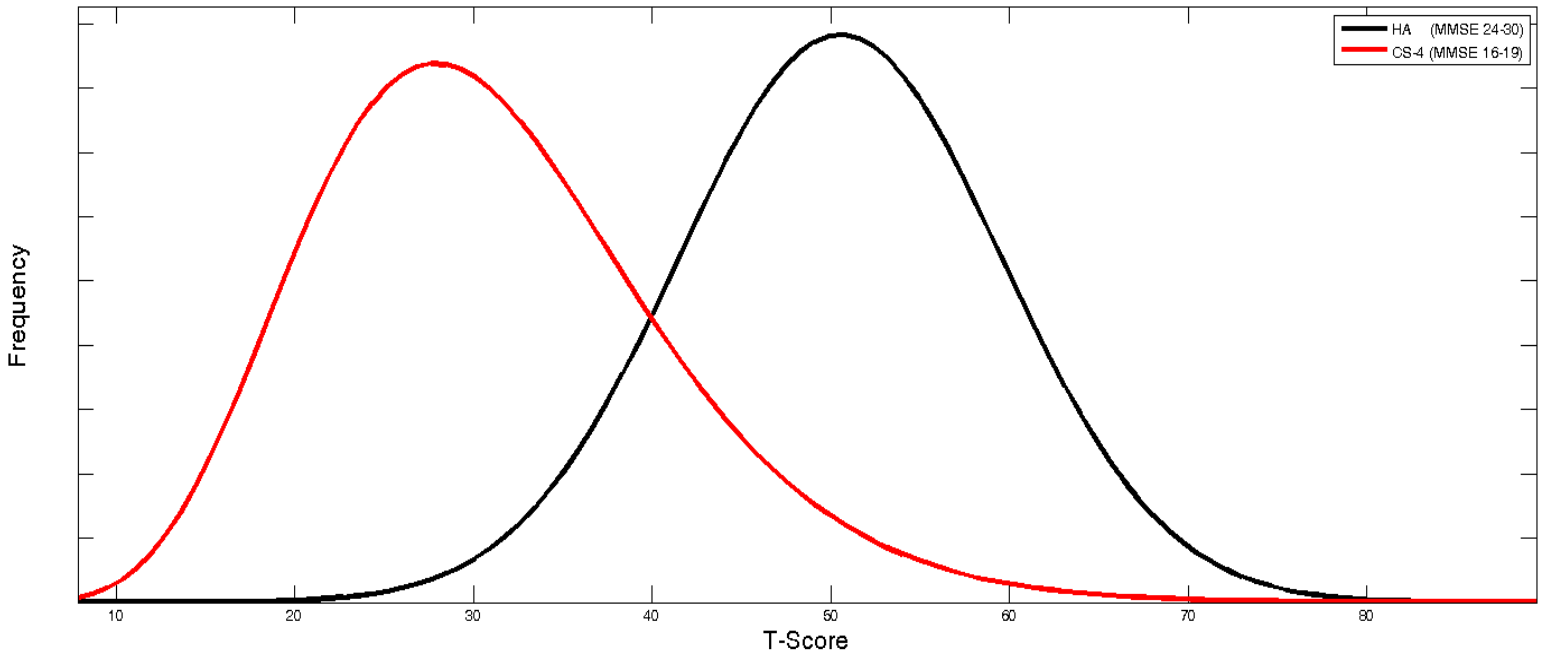
Comparison of within-person test score distributions produced by healthy adults (black line) and patients with MMSE scores of 24–27 (pink line)



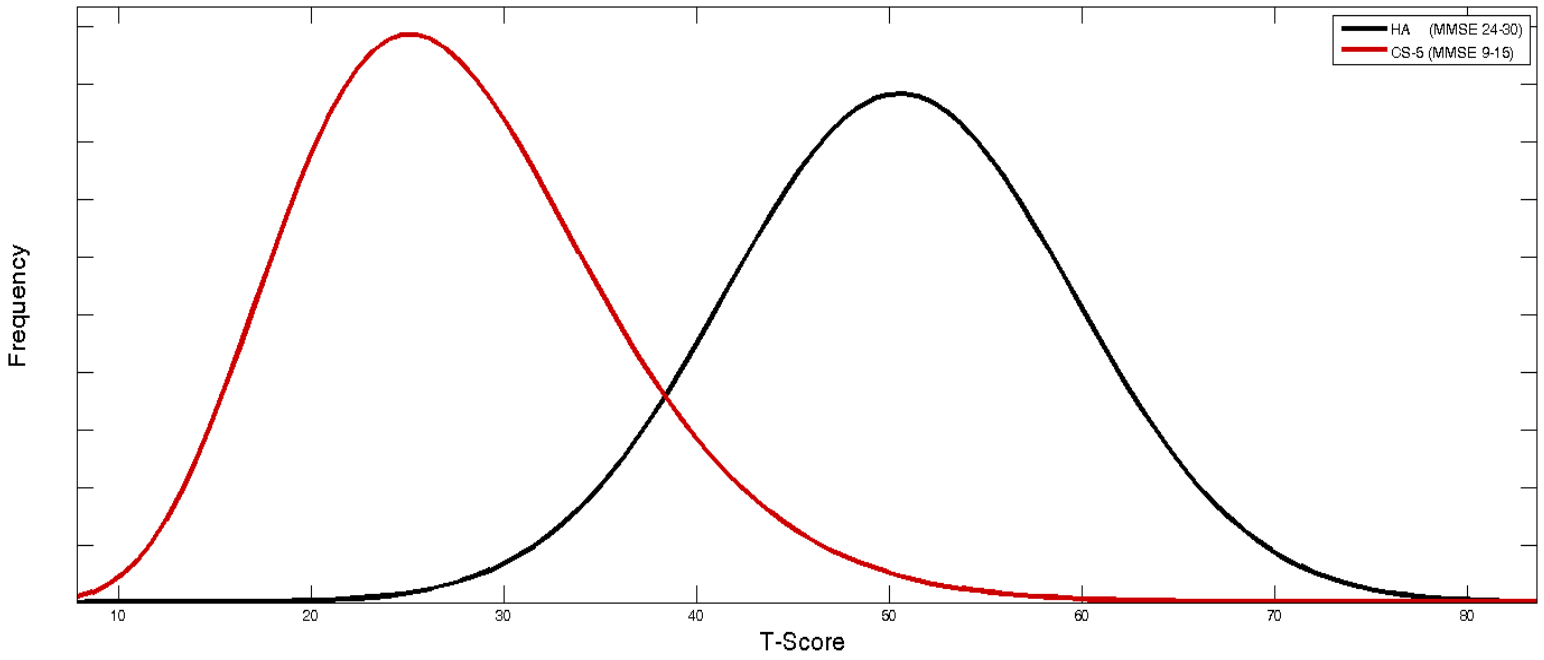
Comparison of within-person test score distributions produced by healthy adults (black line) and patients with MMSE scores of 20–23 (pink line)



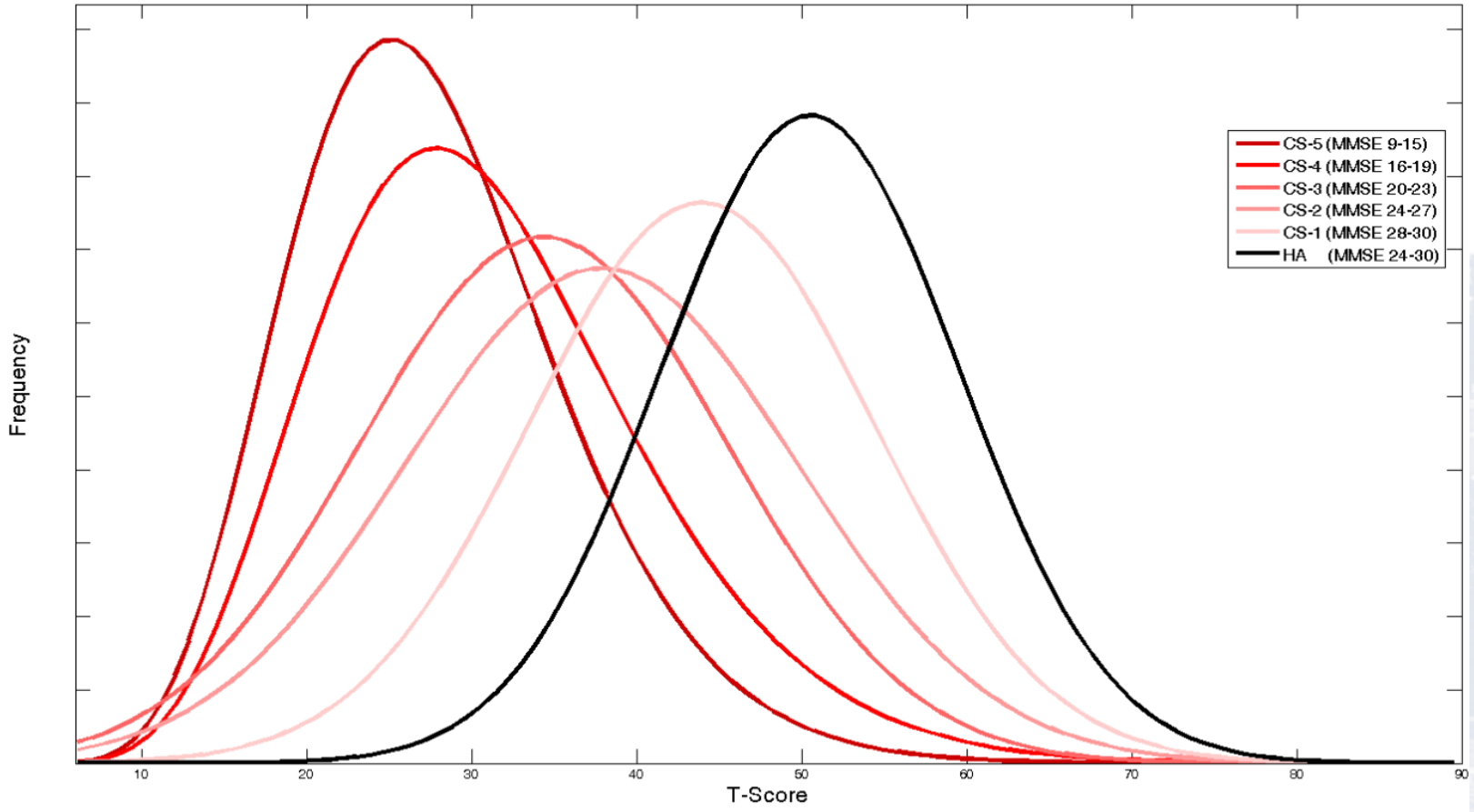
Comparison of within-person test score distributions produced by healthy adults (black line) and patients with MMSE scores of 16–19 (dark pink line)



Comparison of within-person test score distributions produced by healthy adults (black line) and patients with MMSE scores of 9–15 (red line)



Comparison of within-person test score distributions produced by healthy adults (black line) and patients with various MMSE scores (red/pink)



JOHNS HOPKINS MEDICINE

Conclusions

- The distribution of T-scores for healthy adults is, on average, normal
- Cross-sectional results support hypotheses
- This could represent a 4th method of clinical inference