Clinical Inference in the Assessment of Mental Residual Functional Capacity

David J Schretlen, PhD, ABPP

OIDAP Panel Meeting

10 June 2009

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)

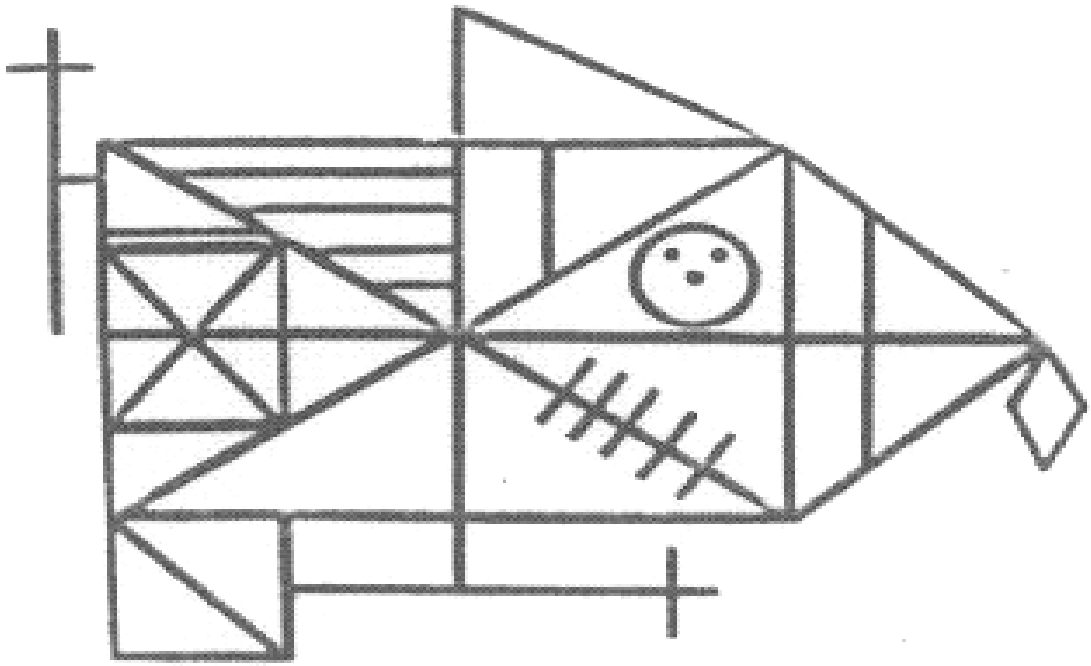


Fig. 4.8 The Complex Figure of Rey (Rey, 1959). Courtesy of Les Editions du Centre de Psychologie Appliquee.

Pathognomonic?

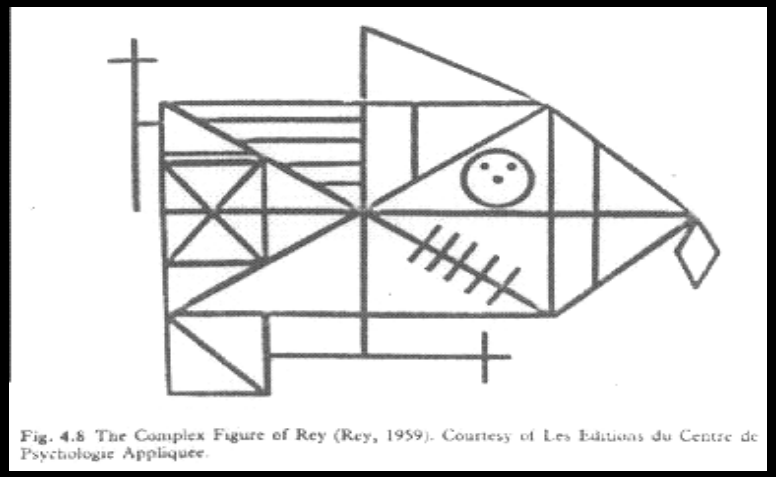


Fig. 4.8 The Complex Figure of Rey (Rey, 1959). Courtesy of Les Editions du Centre de Psychologie Appliquee.

91-year-old Caucasian woman

14 years of educ (AA degree)

Excellent health

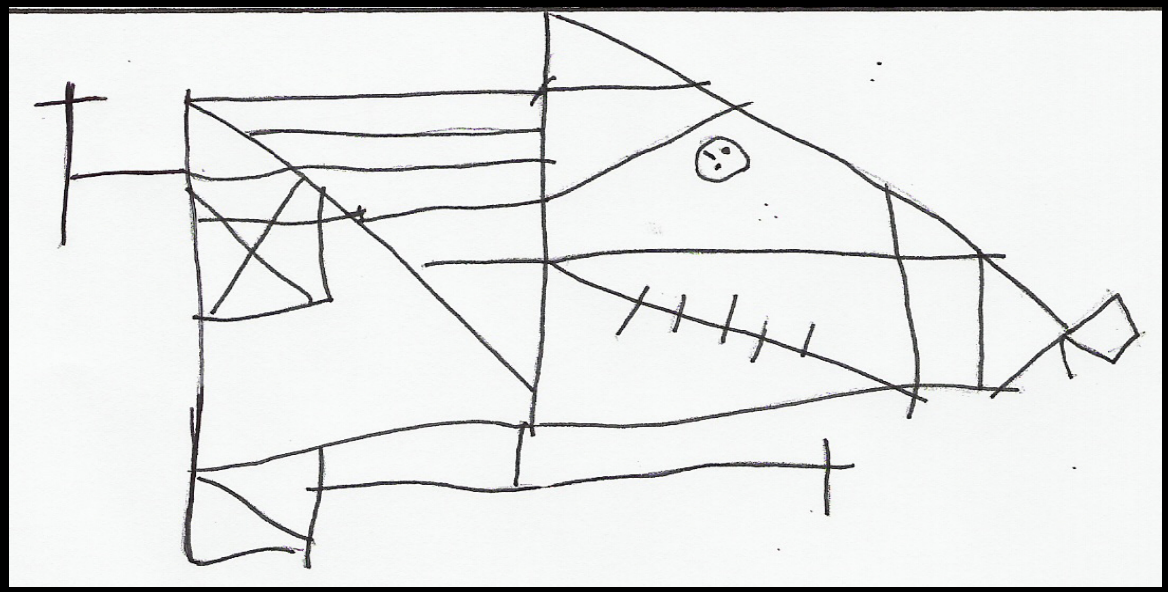
Rx: Floxin, vitamins

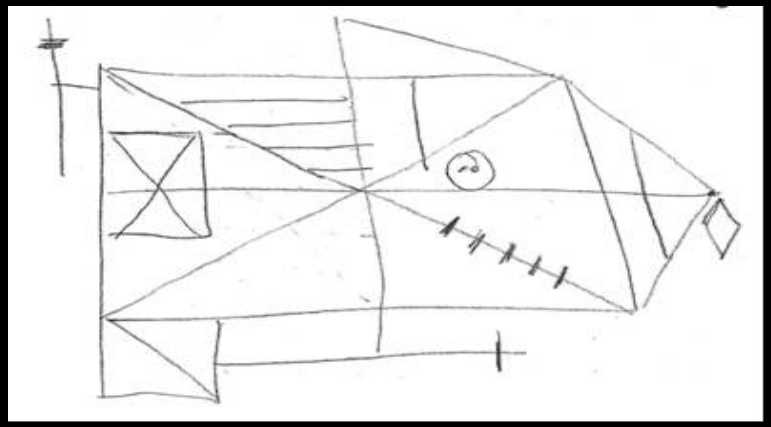
MMSE = 27/30

WAIS-R MOANS IQ = 109

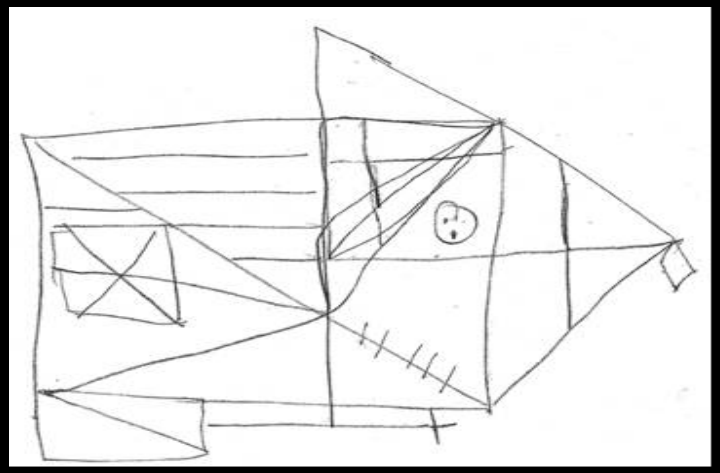
Benton FRT = 22/27

WMS-R VR Immed. SS = 8

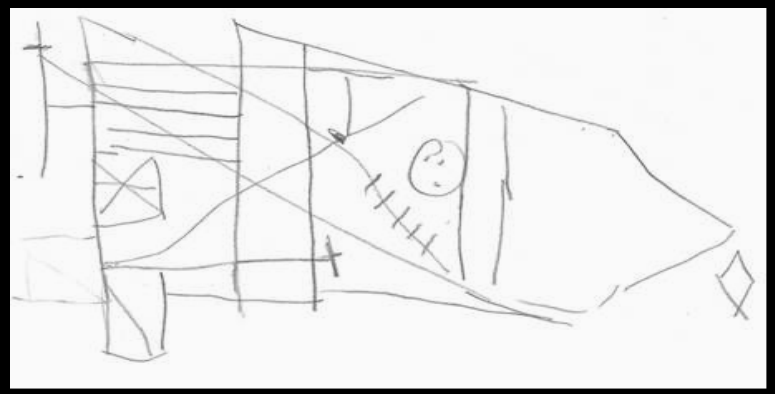




Jan. 2004: 68-year-old retired engineer with reduced arm swing, bradyphrenia & stooped posture. Diagnosed with atypical PD.



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
- ⇨ If we recommend that SSA rely on pathognomonic signs of impairment, we should not assume that successful job incumbents are free of such signs

Methods of Inference

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

Pattern Analysis

- Recognizable gestalt of signs, symptoms, history, laboratory findings, and test results
- Most elaborate approach to inference
- Best for patients with typical presentations

Empirical Basis of Pattern Analysis

- Considerable empirical support
  - But much of it is pieced together from disparate studies
- Studies often involve discriminant function analyses
  - Other designs have been used (eg, comparing AD and HD patients on MMSE after matching for total score)

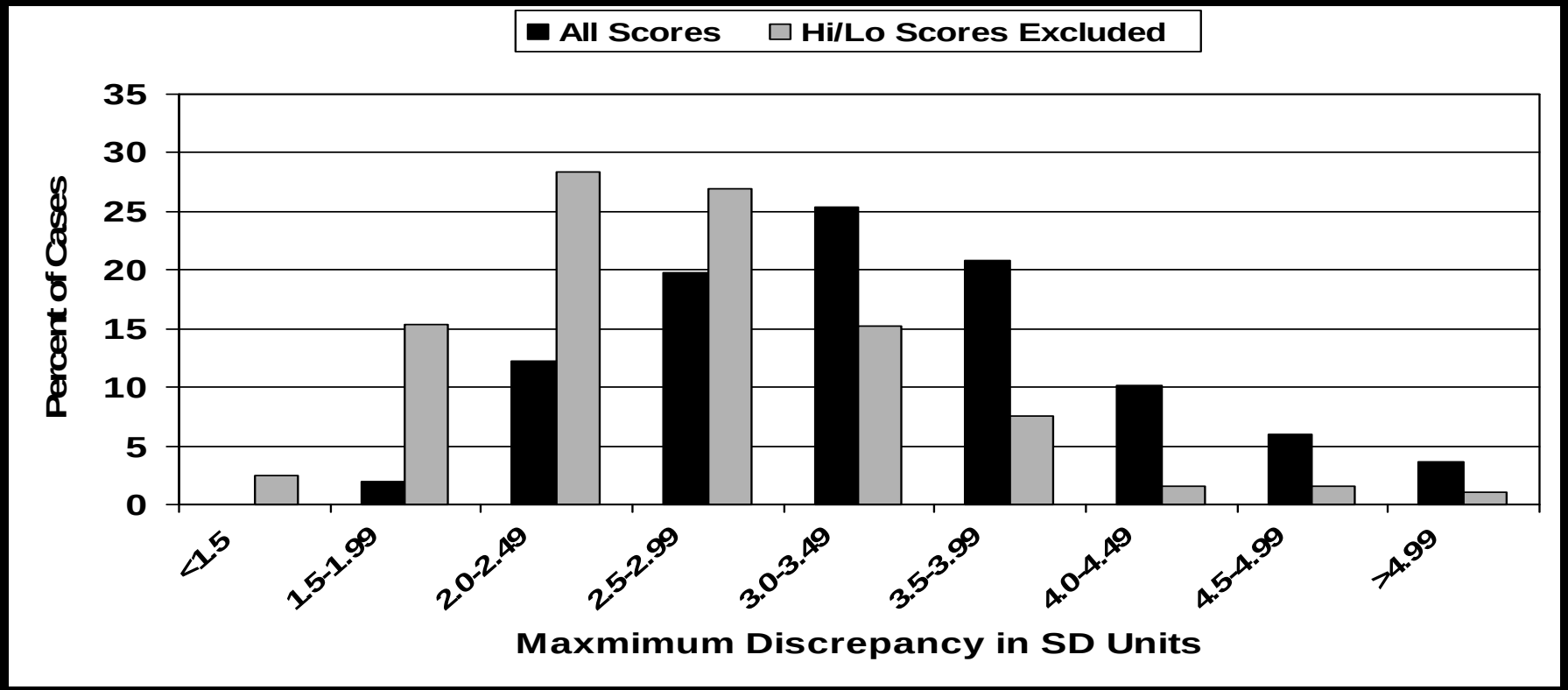
Journal of the International Neuropsychological Society (2003), **9**, 864–870.

DOI: 10.1017/S1355617703960061

Examining the range of normal intraindividual variability in neuropsychological test performance

- 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 ≥3.0; 20% had MDs ≥4.0
- Eliminating each persons' single highest and lowest test scores decreased their MDs, but 27% still produced MS values of 3.0 or greater

Intra-individual variability shown by 197 healthy adults



Pattern Analysis: Limitations & Implications

- Applicability varies with typicality of patient
- Normal variation can be mistaken for meaningful patterns
- ⇨This approach probably mirrors the task of linking specific residual functional capacities to job demands more closely than the others
- ⇨ It might be useful to think about linking specific RFCs to job demands using such statistical methods as cluster analysis or canonical correlation

Methods of Inference

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

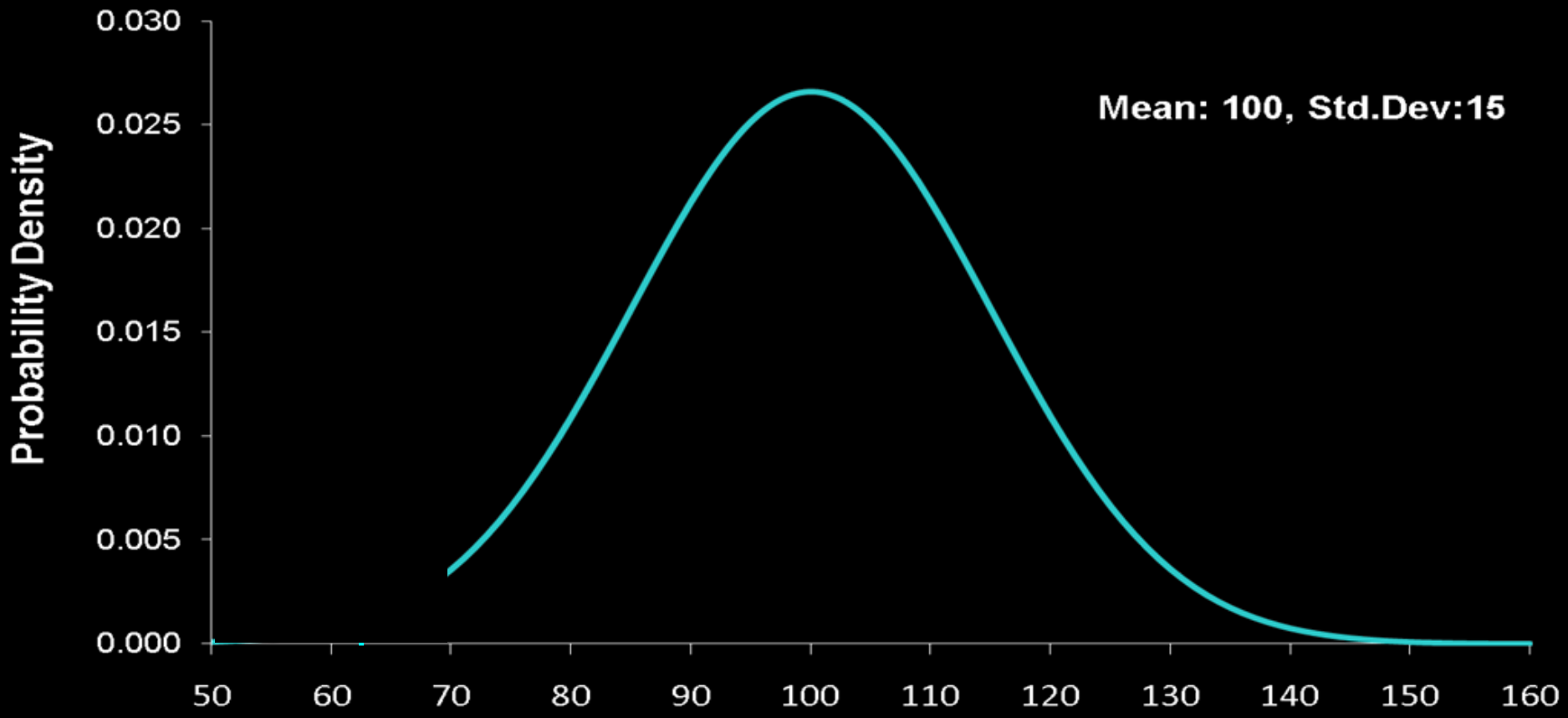
Level of Performance

- 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

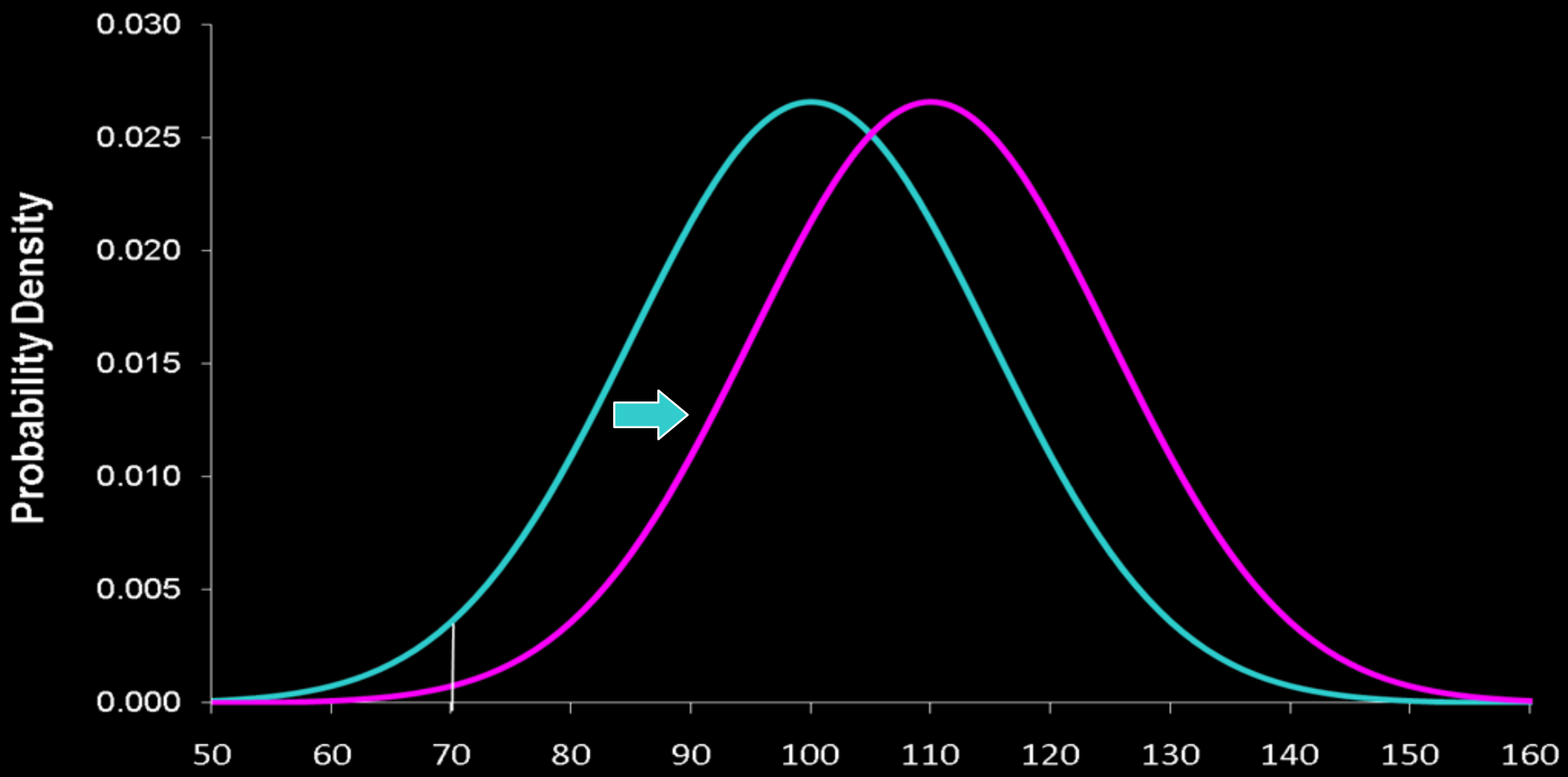
- We infer *ability* from *performance*
  - But factors other than disease (eg, effort) can uncouple them
  - There is no one-to-one relationship between brain dysfunction and abnormal test performance *at any level*
- But even if other factors do not uncouple them, what is an *abnormal* level of performance?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

Would the distribution look like this?



Probably not

More likely, the distribution would be shifted up



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

The binomial distribution can be used to predict how many abnormal scores healthy persons will produce on batteries of various lengths

Probability of obtaining 2 or more "impaired" scores based on selected cut-off criteria & number of tests administered

	Number of Tests Administered		
Cut-off	**10**	**20**	**30**
--1.0 *SD*	.50	.84	.95
--1.5 *SD*	.14	.40	.61
--2.0 *SD*	.03	.08	.16

Ingraham & Aiken (1996)

Journal of the International Neuropsychological Society (2008), **14**, 436–445.

doi: 10.1017/S1355617708080387

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

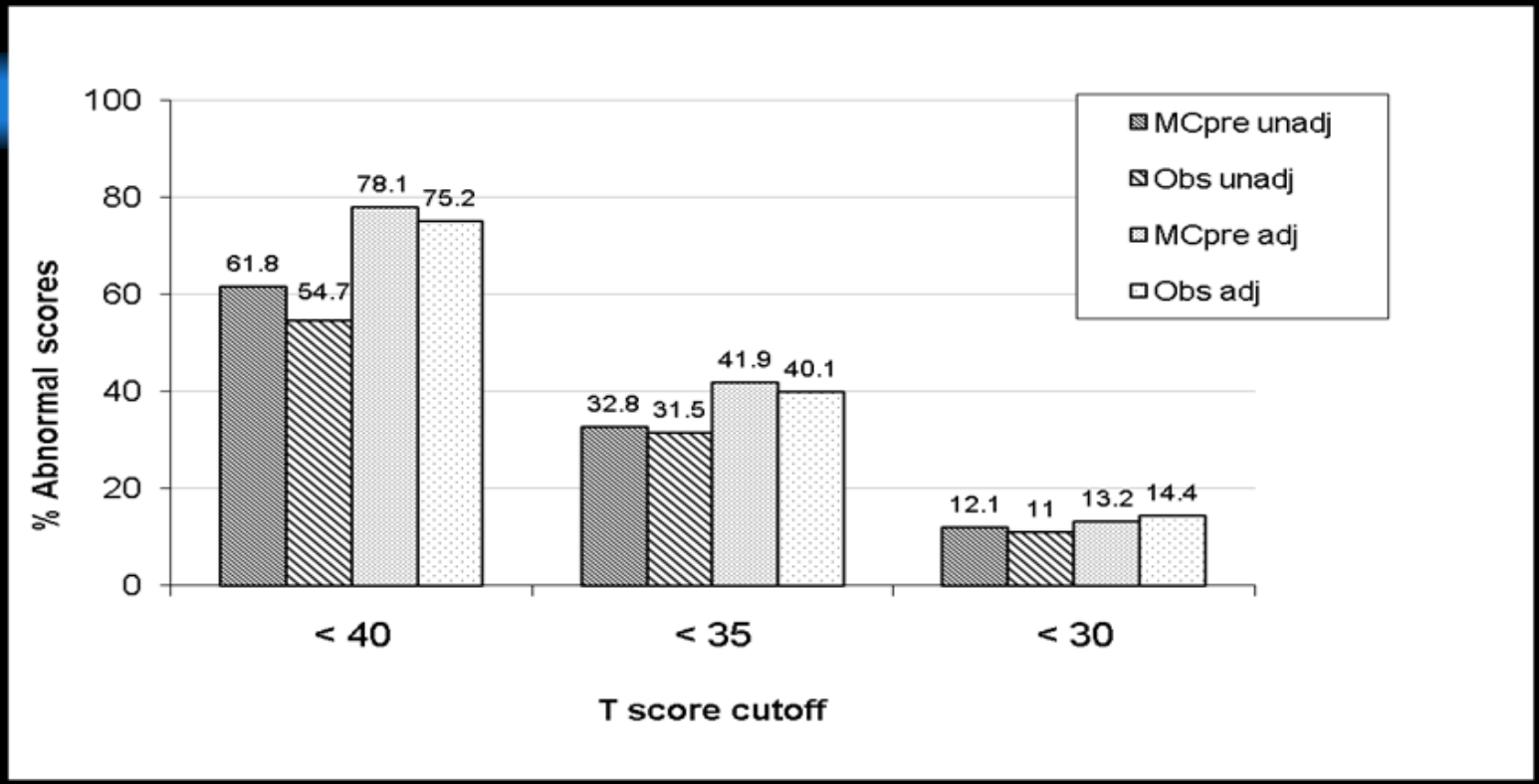
- Participants
  - 327 reasonably healthy adults without current psychiatric illness or substance abuse/dependence
- Procedure
  - Administered 25 cognitive measures; obtained T-scores
  - Classified T-scores as normal or “abnormal” based on three cutoffs: <40, <35, and <30
  - Computed Cognitive Impairment Indices (CII) as the number of abnormal scores each person produced
  - Used both unadjusted and demographically adjusted scores

- 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-dom 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)

Spearman correlations between Cog Imp Index scores based on unadjusted T-scores and age, sex, race, years of 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**

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

This study shows that

- 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

Returning to the question of what cut-off we should use to define abnormal performance…

- Stringent cut-offs decrease test sensitivity
- Liberal cut-offs decrease test specificity
- Adding tests increases the risk of type I errors
- Excluding tests increases the risk of type II error
- As in most endeavors, we must exercise judgment

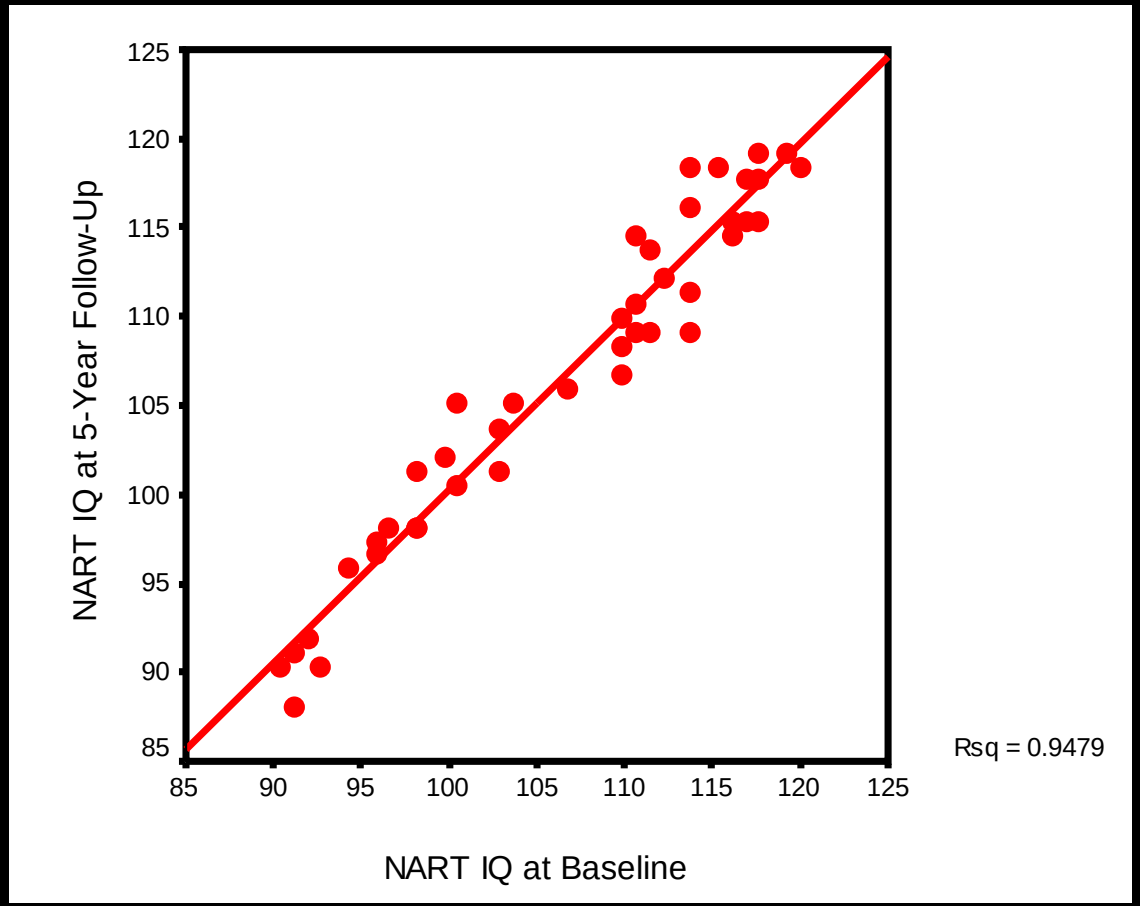
Decline from Premorbid Ability

- 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

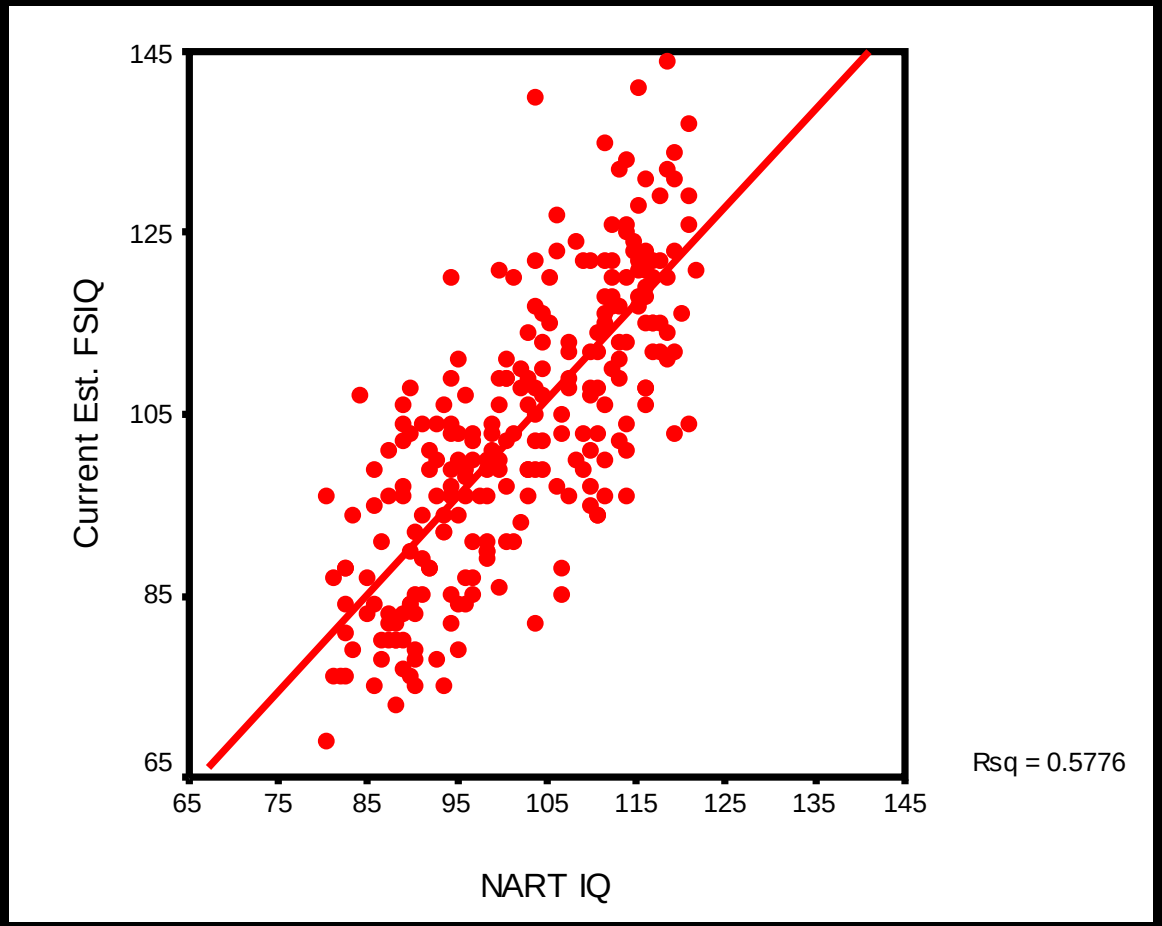
Estimating Premorbid IQ

- 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

Stability of NART-R IQ Estimates



Correlation of NART-R and WAIS-R



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

Journal of the International Neuropsychological Society (2005), **11**, 784–787.

DOI: 10.1017/S1355617705050939

BRIEF COMMUNICATION

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

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

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

Journal of the International Neuropsychological Society (2004), **10**, 82–90.

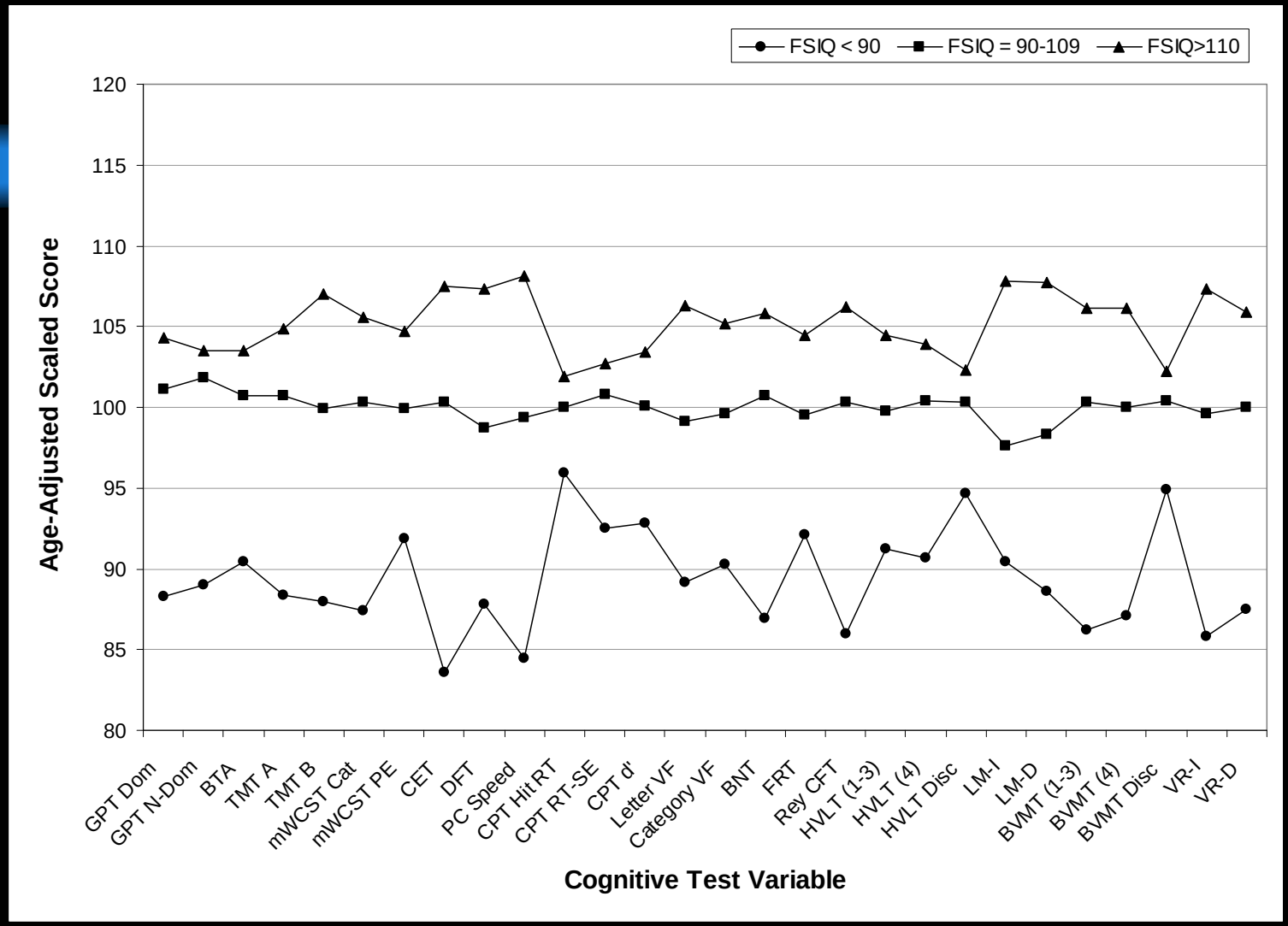
DOI: 10.1017/S1355617704101100

How well does IQ predict neuropsychological test performance in normal adults?

Examined 28 scores derived from 16 cognitive tests that were administered to 221 reasonably healthy adults

Grouped participants by WAIS-R Full Scale IQ into three groups:

N = 37	Below average (BA)	FSIQ < 90	Mean = 83
N =106	Average (A)	FSIQ 90-109	Mean = 101
N = 78	Above average (AA)	FSIQ > 109	Mean = 121


FSIQ < 90
FSIQ = 90-109
FSIQ>110
120
115
110
105
100
95
90
85
80
Age-Adjusted Scaled Score
GPT Dom
GPT N-Dom
BTA
TMT A
TMT B
mWCST Cat
mWCST PE
CET
DFT
PC Speed
CPT Hit RT
CPT RT-SE
CPT d'
Letter VF
Category VF
BNT
FRT
Rey CFT
HVLT (1-3)
HVLT (4)
HVLT Disc
LM-I
LM-D
BVMT (1-3)
BVMT (4)
BVMT Disc
VR-I
VR-D
Cognitive Test Variable

Intelligence and Cognitive Functioning

- Correlations between intelligence and other cognitive abilities are stronger below than above IQ scores of 110
  - *It is less likely that smart people will do well on other tests than it is that dull people will do poorly*
- A normal person with an IQ of 85 is likely to produce “impaired” scores on about 10% of other cognitive tests

Deficit Measurement: Limitations & 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
- NART predicts pre-morbid IQ better than other abilities
- Raising “cut-off” scores for patients of above average IQ can compound the problem of multiple comparisons

Deficit Measurement: Limitations & Implications

- ➡ Many – if not most – successful job incumbents likely fall short of meeting one or more of their job demands
- ➡ What cutoff in the distribution of an ability shown by successful job incumbents should we use to define sufficient RFC for someone to do that job? This will directly affect the percentage of applicants who will be found disabled
- ➡ Factors other than impairment, like effort, can uncouple the linkage between performance and ability
- ➡ Work demands, RFC, and “deficit” vs. “impairment”