Neuropsychological Tests of the Future: How Do We Get There from Here?

Robert M. Bilder, PhD, ABPP-CN

Michael E. Tennenbaum Family Chair in Creativity Research, Chief of Medical Psychology – Neuropsychology, and Distinguished Professor of Psychiatry & Biobehavioral Sciences, Jane & Terry Semel Institute for Neuroscience and Human Behavior, Stewart & Lynda Resnick Neuropsychiatric Hospital, and David Geffen School of Medicine and Psychology, and College of Letters & Science at UCLA

UCLA Semel Institute
Tennenbaum Center for the Biology of Creativity

Journal of the International Neuropsychological Society (2011), **17**, 7–13.

doi:10.1017/S1355617710001396

SHORT REVIEW

Neuropsychology 3.0: Evidence-Based Science and Practice

Robert M. Bilder[1,2,3]

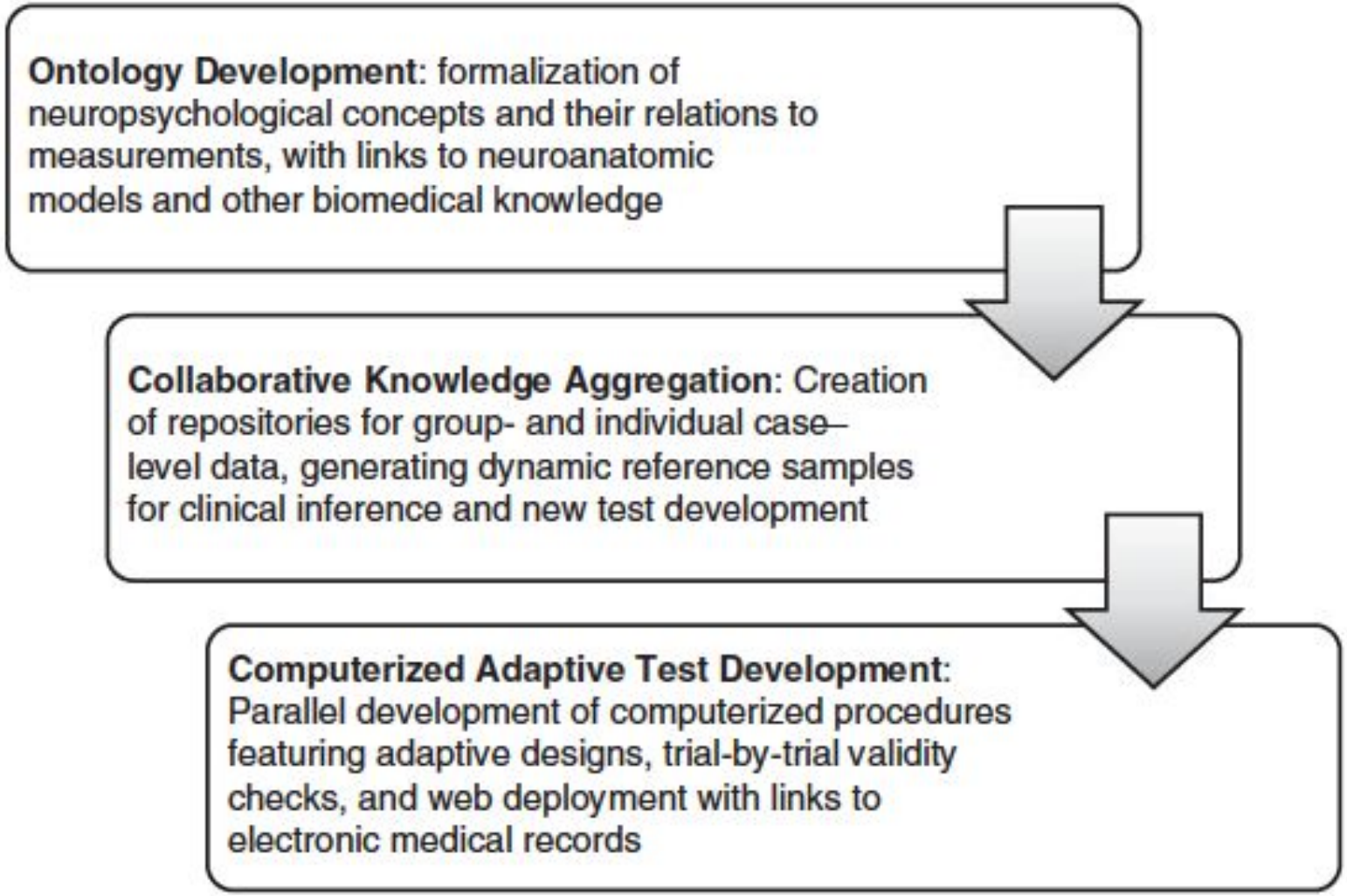
[1]Jane and Terry Semel Institute for Neuroscience & Human Behavior at UCLA, Los Angeles, California
[2]Department of Psychiatry & Biobehavioral Sciences, David Geffen School of Medicine at UCLA, Los Angeles, California
[3]Department of Psychology, UCLA College of Letters & Science, Los Angeles, California

(Received August 13, 2010; Final Revision October 16, 2010; Accepted October 18, 2010)

The Future of NP Assessment?

- Scalable mobile assessments
  - Measure attention to visual, auditory, tactile stimuli; use augmented reality
  - Capture response times, GPS or gyro-captured motion in real world or test space
  - Use other peripheral devices to capture motion or physiological signals, HRV, more…
- IOT (internet of things)
  - Brain sensitive home – measure adl, memory, processing speed, sleep quality, diet
  - Brain sensitive car – measure sensorimotor control, stop signal, harzard avoidance
- Consider all possibilities for: acquisition, analysis, and interpretation of data

Path to Development of Novel NP Paradigms

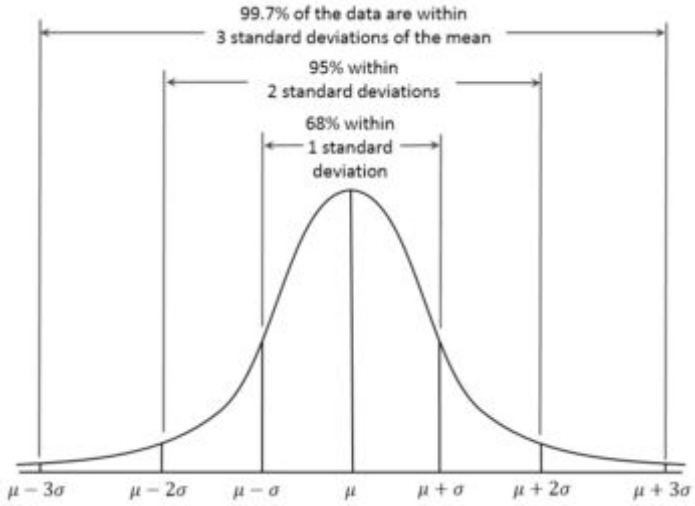


NP 1.0 (1950-1979)

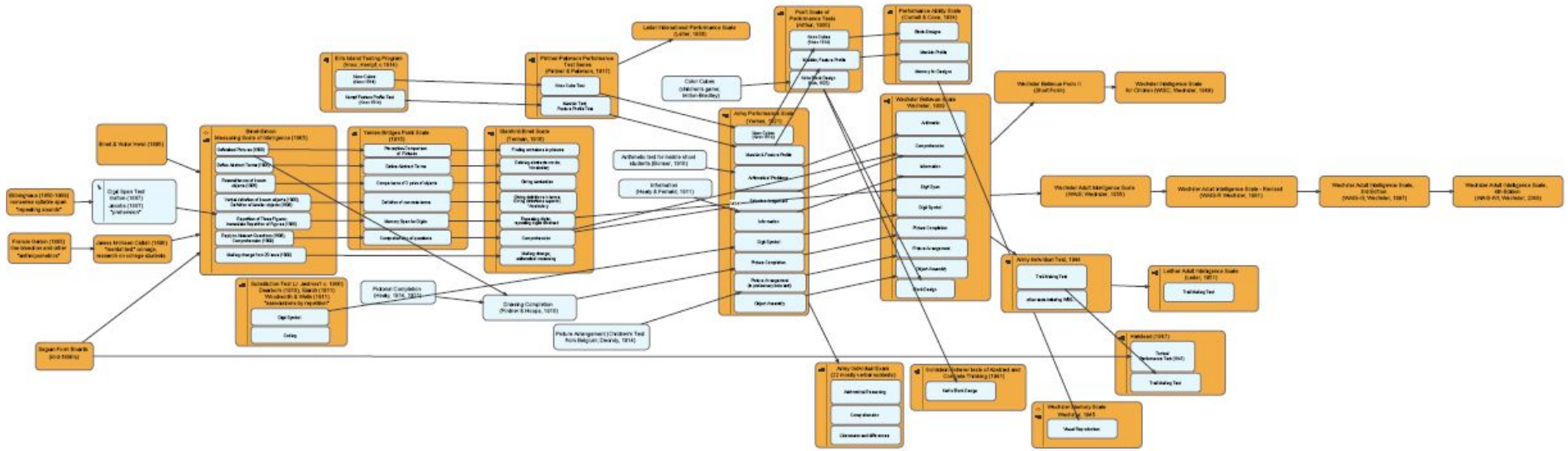
- Term: “Neuropsychology” (Klove, 1963); Founding of INS (1967)
- Effects of discrete brain lesions, use of the “sign” approach and syndrome analysis:
  - Luria
  - Halstead & Reitan
  - Benton and the Iowa school
  - Teuber and the NYC school
  - Kaplan and the Boston school

NP 2.0 (1980-present)

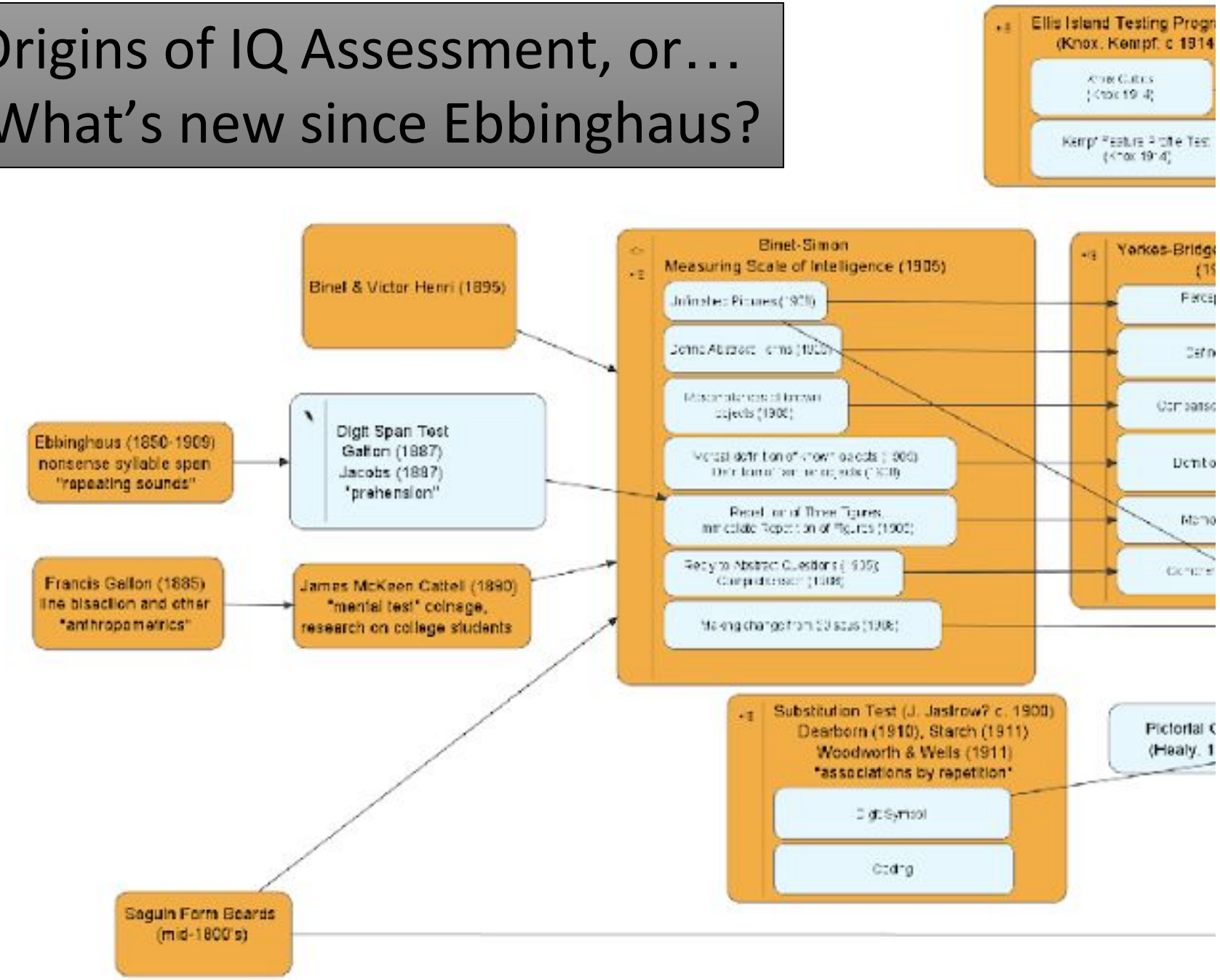
- Neuroimaging displaces localization role (mostly)
- Late 70's – training programs
- Board certification (ABPP) 1981
- Houston Conference (1998)
- "Era of classical psychometrics"
- Characterization of strengths and weaknesses: "Profile" analysis



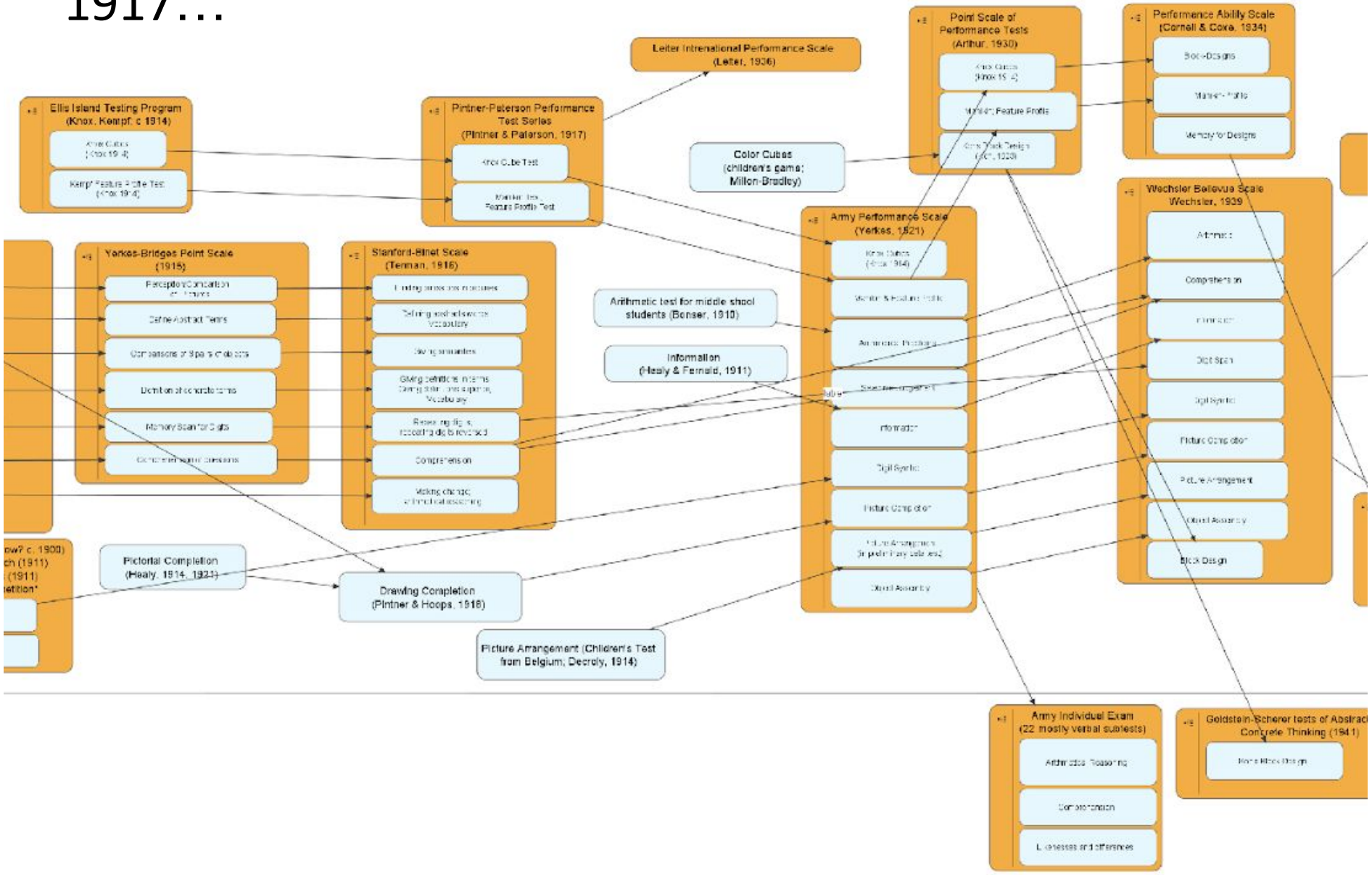
A brief history of intelligence testing!



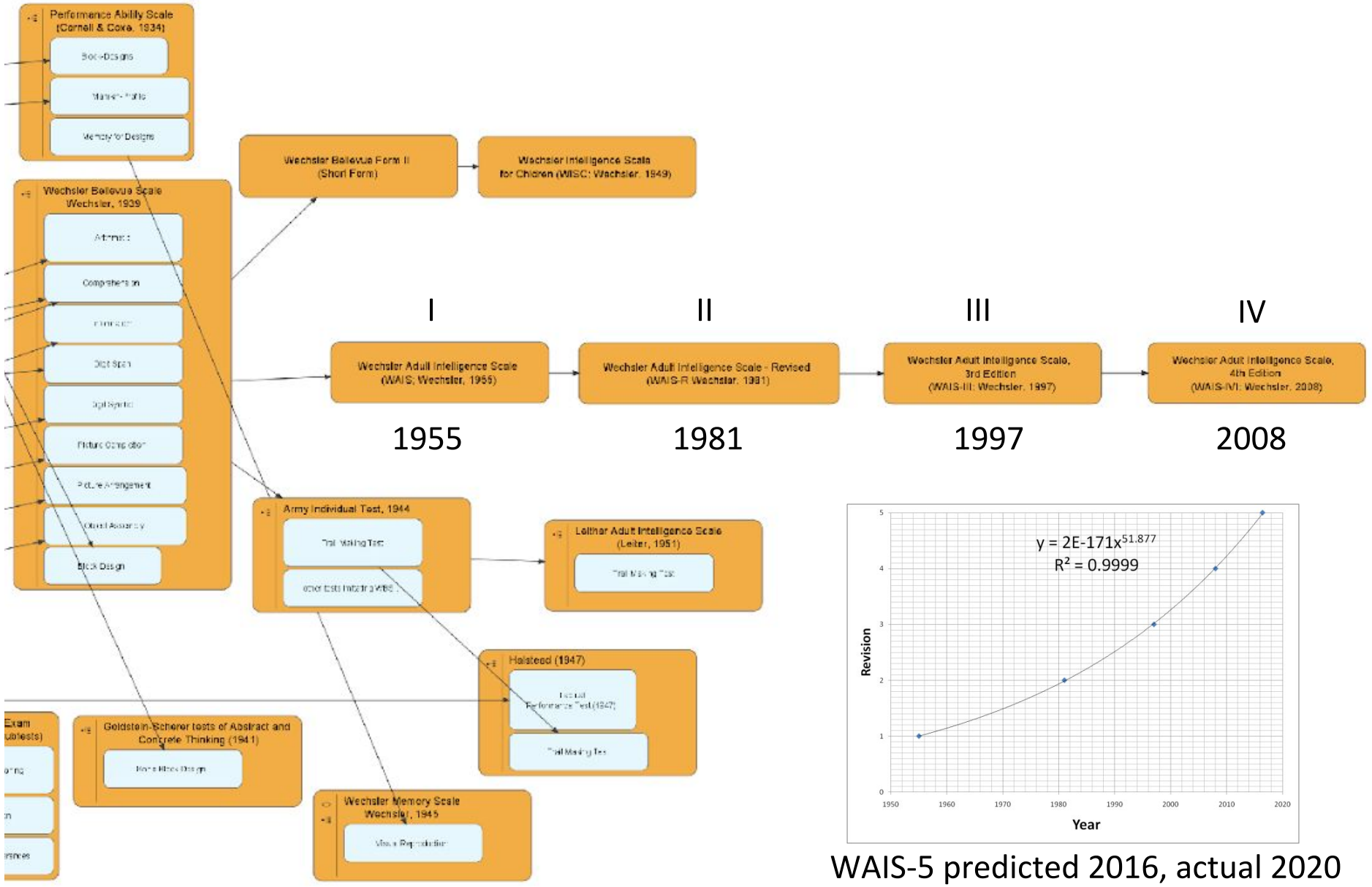
Origins of IQ Assessment, or… What's new since Ebbinghaus?



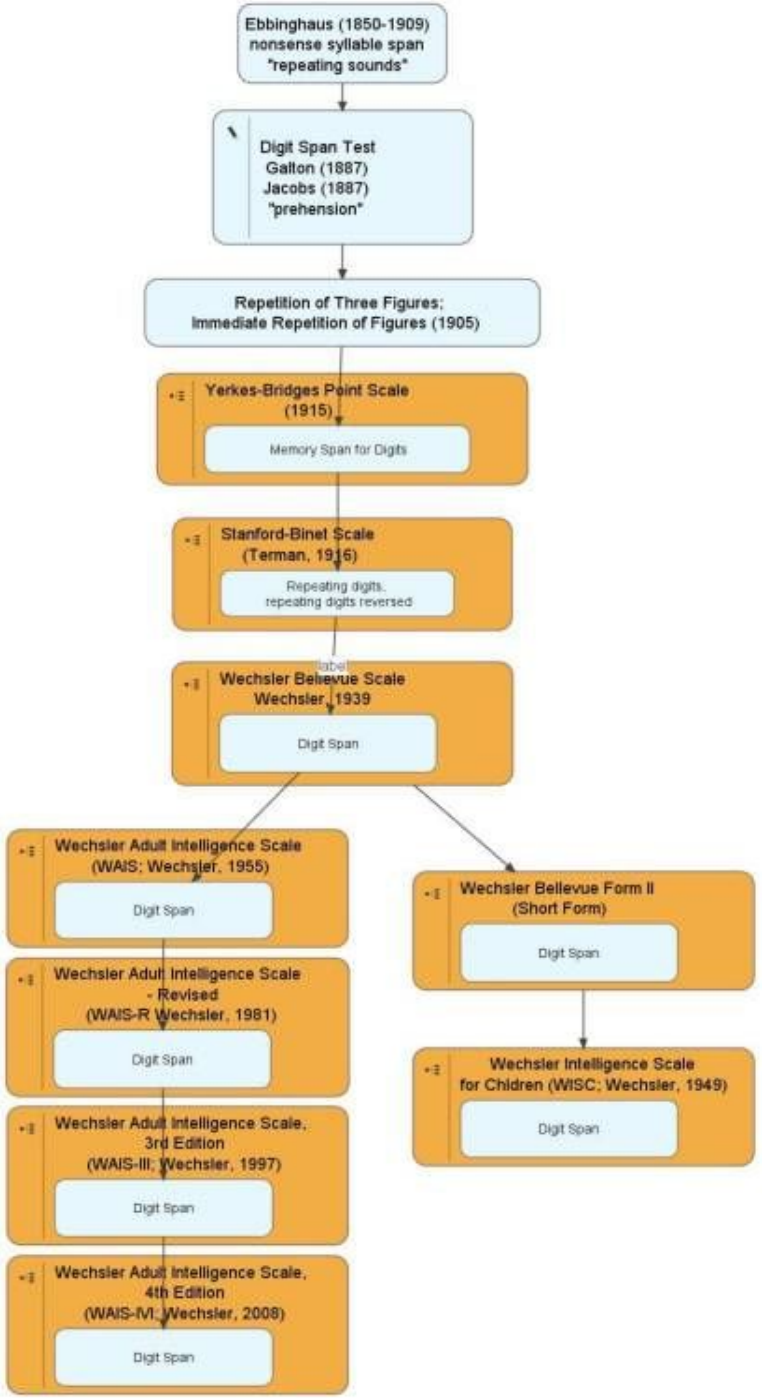
… pretty much settled by 1917…



Sternberg, R. J. (1993). Rocky's back—again: A review of the WISC–III. *Journal of Psychoeducational Assessment* (Monograph). 161–164.



WAIS-5 predicted 2016, actual 2020 or 2021?



Tracking a Task Over Time

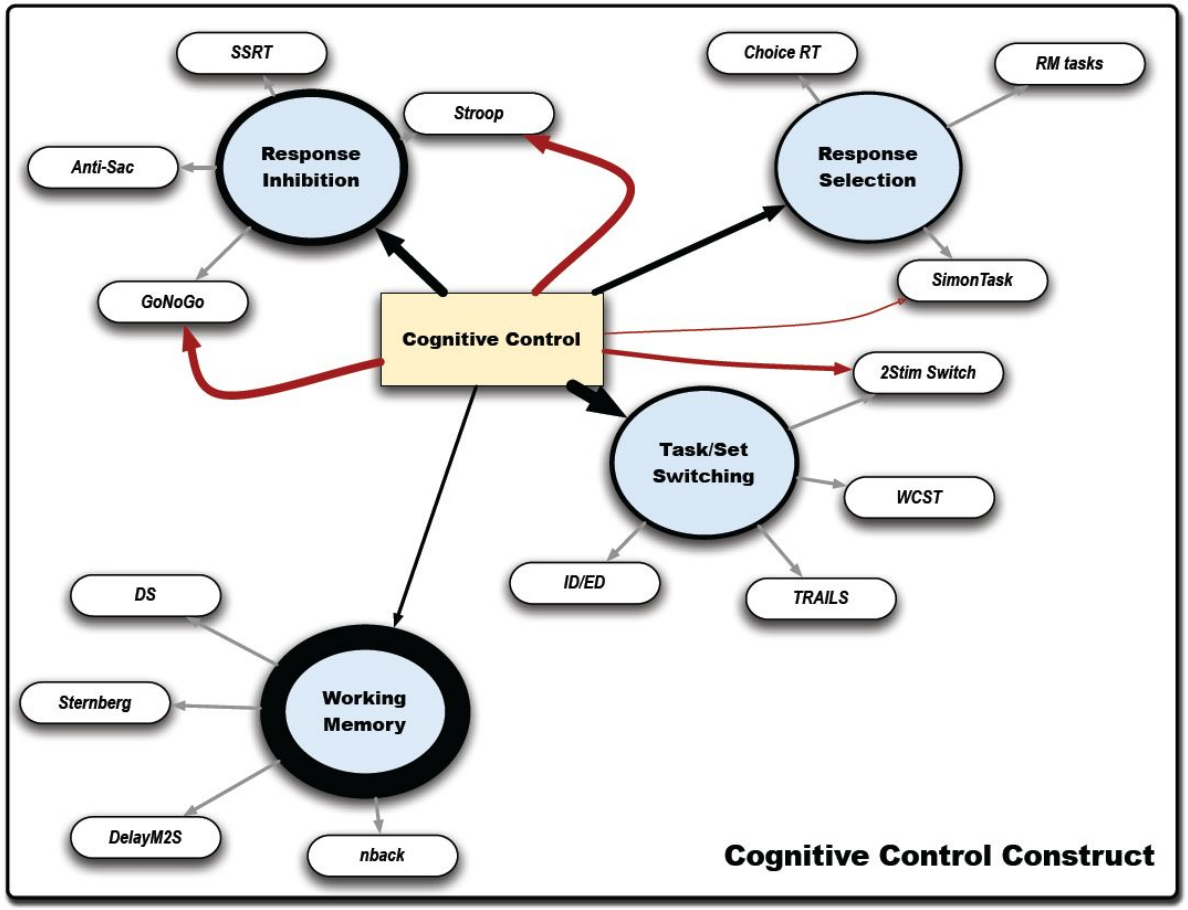
Digit Span - *constructs*

1870's-1880's: *prehension*
1905: *repetition*
1915: *memory*
Subsequently ….

- *Short Term Memory*
- *Attention*
- *Freedom from Distractibility*
- *Working Memory*

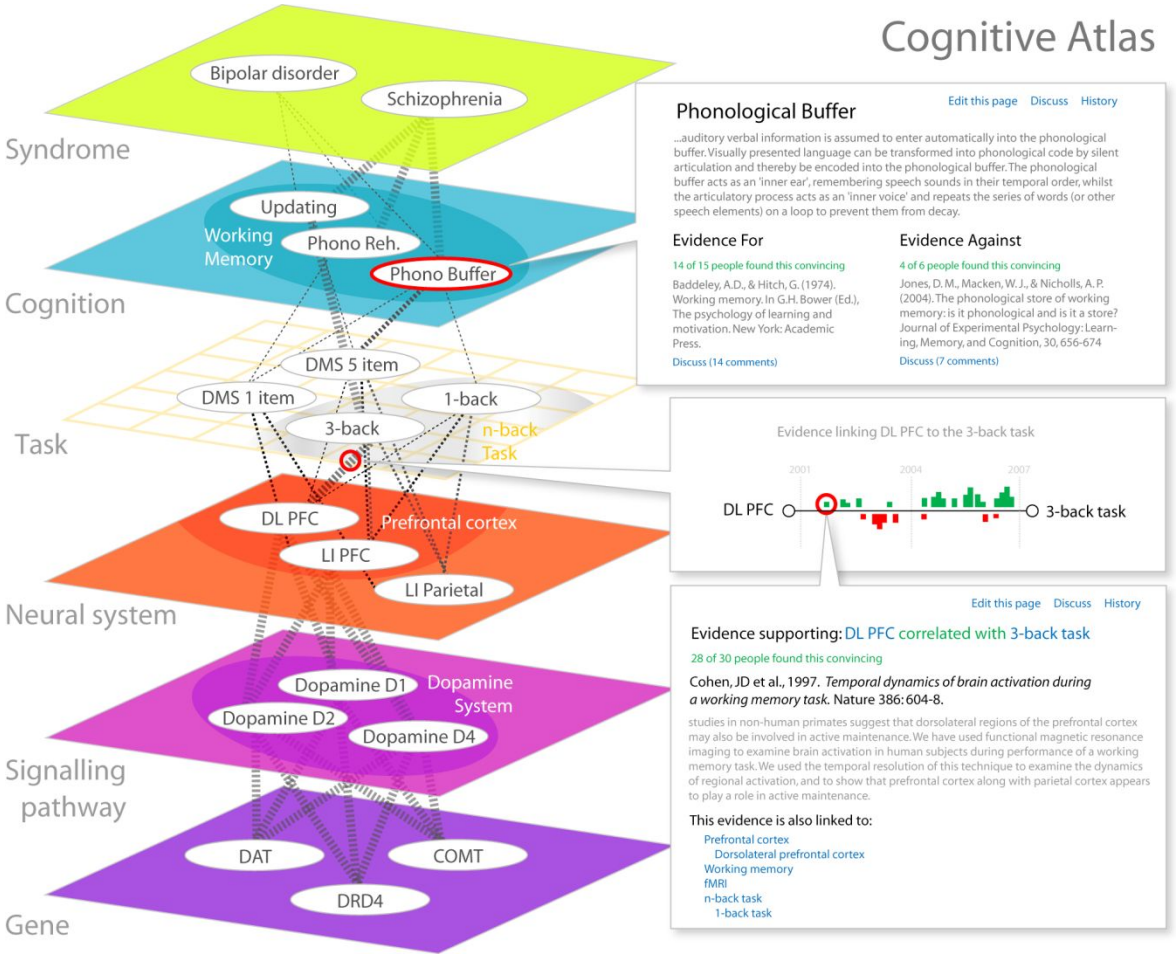
Concept definition via multiple sub-constructs and test indicators

"Cognitive control" has been associated in PubMed literature with RI, RS, TSS and WM. Sampling all 5 concepts (x30 papers) identified the task indicators used to assess these concepts. Note CC itself was measured using the same task indicators as RI and TSS.



Sabb FW, Bearden CE, Glahn DC, Freimer N, Parker DS, Bilder RM. *Molecular Psychiatry*, 2009

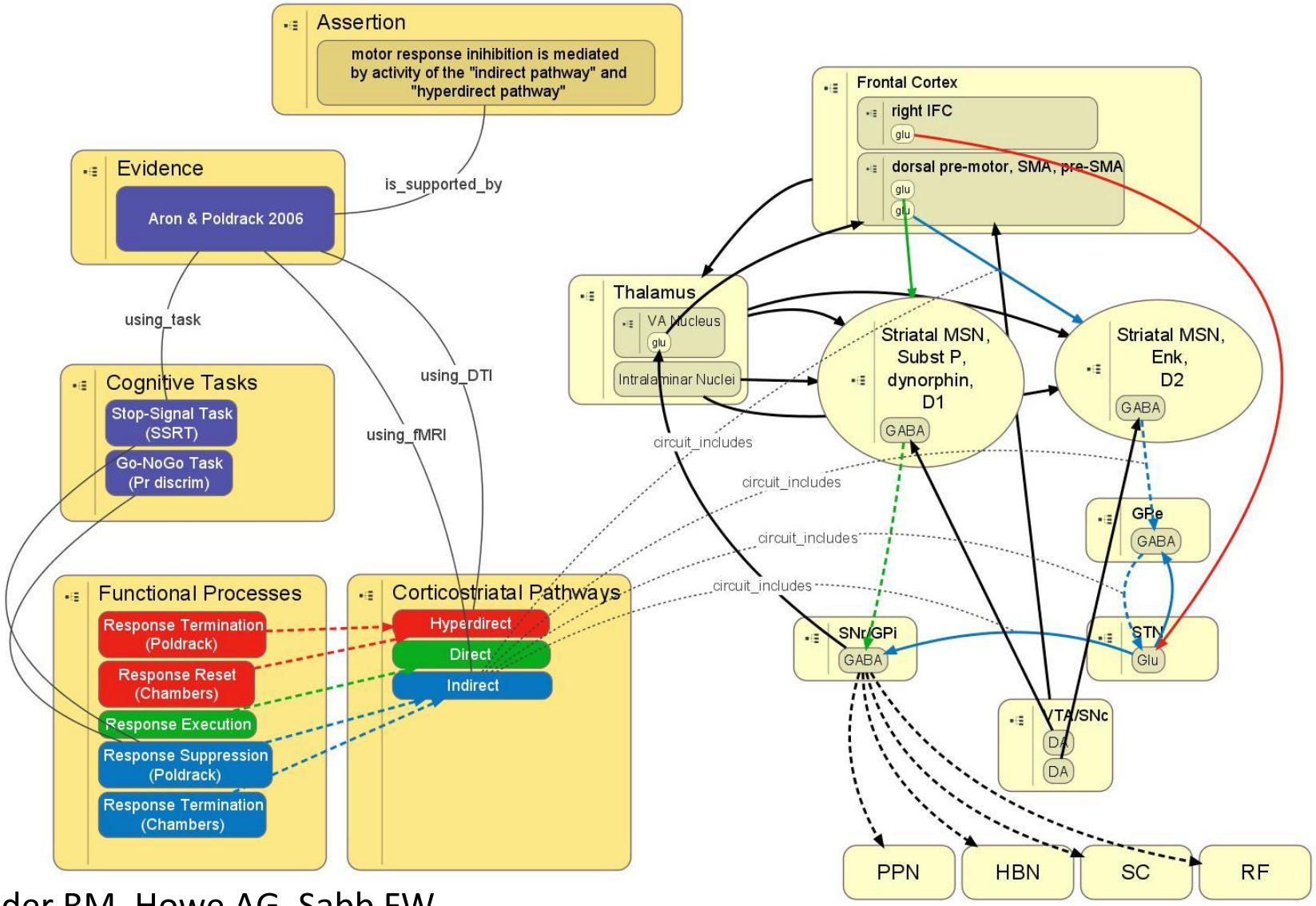
Architectures for cognitive ontology development



The Cognitive Atlas is conceptualized as a related set of maps. A given map may contain sets of related concepts, quantitative models of literature association, annotated effect size statistics, raw data, summaries of voting, and qualitative free-text inputs.

For cognitive concepts (e.g., the "phonological buffer") there are associated cognitive concepts, and a "test" layer comprising objective indicators of the concepts

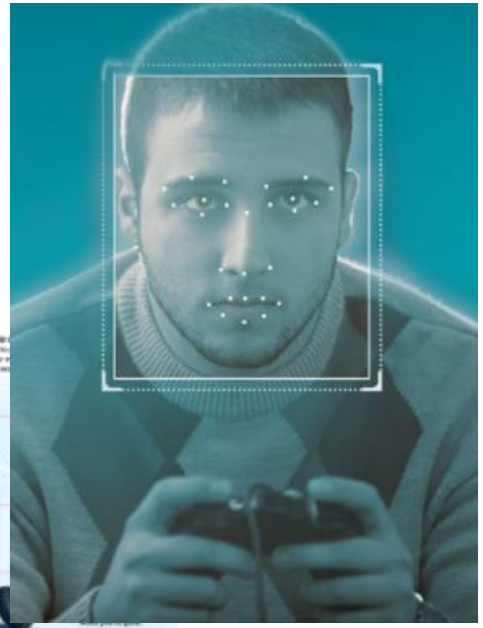
Schema for Validation of Hypotheses about Neurocognitive Concepts



Bilder RM, Howe AG, Sabb FW

Journal of Abnormal Psychology, 2013 Aug;122(3):917-27.


THE SMART INSIDE
Smoke detector
Smart bulbs
Movie night
Digital discipline
Washing machine
Intelligent fridge



Assessment

- Scalable mobile assessments
  - Measure attention to visual, auditory, tactile stimuli; use augmented reality
  - Capture response times, GPS or gyro-captured motion in real world or test space
  - Use other peripheral devices to capture motion or physiological signals, HRV, more…
- IOT (internet of things)
  - Brain sensitive home – measure adl, memory, processing speed, sleep quality, diet
  - Brain sensitive car – measure sensorimotor control, stop signal, harzard avoidance
- Consider all possibilities for: acquisition, analysis, and interpretation of data

Intervention

- Novel brain training tools – real world interface (supermarket “go”)
- Social monitoring and nudging
- Contemplative practice integration
- Meta-cognitive therapy tools
- Integrated neurofeedback, neurostimulation
- Affect monitoring and feedback
- Stress monitoring and feedback
- Any idea capable of transforming and increasing efficacy of brain-oriented interventions

THE CLINICAL NEUROPSYCHOLOGIST
2019, VOL. 33, NO. 2, 220–245
https://doi.org/10.1080/13854046.2018.1521993

Routledge
Taylor & Francis Group

Neuropsychological tests of the future: How do we get there from here?

Robert M. Bilder[a,b] and Steven P. Reise[b]

[a]Departments of Psychiatry & Biobehavioral Science, Jane & Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California, USA; [b]Department of Psychiatry & Biobehavioral Science, Los Angeles, California, USA

Table 1. Overview of potential methodological advances in neuropsychological (NP) assessment.

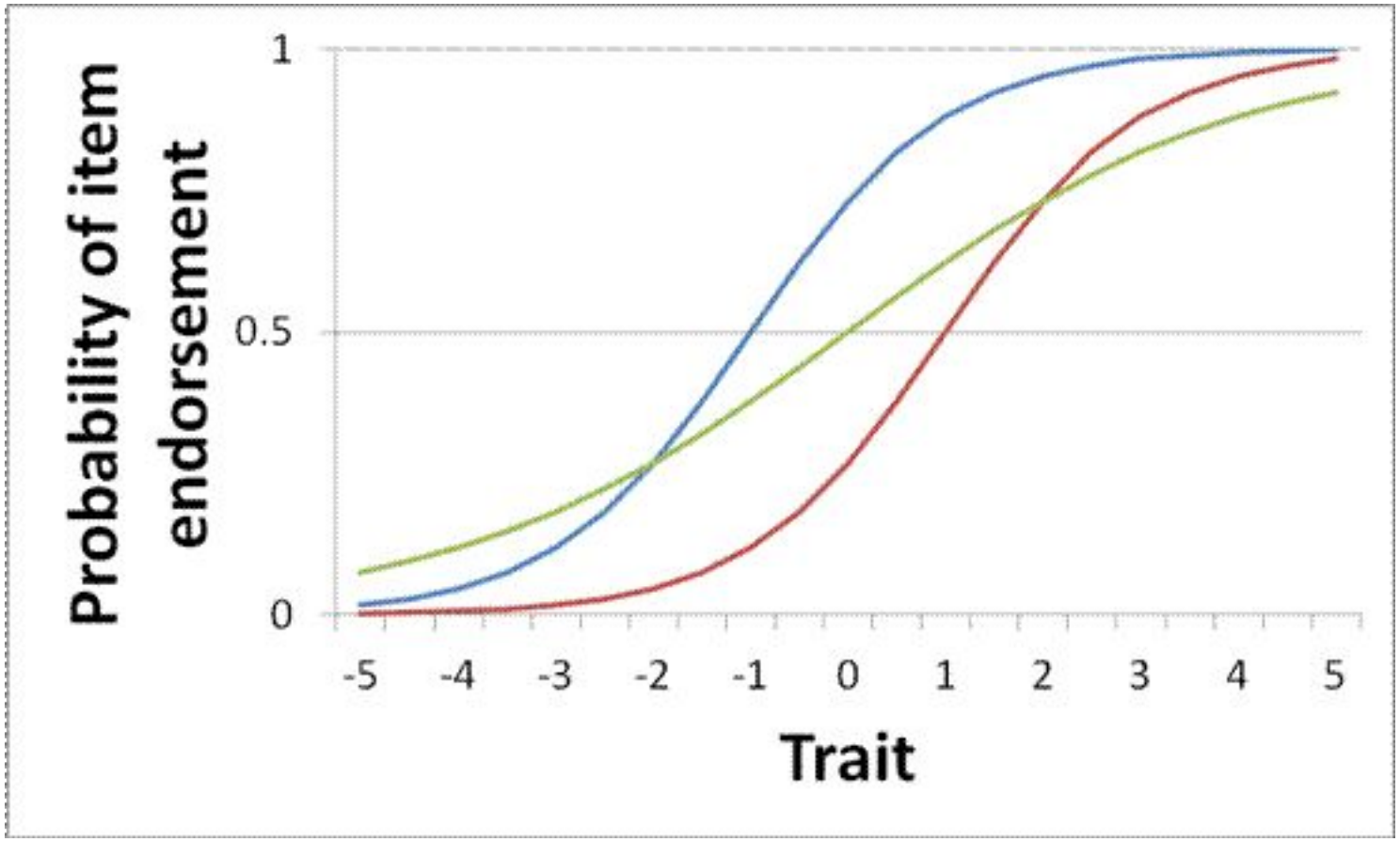
Method	Current	Future	Advantage
NP trait models	Unidimensional	Bifactor models, multidimensional IRT models (mIRT)	Each item can provide information about different traits; a single item or test can help specify both general factors and domain scores
Nominal response model	Different kinds of errors are treated identically	Each wrong response has a different meaning	Each item carries more information, enabling greater precision and/or assessing different constructs
Test linking	Total scores are compared in studies that use both tests	Item banks can be drawn from existing tests and new items, and all items calibrated together	Enables direct comparison of different tests and construction of new tests that are back-compatible with the originals
Computerized adaptive testing	Paper-pencil, fixed administration order, minimal branching	Information from each item response selection and speed used to select next most informative item	Efficiency gain of 50–95% in administration time or precision of measurement
Differential item functioning (DIF)	Effects of group (diagnostic, age, sexual, racial, ethnic, cultural, etc.) determined by comparing total scores	DIF examines group effects for each item	Increased precision in specifying diagnostic and other group differences that may not be apparent in the scores of the whole test
Person fit statistics	Performance validity based on 'cutoff' scores, mostly based on accuracy	Performance validity based on the fit of item response characteristics to the examinees overall estimated trait level	Performance validity can be examined within each test; every item response can be useful in detecting anomalies; increase sensitivity to intentional failure
Non-IRT Item-level strategies	Most emphasis on summary scores not trial-by-trial analysis	Focus on sequential dependence of responses and meaning of response sequences	Increased efficiency in identifying primary constructs; identification of qualitatively distinct response patterns
Evidence-based diagnostic batteries	Batteries with limited flexibility involve redundant testing	Test selection will proceed based on positive predictive power	Testing efficiently focuses time with respect to differential diagnostic questions or recommendations
Computerized testing	Print publishing model; paper-pencil data acquisition and scoring	Computerized tests for stimulus presentation and response acquisition	Precision in timing of stimulus presentation and response collection, automatic recording, scoring and database entry of responses, and automatic updating of software to new versions; acquisition of voice, video, motion
Web-based testing	Testing done in clinic or lab	Testing done at home or wherever convenient for examinee	Scalable assessment at lower cost
Healthcare informatics and bioinformatics	Test results go to file cabinets, report text goes on medical record	Data elements will be part of medical record and integrated with analytics relating these to other health variables	The NP data will be integrated into comprehensive model of patient; implications will be pushed to all care-team members and hypotheses fed back to NP clinicians for follow-up; 'big data' analytics will find new patterns to inform future evidence-based practice
Mobile platforms	Not used; not trusted	Passive monitoring will dramatically increase data flow; experience sampling will augment self reports	Marked increase in longitudinal repeated measures for self-reports and tests; new variables extracted from passive monitoring
Wearables	Not used; not trusted	Passive monitoring of diverse physiological, activity, and experiential data	Data previously available only in cross-sectional lab studies (sleep, EEG, cardiovascular) will be widely available and assessed longitudinally
Internet of Things (IOT)	Not used; not trusted	Passive monitoring of activities across multiple environments	Ecologically valid assessments will be done in real-world contexts; and environment can 'respond' with appropriate cues and assistance

Method	Current	Future	Advantage
NP trait models	Unidimensional	Bifactor models, multidimensional IRT models (mIRT)	Each item can provide information about different traits; a single item or test can help specify both general factors and domain scores
Nominal response model	Different kinds of errors are treated identically	Each wrong response has a different meaning	Each item carries more information, enabling greater precision and/or assessing different constructs
Test linking	Total scores are compared in studies that use both tests	Item banks can be drawn from existing tests and new items, and all items calibrated together	Enables direct comparison of different tests and construction of new tests that are back-compatible with the originals
Computerized adaptive testing	Paper-pencil, fixed administration order, minimal branching	Information from each item response selection and speed used to select next most informative item	Efficiency gain of 50-95% in administration time or precision of measurement.
Differential item functioning (DIF)	Effects of group (diagnostic, age, sexual, racial, ethnic, cultural, etc.) determined by comparing total scores	DIF examines group effects for each item	Increased precision in specifying diagnostic and other group differences that may not be apparent in the scores of the whole test
Person fit statistics	Performance validity based on "cutoff" scores, mostly based on accuracy	Performance validity based on the fit of item response characteristics to the examinees overall estimated trait level	Performance validity can be examined within each test; every item response can be useful in detecting anomalies; increase sensitivity to intentional failure
Non-IRT Item-Level Strategies	Most emphasis on summary scores not trial-by-trial analysis	Focus on sequential dependence of responses and meaning of response sequences	Increased efficiency in identifying primary constructs; identification of qualitatively distinct response patterns

Method	Current	Future	Advantage
Evidence-based diagnostic batteries	Batteries with limited flexibility involve redundant testing	Test selection will proceed based on positive predictive power	Testing efficiently focuses time with respect to differential diagnostic questions or recommendations
Computerized testing	Print publishing model; paper-pencil data acquisition and scoring	Computerized tests for stimulus presentation and response acquisition	precision in timing of stimulus presentation and response collection, automatic recording, scoring and database entry of responses, and automatic updating of software to new versions; acquisition of voice, video, motion.
Web-based testing	Testing done in clinic or lab	Testing done at home or wherever convenient for examinee	Scalable assessment at lower cost
Healthcare informatics and bioinformatics	Test results go to file cabinets, report text goes on medical record	Data elements will be part of medical record and integrated with analytics relating these to other health variables	NP data integrated into comprehensive model of patient; implications pushed to all care-team members and hypotheses fed back to NP clinicians for follow-up; "big data" analytics will find new patterns to inform future evidence-based practice
Mobile platforms	Not used; not trusted	Passive monitoring will dramatically increase data flow; experience sampling will augment self reports	Marked increase in longitudinal repeated measures for self-reports and tests; new variables extracted from passive monitoring
Wearables	Not used; not trusted	Passive monitoring of diverse physiological, activity, and experiential data	Data previously available only in cross-sectional lab studies (sleep, EEG, cardiovascular) will be widely available and assessed longitudinally)
Internet of Things (IOT)	Not used; not trusted	Passive monitoring of activities across multiple environments	Ecologically valid assessments will be done in real-world contexts; and environment can "respond" with appropriate cues and assistance

Paradigm Shift from CTT to IRT-CAT

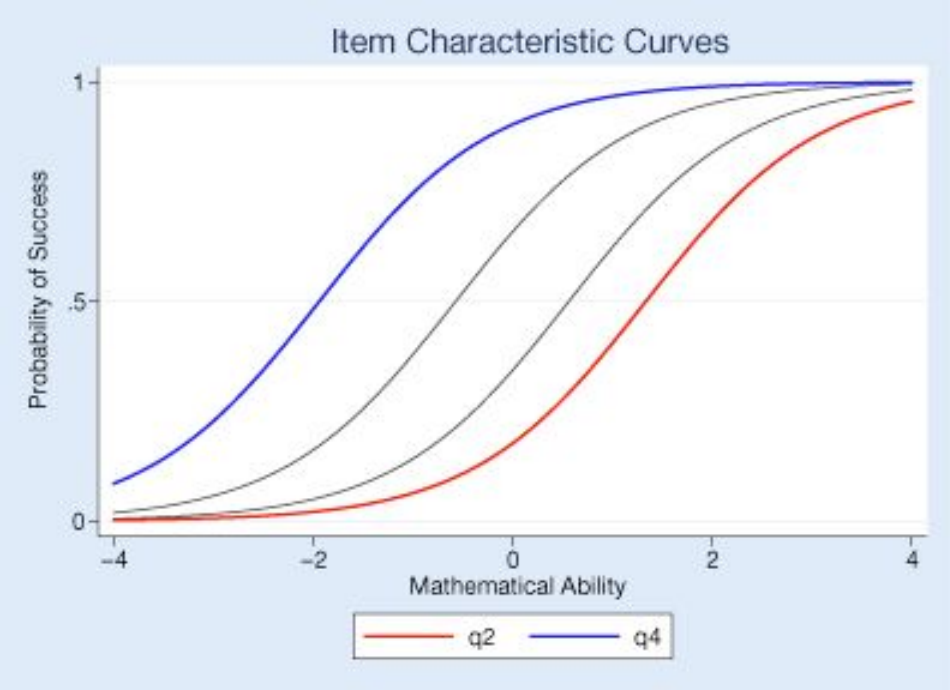
- Traditional Measurement
  - Fix items allow precision to vary
- IRT-Based CAT
  - Fix precision allow items to vary
  - 2x – 5x efficiency gain
- Change precision depending on application
  - Epidemiology – fewer items lower precision (se=.4)
  - Primary care screening – medium precision (se=.3)
  - RCTs – more items high precision (se=.2)

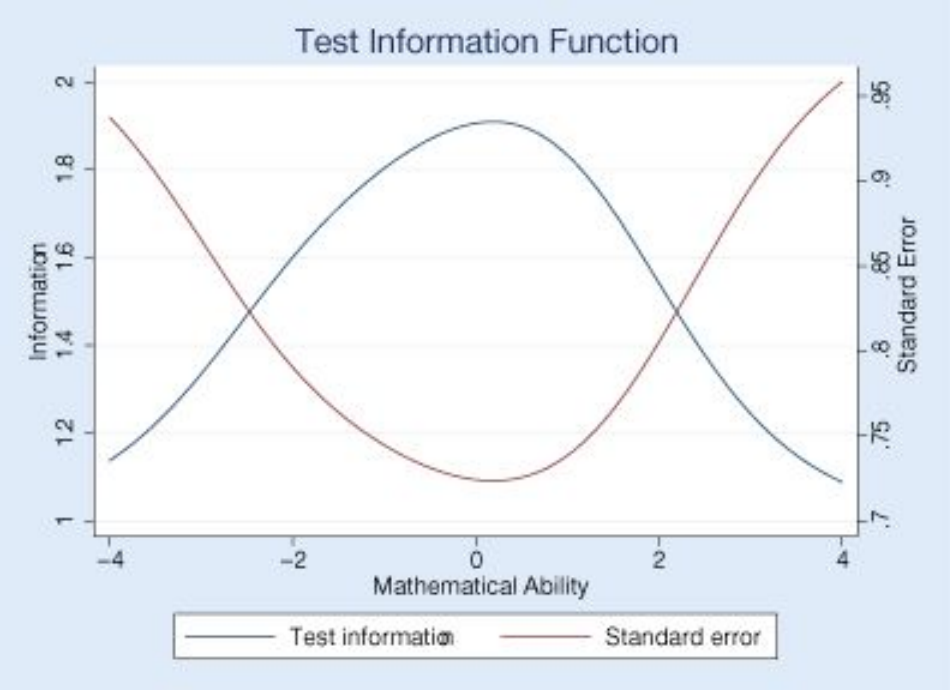


The item parameters can be interpreted as changing the shape of the standard logistic function, with the following parameters:

difficulty, the half-way point between (min) and (max), where the slope is maximized.
discrimination, slope: the maximum slope
pseudo-guessing, or chance, asymptotic minimum

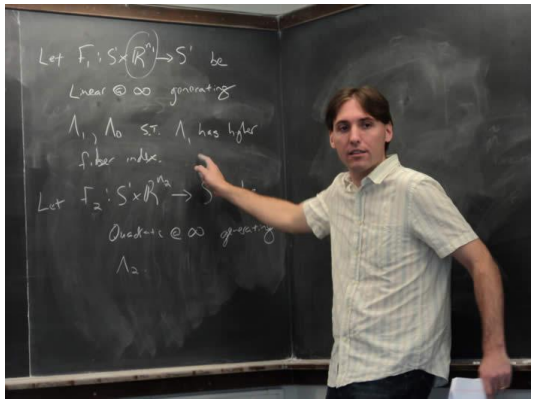
IRT – 4 sample items and test information based on those 4 items





http://www.stata.com/stata-news/news30-1/irt-item-response-theory/

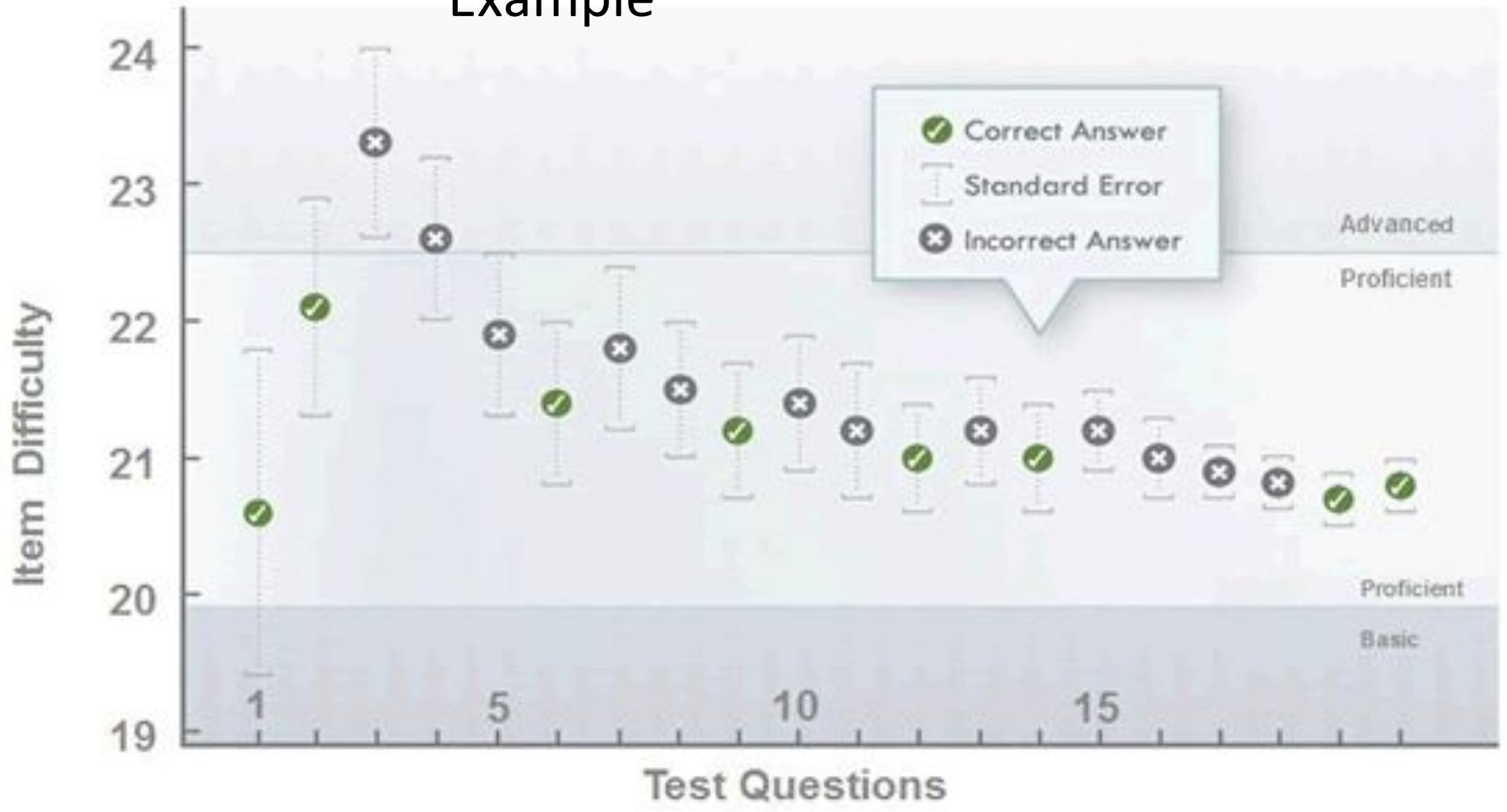
What is CAT?



Arithmetic Algebra Calculus

Imagine a 1000 Item Math Test

Adaptive Test Example



http://www.ascd.org/publications/educational-leadership/mar14/vol71/num06/The-Potential-of-Adaptive-Assessment.aspx

Bi-Factor IRT Model

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}$$

$$P = \int_{-\infty}^{\infty} \left\{ \prod_{v=2}^{d} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n} \left(\Phi \left[\frac{\gamma_j - \alpha_{j1}\theta_1 - \alpha_{jv}\theta_v}{\sqrt{1-\alpha_{j1}^2-\alpha_{jv}^2}} \right] \right)^{u_{jv}} \right] g(\theta_v)d\theta_v \right\} g(\theta_1)d\theta_1 ,$$

$$\hat{\theta}_{1i} = E(\theta_{1i} \mid \mathbf{u}_i, \theta_{2i} \square\ \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} \theta_{1i} \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v)d\theta_v \right\} g(\theta_1)d\theta_1 .$$

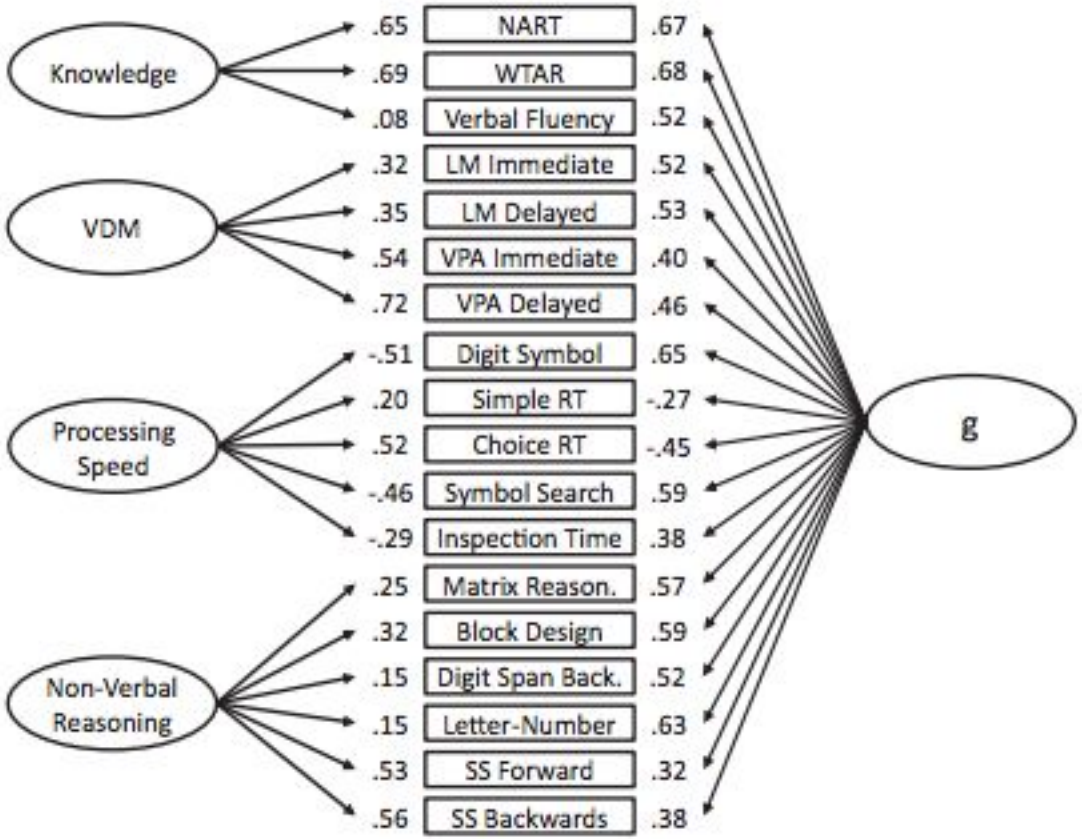
$$V(\theta_{1i} \mid \mathbf{u}_i, \theta_{2i} \square\ \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} (\theta_{1i} - \hat{\theta}_{1i})^2 \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v)d\theta_v \right\} g(\theta_1)d\theta_1 .$$

Gibbons and Hedeker, 1992, *Psychometrika*
Gibbons et.al., 2007, *Applied Psychological Measurement*

Neuropsychology
2013, Vol. 27, No. 5, 595–607

0894-4105/13/$12.00 DOI: 10.1037/a0033354

Brain White Matter Tract Integrity and Cognitive Abilities in Community-Dwelling Older People: The Lothian Birth Cohort, 1936

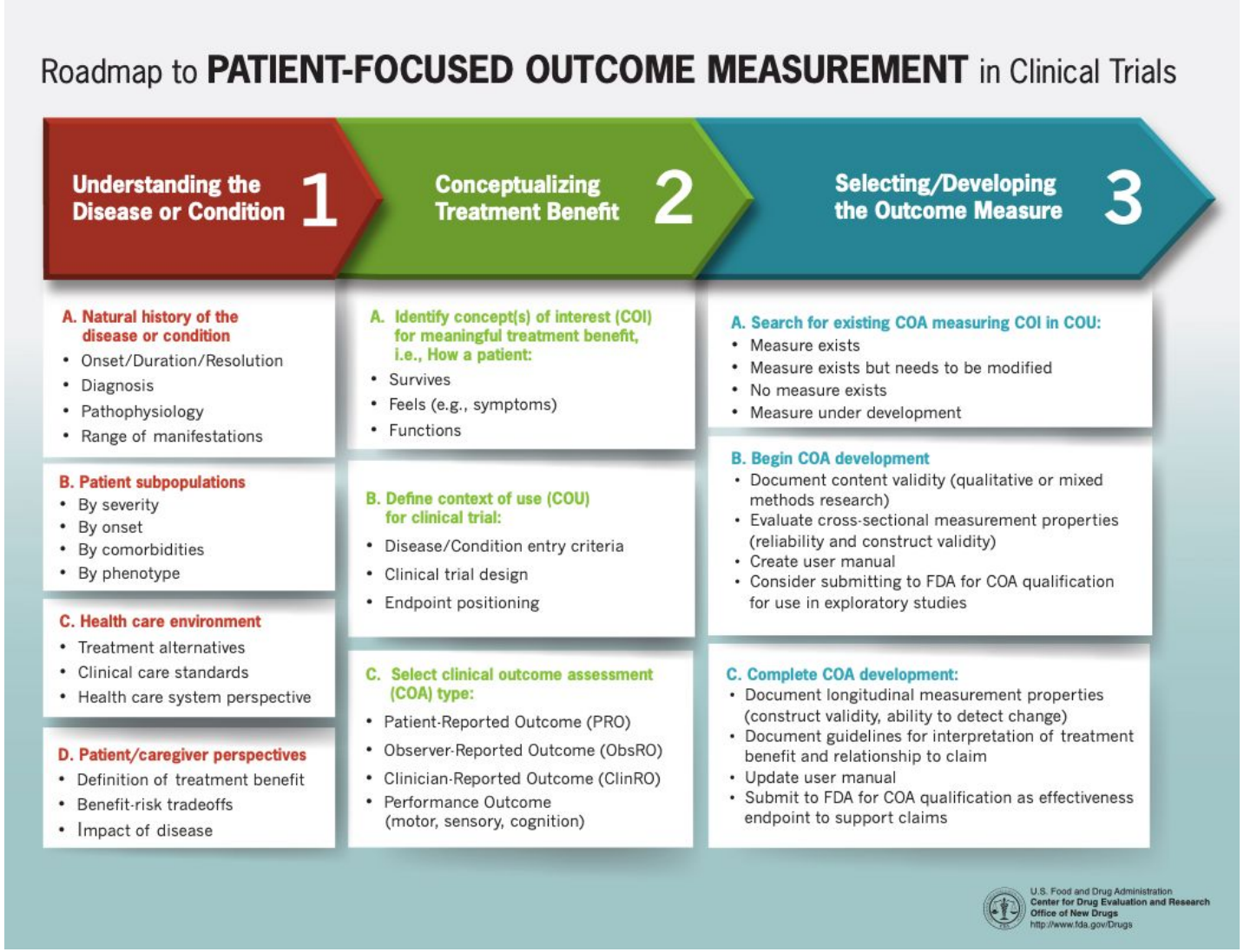


Booth et al., 2013

Adaptive Testing for Clinical NP Assessment

- Prior probabilities of exam outcome (diagnoses, descriptives, recommendations) based on:
  - Referral Question
  - Demographics
  - History and Lab Results
- Stage 1: Select next procedure based on positive predictive power for each exam outcome
- Stage 2: Within procedure, select relevant precision and next item that maximizes information content
- Repeat until exit criteria are satisfied

What are we learning from clinical trials research – clinical outcome assessment (COA)?



New clinical outcomes assessment methods require new strategies

- Changes compared to old-fashioned RCTs
- Traditional RCT - primary endpoint was usually:
  - A test summary score…
  - Reflecting performance across a bunch of items…
  - From a single test instrument…
  - That was administered by a trained human…
  - With results recorded on a clinical record form and…
  - Then transcribed into a database for analysis…

New behavior sampling methods require new strategies

- Changes compared to old-fashioned RCTs with primary endpoint include:
  - Dense temporal sampling
  - Multivariate sampling
  - Passive sampling
  - Machine sampling
  - More direct sampling of biological variables

Temporal sampling density

- Increased density of observations (from mobile, wearable or IOT)
- Sampling may occur more than 1 per second
  - consider: 8 weeks x 7 days x 24 hours x 60 minutes x 60 seconds = 4.84M measures
- Analyze trajectories rather than simple changes from baseline to endpoint
- [give example from NIDDK study]

Assessment
2018, Vol. 25(1) 14–30

DOI: 10.1177/1073191116643164
journals.sagepub.com/home/asm
SAGE

Reliability and Validity of Ambulatory Cognitive Assessments

Martin J. Sliwinski[1], Jacqueline A. Mogle[1], Jinshil Hyun[1], Elizabeth Munoz[1], Joshua M. Smyth[1], and Richard B. Lipton[2]

Abstract

Mobile technologies are increasingly used to measure cognitive function outside of traditional clinic and laboratory settings. Although ambulatory assessments of cognitive function conducted in people's natural environments offer potential advantages over traditional assessment approaches, the psychometrics of cognitive assessment procedures have been understudied. We evaluated the reliability and construct validity of ambulatory assessments of working memory and perceptual speed administered via smartphones as part of an ecological momentary assessment protocol in a diverse adult sample (N = 219). Results indicated excellent between-person reliability (≥0.97) for average scores, and evidence of reliable within-person variability across measurement occasions (0.41-0.53). The ambulatory tasks also exhibited construct validity, as evidence by their loadings on working memory and perceptual speed factors defined by the in-lab assessments. Our findings demonstrate that averaging across brief cognitive assessments made in uncontrolled naturalistic settings provide measurements that are comparable in reliability to assessments made in controlled laboratory environments.

Table 2. Reliabilities for Individual and Aggregated Ambulatory Cognitive Test Scores.

	Symbol search	Dot memory	*n*-Back
Between-person variance	0.44	0.77	0.02
Within-person variance	0.36	1.19	0.01
Reliability of 1 occasion (ICC)	0.54	0.39	0.59
Reliability of average (1 day)	0.90	0.75	0.92
Reliability of average (2 days)	0.95	0.86	0.95
Reliability of average (3 days)	0.96	0.90	0.96
Reliability of average (14 days)	0.98	0.97	0.99

Note. ICC = intraclass correlation. There are five assessment occasions per day, so 1 day reflects 5 assessments, 2 days reflect 10 assessments, and so on.

Multivariate sampling

- Single mobile device yields multiple outputs in different modalities
  - GPS
  - Motion
  - Voice
  - Video: light/dark, facial affect, oxygenation
  - EMA
  - GSR
  - HR, HRV
- Or data may be integrated across multiple devices
  - Smart watch or actigraphy
  - Skin patch sensor
  - Sleep respiration monitor
  - EEG, EKG, etc…
- Methods to aggregate all these data types into composite COAs under development…

Passive sampling = more objective

- Less censoring and bias of data related to:
  - Compliance
  - Effort
  - Intent
- Examinee less *prepared* for assessment
  - Measures less likely to be affected by expectancy biases
  - Presumably better at overcoming placebo effects

Machine sampling

- Increased precision
- Probably decreased flexibility
  - All flexibility must be programmed in advance (there is no “on the fly” flexibility that occurs with humans, for better or worse)
  - Interaction monitoring still early (e.g., interactive video monitoring of engagement during assessment)
- Unclear impacts on human responders
  - Tech naïve older adults vs early adopters
  - Consider “rod & frame” studies…

Reliability and Validity Issues

- Reliability
  - Internal consistency, construct validity
  - Test-retest reliability: stability, bias, effects of repeated measurement
  - Inter-rater, Inter-site, Inter-national reliability
  - At least as good as conventional measures?
- Criterion validity
  - With respect to existing measures
  - With respect to clinical outcomes
  - At least as good as conventional measures?

Using IRT for co-calibration of tests and longitudinal assessment

- Test linking
  - Quantify how different methods identify individuals with respect to a shared latent trait that both instruments measure
  - Typically requires at least some linking or “anchor” items
  - Examine differential item functioning (DIF) for anchor items
  - Summaries include:
    - Test characteristic curves: plot most likely score for each level of ability
    - Test information curves: plot measurement precision at each level of ability
- Assumption that test characteristics are constant over time is probably wrong
  - Regression and change score approaches all assume linearity across scale – not true for virtually any test

ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 1018–1027

Journal of Clinical Epidemiology

Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline

Paul K. Crane[a,*], Kaavya Narasimhalu[a], Laura E. Gibbons[a], Dan M. Mungas[b], Sebastien Haneuse[c], Eric B. Larson[c], Lewis Kuller[d], Kathleen Hall[e], Gerald van Belle[f]

[a]*Department of Medicine, University of Washington*
[b]*Department of Neurology, University of California at Davis*
[c]***Center for Health Studies, Group Health Cooperative***
[d]*Department of Epidemiology, University of Pittsburgh*
[e]*Department of Psychiatry, Indiana University – Purdue University in Indianapolis*
[f]*Departments of Biostatistics and Environmental and Occupational Health Sciences, University of Washington*

Accepted 18 November 2007

Abstract

Objective: To cocalibrate the Mini-Mental State Examination, the Modified Mini-Mental State, the Cognitive Abilities Screening Instrument, and the Community Screening Instrument for Dementia using item response theory (IRT) to compare screening cut points used to identify cases of dementia from different studies, to compare measurement properties of the tests, and to explore the implications of these measurement properties on longitudinal studies of cognitive functioning over time.

Study Design and Setting: We used cross-sectional data from three large ($n > 1000$) community-based studies of cognitive functioning in the elderly. We used IRT to cocalibrate the scales and performed simulations of longitudinal studies.

Results: Screening cut points varied quite widely across studies. The four tests have curvilinear scaling and varied levels of measurement precision, with more measurement error at higher levels of cognitive functioning. In longitudinal simulations, IRT scores always performed better than standard scoring, whereas a strategy to account for varying measurement precision had mixed results.

Conclusion: Cocalibration allows direct comparison of cognitive functioning in studies using any of these four tests. Standard scoring appears to be a poor choice for analysis of longitudinal cognitive testing data. More research is needed into the implications of varying levels of measurement precision.

Keywords: Cognition; Cocalibration; Item response theory; Psychometrics; Longitudinal data analysis; Simulation

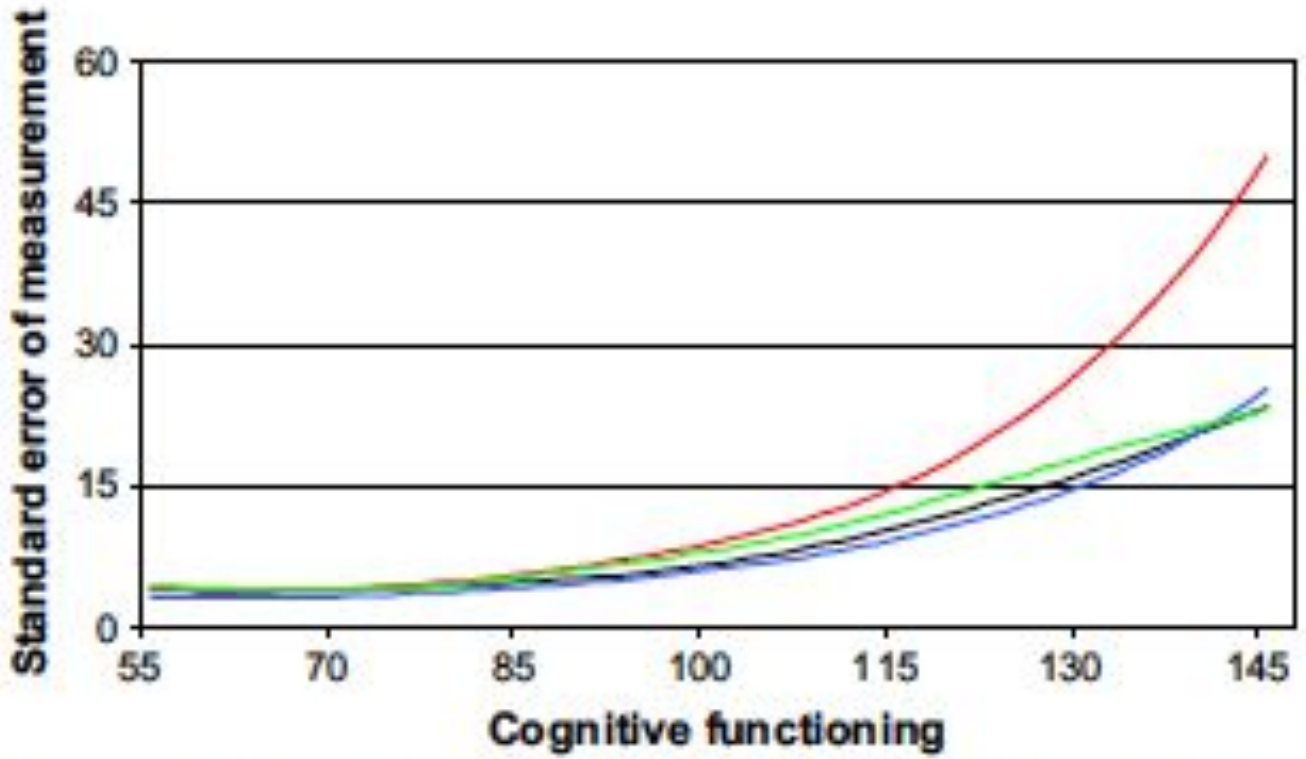


Fig. 2. Standard error of measurement for tests of global cognitive functioning. Red = MMSE, blue = 3MS, black = CASI, green = CSI 'D'. The *x*-axis is the rescaled IRT score, with a mean of 100 and a standard deviation of 15.

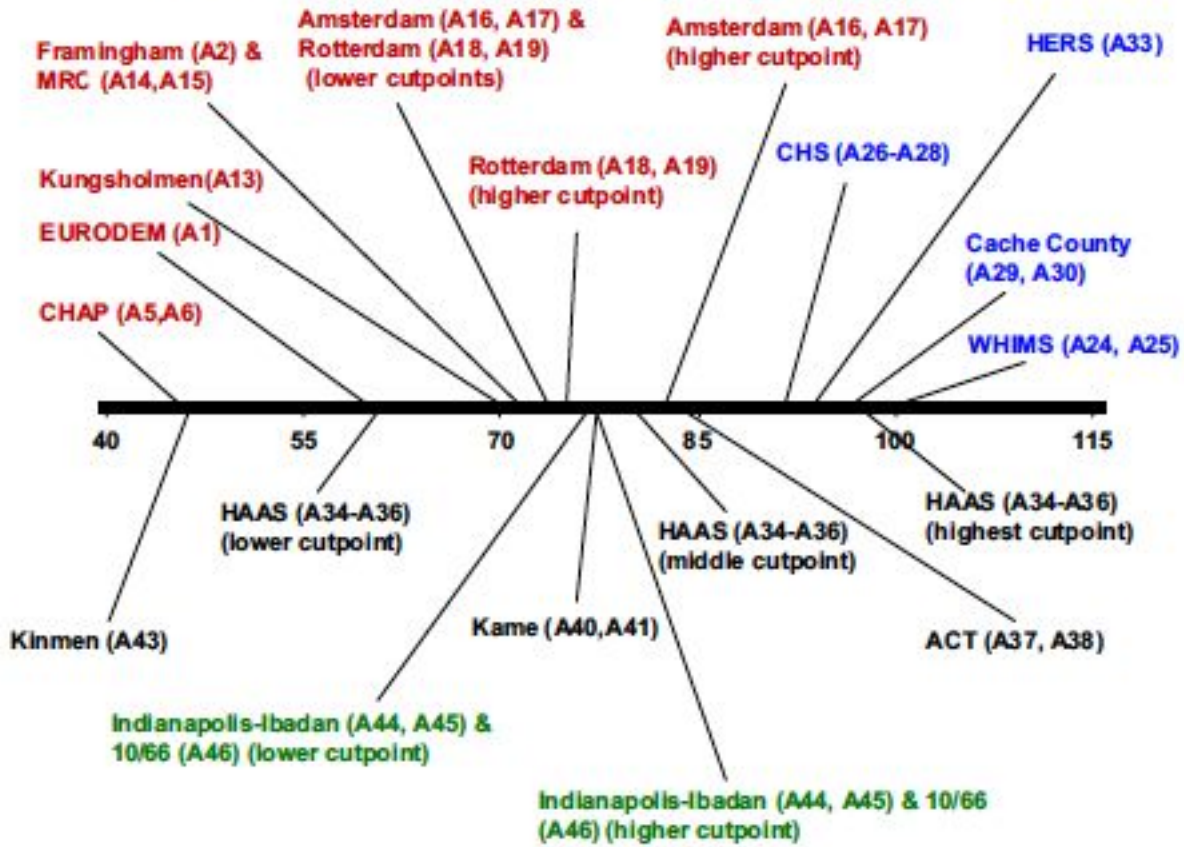


Fig. 3. Screening cut points used in selected studies. Red = MMSE, blue = 3MS, black = CASI, green = CSI 'D'. The *x* axis is the rescaled IRT score, with a mean of 100 and a standard deviation of 15. Screening cut points were abstracted from the sources cited in the figure. Thus, there is variability in screening cut points across studies that use the same instrument (identified in the figure by ink color). The unique contribution here is the ability to compare cut points across studies that used different tests, such as the MMSE cut point used by the Framingham study and the CASI cut point used by the ACT study. The cut points used in the different studies varied by nearly 4 standard deviations, a huge variation.

Methods to Assure Equivalency

- General measurement invariance issues, using multiple group confirmatory factor analysis (CFA)
  - Equal form: The number of factors and the pattern of factor-indicator relationships are identical across groups.
  - Equal loadings: Factor loadings are equal across groups.
  - Equal intercepts: When observed scores are regressed on each factor, the intercepts are equal across groups.
  - Equal residual variances: The residual variances of the observed scores not accounted for by the factors are equal across groups.

Measurement Invariance: Levels of Equivalence

- Configural equivalence: The factor structure is the same across groups in a multi-group confirmatory factor analysis.
- Metric equivalence: Factor loadings are similar across groups.
- Scalar equivalence: Values/Means are also equivalent across groups.

Challenges for Measurement Invariance for Introducing New Methods into Clinical Trials

- Assessment of measurement invariance typically requires:
  - Shared “linking” items across instruments that serve as “anchors” against which other aspects of covariance can be judged
- Absent linking items, comparability can be established by studying the same people with both methods. This is the conventional criterion validity approach or assessment of “concurrent validity.”
- Other strategies are possible for integrative data analysis:
  - Variable network harmonization
  - Covariance structure harmonization
  - Factor alignment

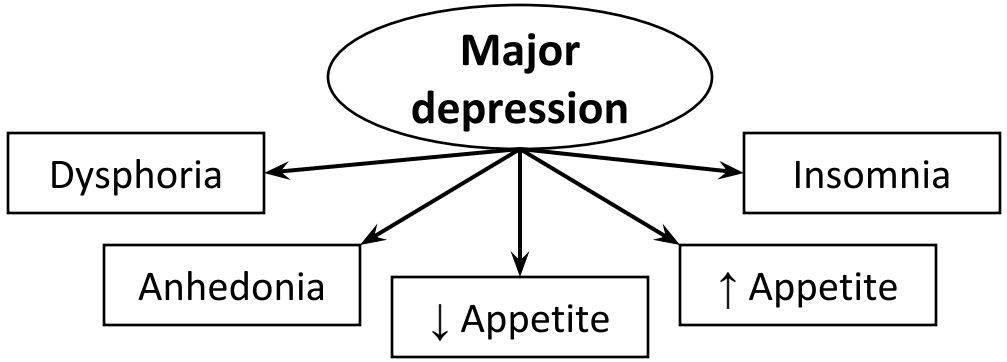
Classical psychometric and network approaches to measurement invariance

Psychometric model

Assumes latent variable

- Constrains correlations

Psychometric



Network model

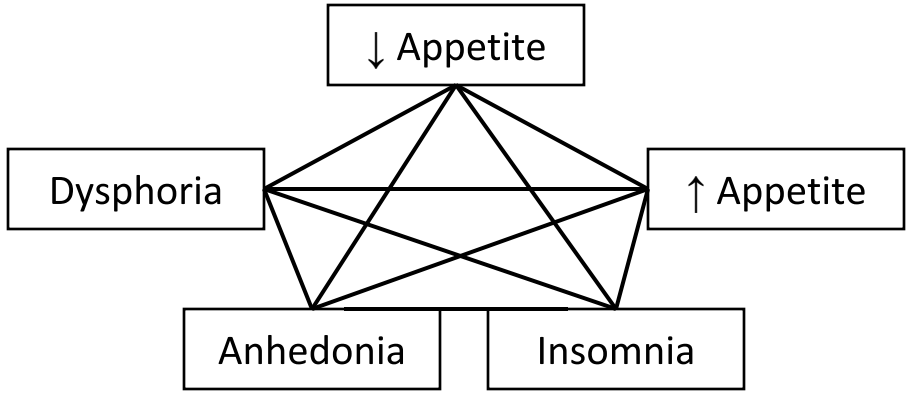
No constraints on correlations

- Saturated model

If networks harmonize…

- … so will factor model
- … so will composites

Network



Harmonization – important?

- Increased study size and global scope: increased need for accurate phenotype specification
- NIMH WGSPD Consortium
  - 20K+ participants, 4 projects, 10+ countries over 5 continents, diverse methods for phenotyping of diagnoses, symptoms, and cognitive function; 5 diagnostic categories (schizophrenia, schizoaffective disorder, bipolar disorder, major depressive disorder, and autism spectrum disorders), plus "controls"
- What phenotypes can we pool *with confidence* across studies?
- Prior art: Ruderfer et al. 2014 – EFAs/CFAs in different samples to arrive at SCZ and BP factors, standardized *within* each site

Molecular Psychiatry (2014) 19, 1017–1024

www.nature.com/mp

ORIGINAL ARTICLE

Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia

DM Ruderfer[1,23], AH Fanous[2,3,4,23], S Ripke[5,6,23], A McQuillin[7], RL Amdur[2] Schizophrenia Working Group of the Psychiatric Genomics Consortium[24] Bipolar Disorder Working Group of the Psychiatric Genomics Consortium[24] Cross-Disorder Working Group of the Psychiatric Genomics Consortium[24], PV Gejman[8], MC O'Donovan[9], OA Andreassen[10], S Djurovic[10], CM Hultman[11], JR Kelsoe[12,13], S Jamain[14], M Landén[11,15], M Leboyer[14], V Nimgaonkar[16], J Nurnberger[17], JW Smoller[18], N Craddock[9], A Corvin[19], PF Sullivan[20], P Holmans[9,21], P Sklar[1,25] and KS Kendler[4,22,25]

Leaves unclear:

- how comparable are phenotypes across sites, in pattern, level?

Our aim:

- identify phenotypes that are similar *across* sites, using quantitative thresholds for harmonization.

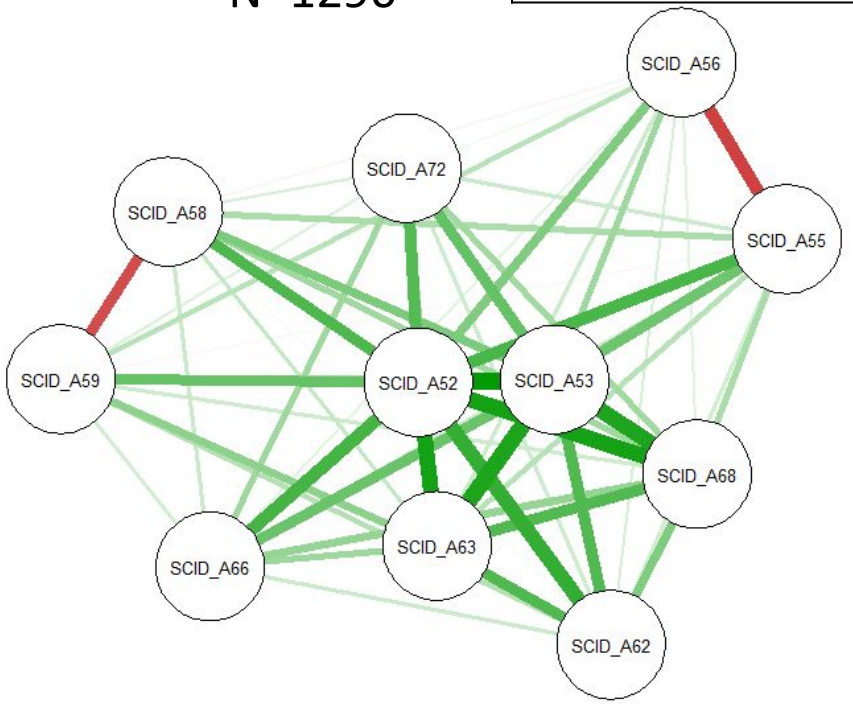
Depression – Matching symptoms

Model fit: CFI=.992, RMSEA=.061, SRMR=.089

Symptom Name	SCID	DI-PAD
Dysphoria (Depression)	A52	OPCRIT37
Loss of pleasure	A53	OPCRIT39
Weight loss/decreased appetite	A55	OPCRIT489
Weight gain/increased appetite	A56	OPCRIT501
Insomnia	A58	OPCRIT4456
Excessive sleep	A59	OPCRIT47
Slowed activity	A62	OPCRIT24
Loss of energy or fatigue	A63	OPCRIT25
Inappropriate guilt	A66	OPCRIT42
Impaired Concentration	A68	OPCRIT41
Suicidal ideation	A72	OPCRIT43

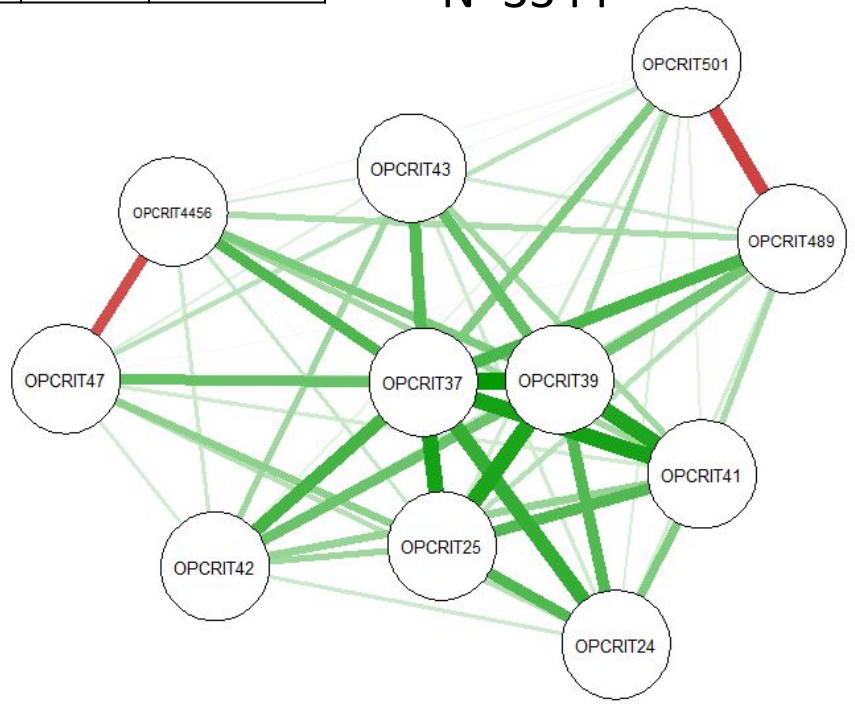
SCID

N=1290



DI-PAD

N=3344



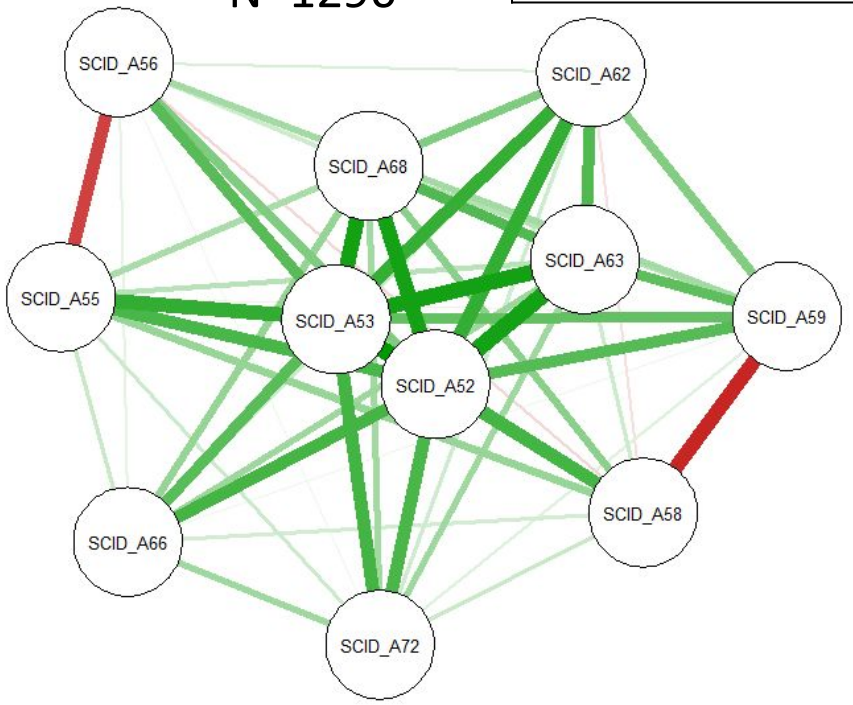
Depression – Matching symptoms

Model fit: CFI=.999, RMSEA=.032, SRMR=.038

Symptom Name	SCID	DI-PAD	MAD_r
Dysphoria (Depression)	A52	OPCRIT37	
Loss of pleasure	A53	OPCRIT39	.256
Weight loss/decreased appetite	A55	OPCRIT489	
Weight gain/increased appetite	A56	OPCRIT501	
Insomnia	A58	OPCRIT4456	.114
Excessive sleep	A59	OPCRIT47	.161
Slowed activity	A62	OPCRIT24	
Loss of energy or fatigue	A63	OPCRIT25	
Inappropriate guilt	A66	OPCRIT42	
Impaired Concentration	A68	OPCRIT41	
Suicidal ideation	A72	OPCRIT43	

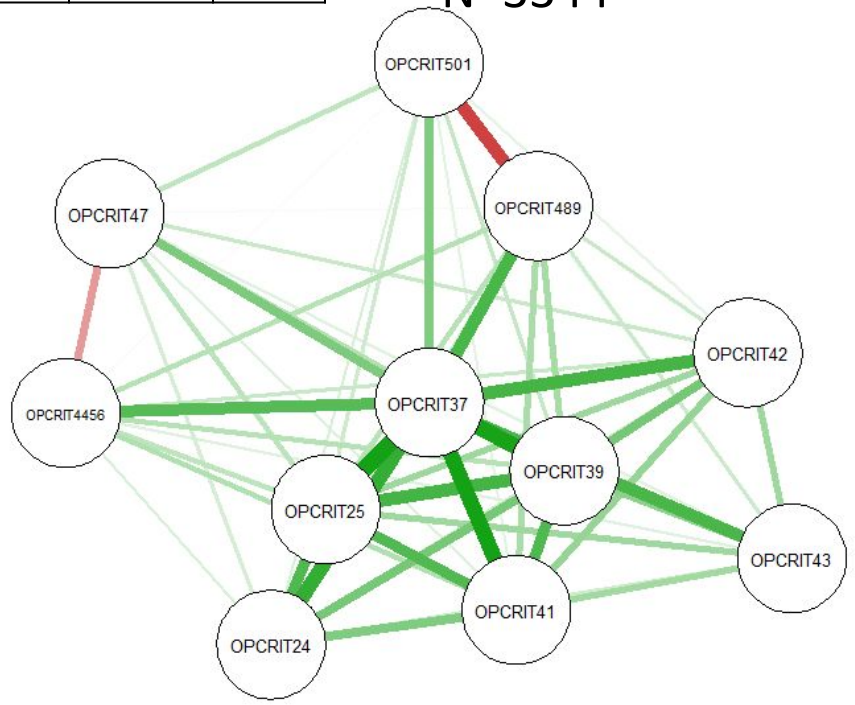
SCID

N=1290



DI-PAD

N=3344



Depression – Non-matching symptoms

Symptom Name	SCID	DI-PAD
Psychomotor agitation	A61	
Feelings of worthlessness	A65	
Indecisiveness	A69	
Recurrent thoughts of death	A71	
Specific plan	A73	
Suicide attempts	A74	
Altered libido		OPCRIT40
Diurnal variation		OPCRIT38

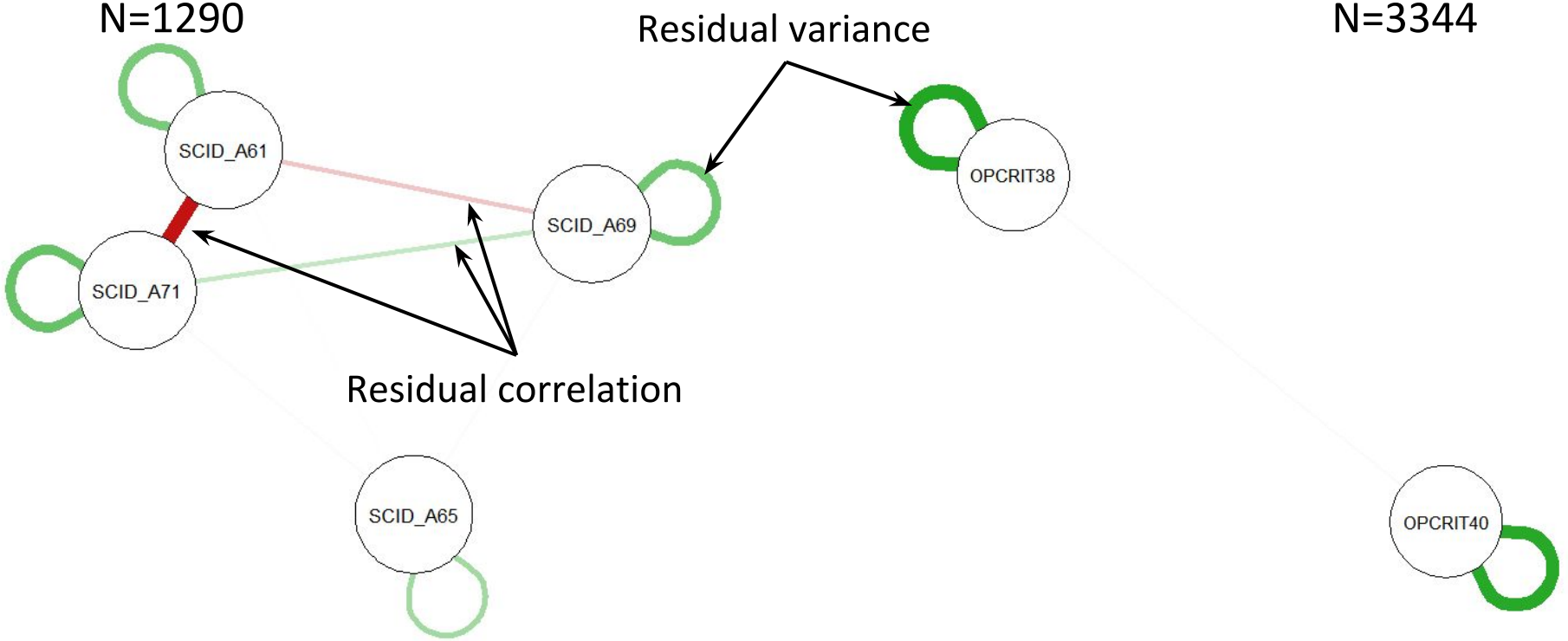
Low residual variance

SCID

N=1290

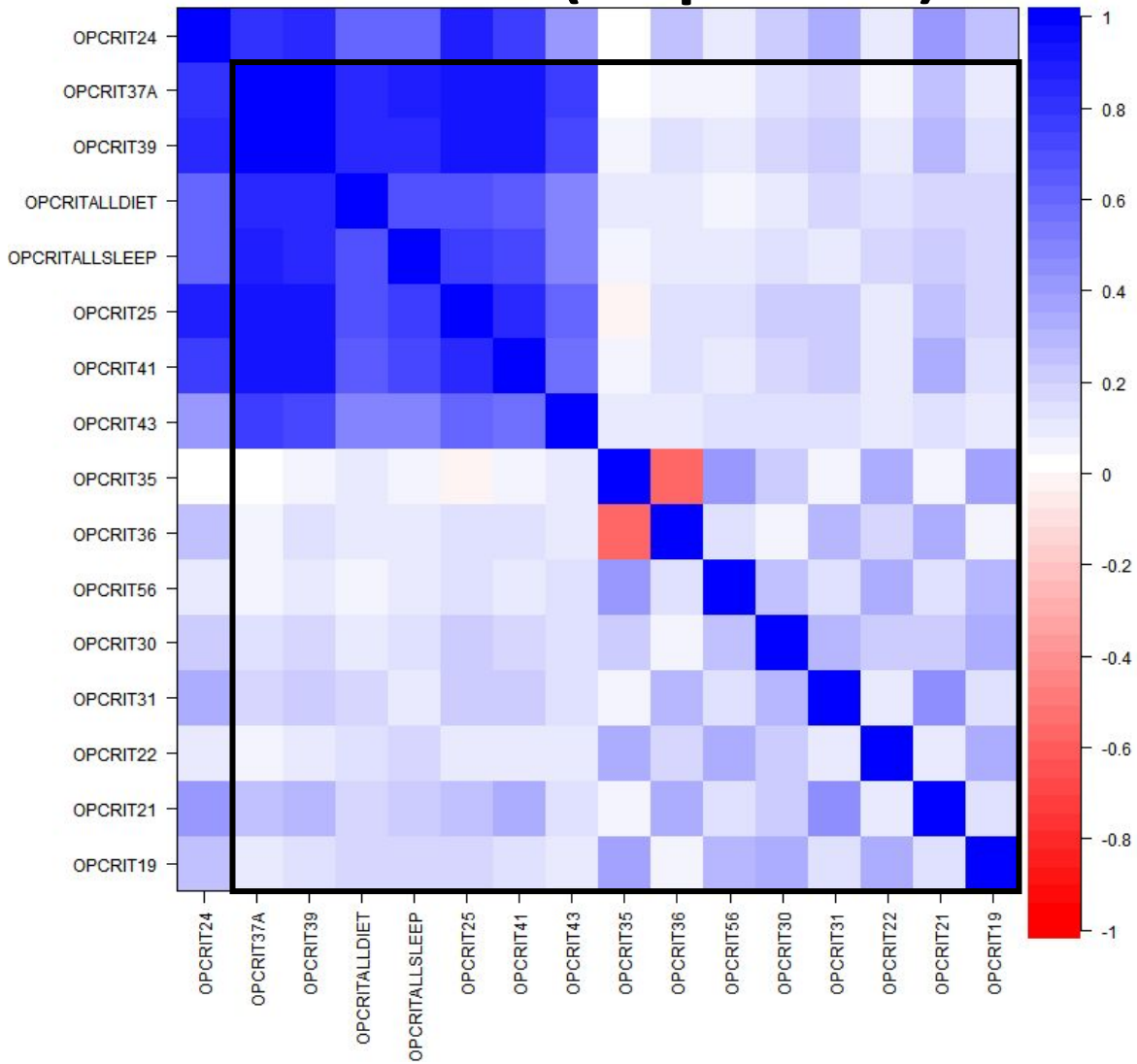
DI-PAD

N=3344

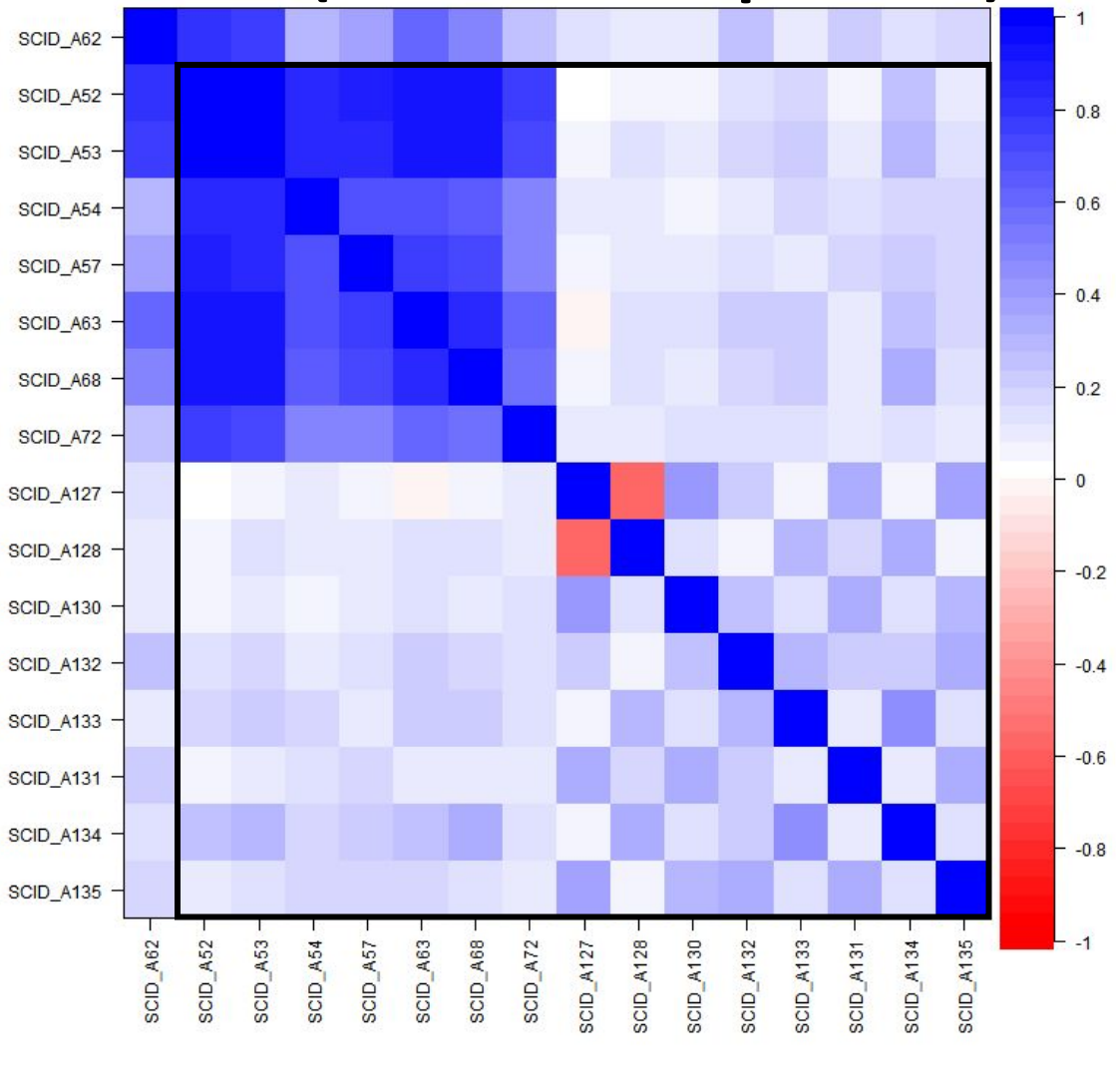


Harmonized Model Results

DI-PAD (Bipolar)



SCID (Dutch bipolar)



GCTA results: Harmonization factors

including 10PCs

Quantile normalization
of phenotypes

Increase to 230k,

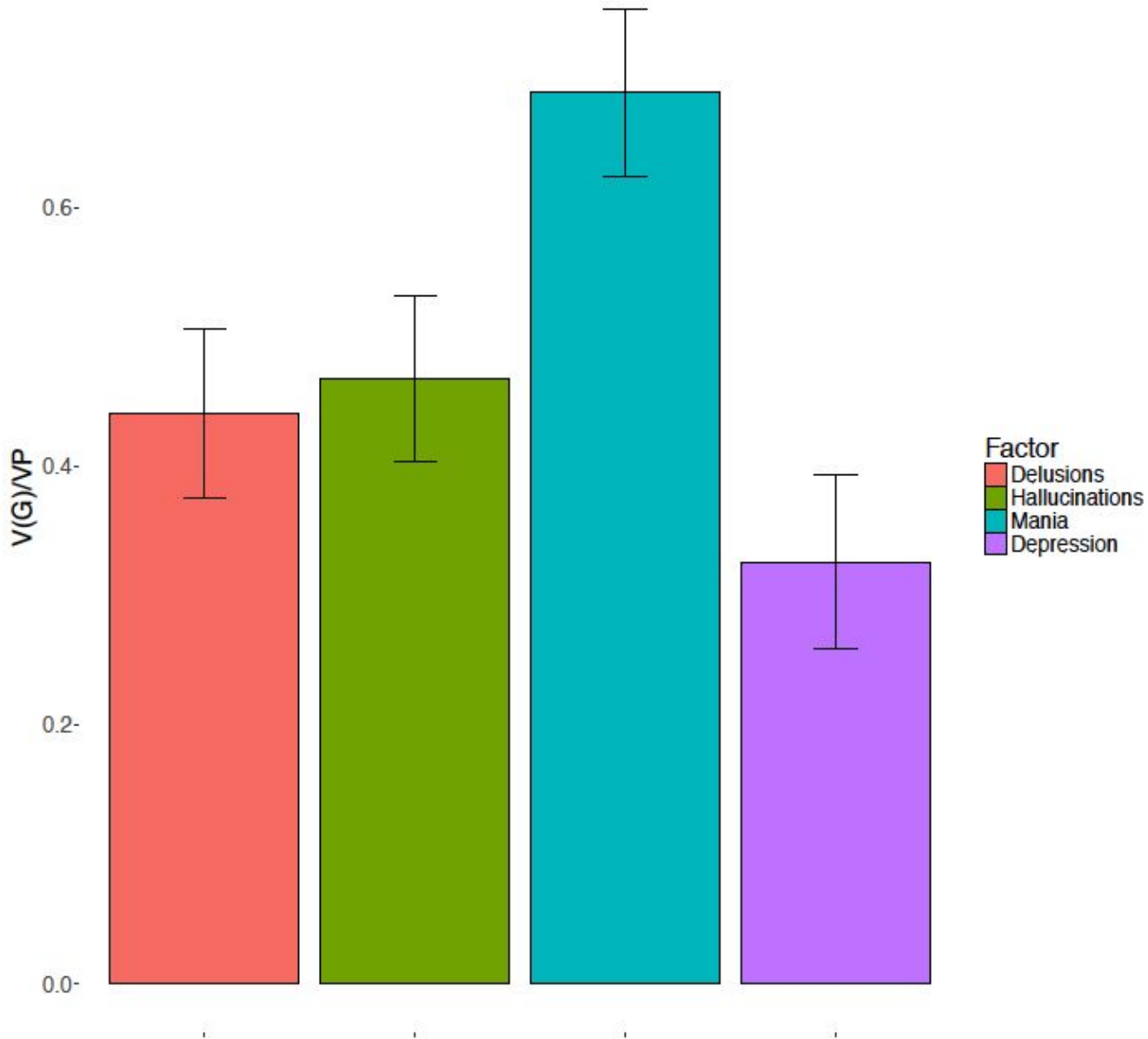
drop PatoGlobal

-> similar results

genetic correlation

between factors,

but not enough power



frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 September 2014
doi: 10.3389/fpsyg.2014.00978

IRT studies of many groups: the alignment method

Bengt Muthén* and Tihomir Asparouhov

Mplus, Los Angeles, CA, USA

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Eldad Davidov, University of Zurich, Switzerland
Joop J. Hox, Utrecht University, Netherlands
Aleksandra Bujacz, Stockholm University, Sweden

****Correspondence:***
Bengt Muthén, Mplus, 3463 Stoner Ave, Los Angeles, CA 90066, USA
e-mail: bmuthen@statmodel.com

Asparouhov and Muthén (2014) presented a new method for multiple-group confirmatory factor analysis (CFA), referred to as the alignment method. The alignment method can be used to estimate group-specific factor means and variances without requiring exact measurement invariance. A strength of the method is the ability to conveniently estimate models for many groups, such as with comparisons of countries. This paper focuses on IRT applications of the alignment method. An empirical investigation is made of binary knowledge items administered in two separate surveys of a set of countries. A Monte Carlo study is presented that shows how the quality of the alignment can be assessed.

Keywords: factor means invariance testing country comparisons, approximate invariance maximum-likelihood, Bayesian inference, invariance testing, maximum likelihood estimation

Next, alignment proceeds as in the continuous case by minimizing the graded response model (GRM) complexity function:

$$F_{GRM} = \sum_{p} \sum_{g_1<g_2} w_{g_1,g_2} f\left(\lambda_{pg_1,1} - \lambda_{pg_2,1}\right) + \sum_{p} \sum_{g_1<g_2} \sum_{q} w_{g_1,g_2} f\left(\nu_{pqg_1,1} - \nu_{pqg_2,1}\right)$$

Note the extra summation in the second term, which accounts for multiple measurement intercepts in the graded response model. After the model parameters are aligned in the factor analytic metric, the aligned IRT model parameters are given by the following transformations:

$$a_{pg_1,1} = \frac{1.7 * \lambda_{pg_1,1} * \sqrt{\psi_{pg}}}{\sqrt{1 - \lambda_{pg_1,1}^2 \psi_{pg}}}$$

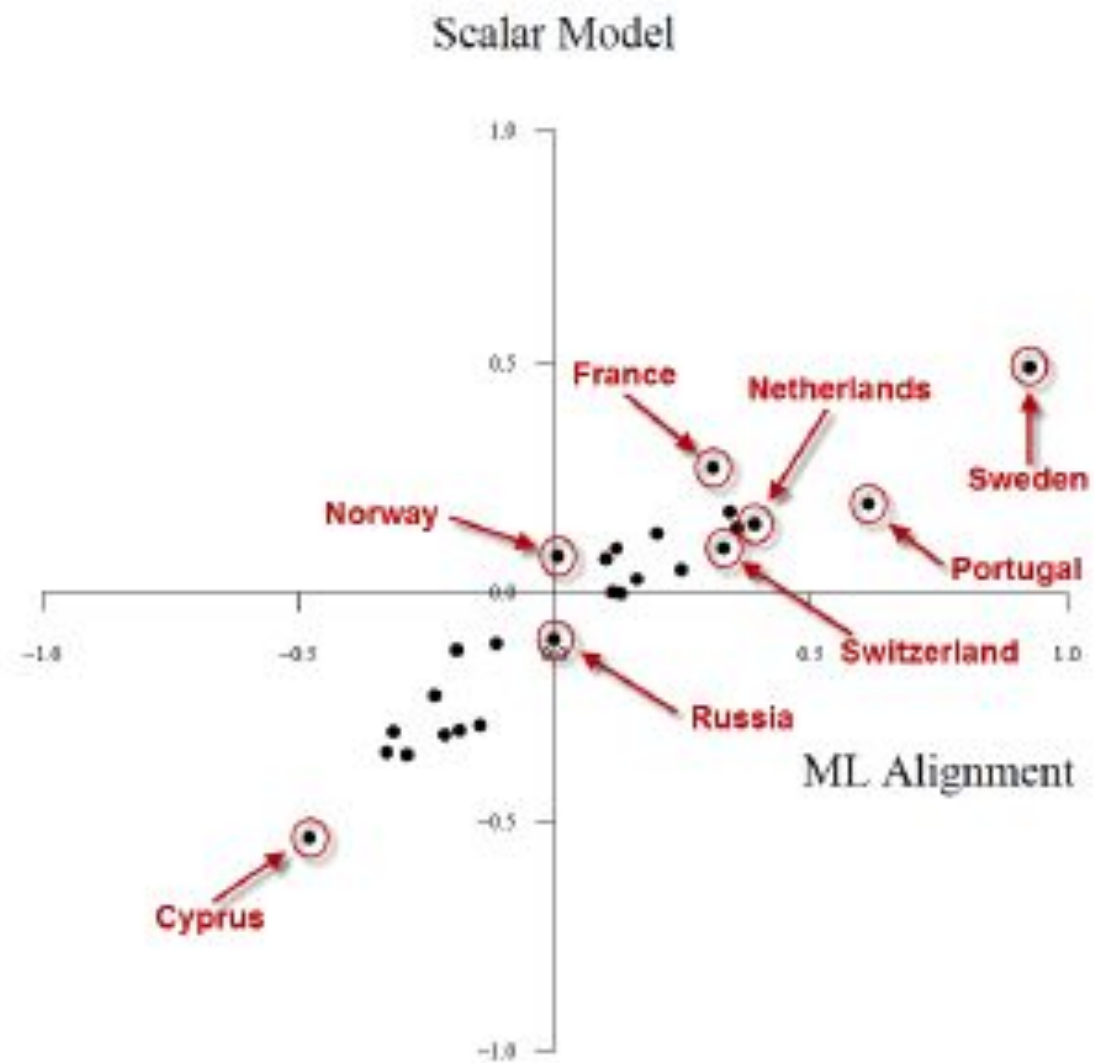
$$d_{pqg_1,1} = d_{pqg_1,1} - a_{pg_1,1} * \alpha_g$$

With these modifications, the final alignment complexity function is given by

$$F_{GRM}^{*} = \sum_{g_1<g_2} \sum_{p \in I_1, p \in I_2} w_{g_1,g_2} f\left(\lambda_{pg_1,1} - \lambda_{pg_2,1}\right) + \sum_{g_1<g_2} \sum_{p \in I_1, p \in I_2} w_{g_1,g_2} f\left(\nu_{p0g_1,1} - \nu_{p0g_2,1}\right)$$

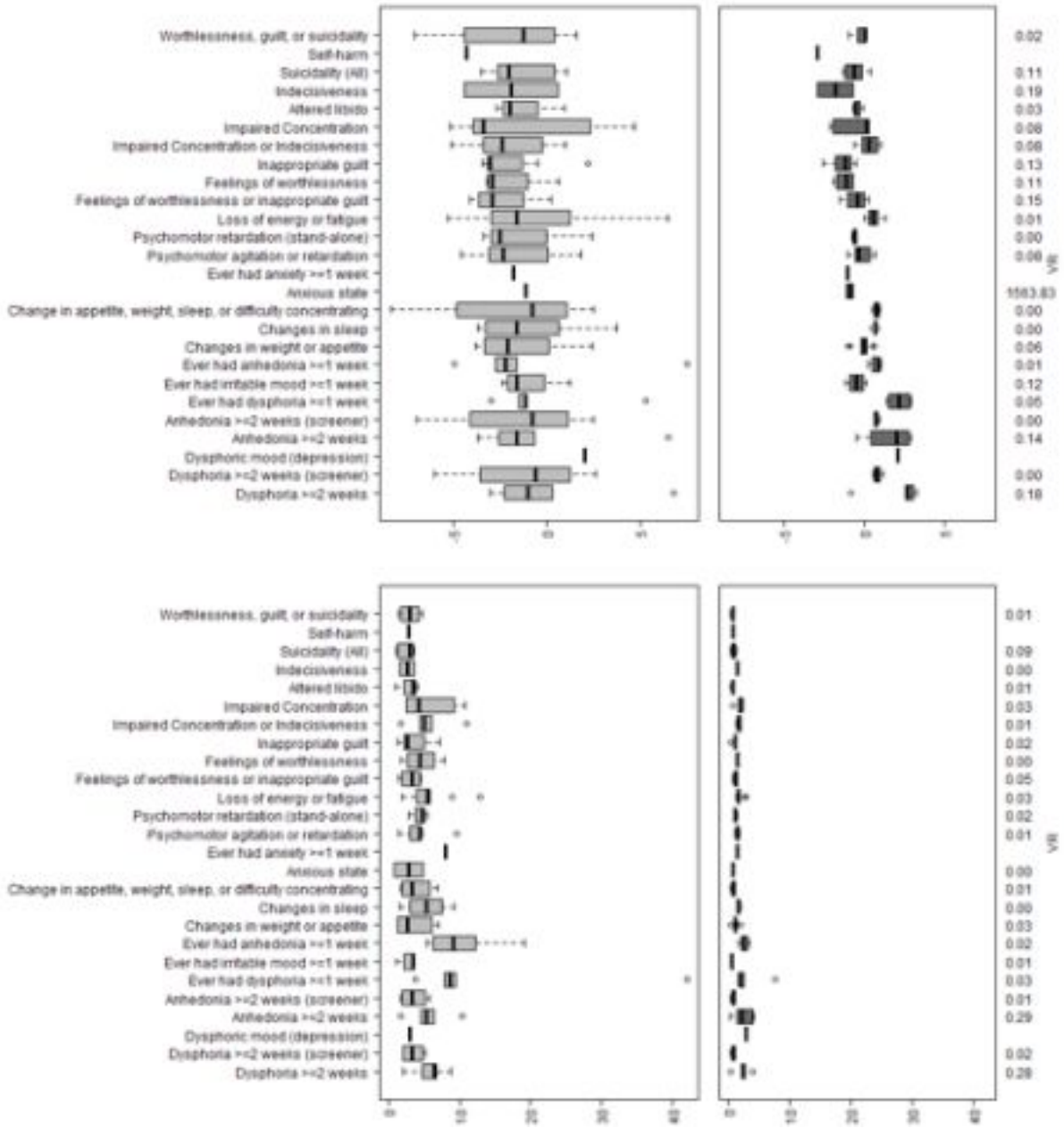
As described above, measurement non-invariance is only minimized for items which appear in each pair of instruments, and only the first measurement intercept is considered.

Figure 1: Factor Means for Tradition-Conformity in 26 Countries: Alignment Method vs Scalar Model



IRT PARAMETER ESTIMATES BEFORE AND AFTER ALIGNMENT: FACTOR = DEPRESSION

Distributions of parameter estimates in the configural and aligned models for depression. *Top left*: Intercept parameters in the configural model. *Top right*: Intercept parameters in the aligned model. *Bottom left*: Slope parameters in the configural model. *Bottom right*: Slope parameters in the aligned model. *VR* = variance ratio, calculated as the ratio of post-alignment parameter variance to pre-alignment parameter variance.



How do we get from here to there?

- Barriers
  - Legacy instruments … have a legacy
  - Inertia in the NP customer base:
    - concerns about validity of new methods
    - Prefer “tried and true” or familiar methods
  - CATs require large samples to calibrate items for IRT analysis
  - Evaluation of positive predictive power for different exam outcomes requires large samples
  - Assuming we can get enough data to generate a CAT approach to the NP exam, how would it be implemented?

National Neuropsychology Network

- National Data Archives (NDA) now aggregating item-level test data for NIH projects (autism, RDoC, ADNI), n's increasing (RDoC=12k total), BUT…
  - Patient selection follows grant inclusion/exclusion criteria – how representative?
  - Test selection follows grant protocols, usually selected experimental measures, often not tests most widely used in practice
- Meanwhile: ***Clinical NP exams = 500K/year (!)***
- Proposed:
  - National Neuropsychology Network: clinical sites sharing item-level data with NDA for open analysis, generation of back-compatible, efficient assessments, and forward-looking introduction of novel items to expand banks for existing and novel construct measurement

How to launch the Neuropsychology Liberation Front?

- Collaborative data aggregation at the item level across clinics, nationwide
- Need to provide shared access to item-level data in a way that provides appropriate:
  - Privacy
  - Data security
  - Practicality for busy clinicians and staff
- Solutions:
  - Leverage current methods for data collection (e.g., Pearson Q-Interactive)
  - Develop novel software for point of testing data acquisition
  - Use existing privacy/security protocols developed by NIH for data archives (GUID)
  - Proposal: Submitted February 2018, recommended for funding early in FY19
- GOAL: simultaneously make life *easier* for clinicians AND share data to support assessments of the future.

The Towers of Babel, London, Hanoi…: Which instruments should be included?

- Daunting challenge: how to accommodate the broad range of tests used?
- Surprise: despite flexible approaches to NP there is considerable homogeneity of actual tests used
- Rabin et al (2016) survey – 80% of exams covered by:
  - Wechsler Adult Intelligence Scale, 4th Edition (WAIS-IV)
  - Wechsler Memory Scale, 4th Edition (WMS-IV)
  - California Verbal Learning Test, 2nd Edition (CVLT-2)
  - Delis-Kaplan Executive Function System (D-KEFS)
    - Trail Making, Verbal Fluency, Design Fluency, Color Word Interference Test
  - OTHERS: Rey Auditory Verbal Learning Test [RAVLT], Hopkins Verbal Learning Test [HVLT], Rey Osterrieth Complex Figure Test (ROCFT), Wide Range Assessment of Memory & Learning, 2nd edition (WRAML-2), Brief Visuospatial Memory Test, Revised (BVMT-R), Wisconsin Card Sorting Test (WCST); Boston Naming Test; Mini Mental State Exam (MMSE), Montreal Cognitive Assessment (MoCA)

R01MH118514 – (3/4/19 to 1/31/24): National Neuropsychology Network

- Sites
  - University of Florida, c/o Russell Bauer, ABPP-CN
  - Medical College of Wisconsin, c/o David Sabsevitz, ABPP-CN
  - Emory, c/o Daniel Drane, ABPP-CN, David Loring, ABPP-CN
  - UCLA, c/o Robert Bilder, ABPP-CN
- UCLA – coordinating, statistical expertise including:
  - Steve Reise: head of quantitative area, UCLA Psychology; Catherine Sugar, head of Semel Institute Biostatistics Core
- Pearson – collaborative deposition of Q-interactive results into NIMH Data Archive for shared use by NP community
  - Dustin Wahlstrom (Director of Portfolio Management and Delivery - Therapeutics) Kristen Getz (Research Director, Digital Products/Platforms, Clinical Assessment)
- Total Budget: ~$4.4 M/5 years

Table 2. Tests Most Frequently Administered by NNN Sites

Battery or Domain	Test	Total x 4 years	QI		Battery or Domain	Test	Total x 4 years	QI
WAIS-IV	Digit Span	14900	*		General	MOCA	4000	
WAIS-IV	Coding	11140	*		Symptom	Beck Depression Inventory	3700	
WMS-IV	Logical Memory	10300	*		WMS-IV	Verbal Paired Associates	3620	*
WAIS-IV	Block Design	10200	*		Memory	Hopkins Verbal Learning Test	3520	
Language	Boston Naming Test	10200			WAIS-IV	Letter-Number Sequencing	3420	*
WMS-IV	Visual Reproduction	10020	*		Memory	Brief Vis Memory Test-Revised	2920	
Executive	Wisconsin Card Sorting Test	9320			Visuospatial	Facial Recognition Test	2600	
WAIS-IV	Symbol Search	8140	*		General	Mini-Mental State Exam	2000	
WAIS-IV	Similarities	8100	*		Language	WMS-III Mental Control	2000	
WAIS-IV	Matrix Reasoning	7940	*		Language	Test of Memory Malingering	1916	
WAIS-IV	Information	7620	*		Memory	Rey Auditory Verbal Learning Test	1900	
Memory	Rey Complex Figure Test	6420			PVT	Green's Word Memory Test	1640	
D-KEFS	Verbal Fluency Test	6220	*		D-KEFS	Design Fluency Test	1600	*
WAIS-IV	Arithmetic	6140	*		Exec	EXIT25	1600	
WAIS-IV	Vocabulary	6060	*		Symptom	Beck Anxiety Inventory	1500	
D-KEFS	Color-Word Interference Test	5720	*		WAIS-IV	Picture Completion	1440	*
Motor	Grooved Pegboard Test	5500			PVT	Medical Symptom Validity Test	1400	
D-KEFS	Trail Making Test	5420	*		Executive	Symbol Digit Modalities Test	1320	
General	ACS-Test of Premorbid Function	4820	*		WMS-IV	Design Memory	1180	*
Memory	California Verbal Learning Test	4820	*		Achievement	Woodcock Johnson-subtests	1060	
WAIS-IV	Visual Puzzles	4720	*		General	NIH Toolbox	1000	
Motor	Finger Tapping Test	4500			Language	Emory Semantic Fluency Paradigm	800	
Visuospatial	Judgment of LIne Orientation	4120			Language	Columbia Auditory Naming Test	800	
					General	RBANS	800	

Note. QI: * test administered on Q-interactive platform. The rest will be administered via a new, tablet-based/web-based point-of-testing data acquisition program.

Structured Clinical Protocol/ Common Data Elements

- Clinical measures will include structured demographic, diagnostic, and dimensional ratings of key symptoms using instruments proposed as common data elements by the NIMH Research Panel (Barch et al., 2016):
  - Structured History Protocol for Neuropsychology (SHiP-NP)
  - Patient Reported Outcome Measures (Self-Reports)
    - DSM-5 Self-Rated Level 1 Cross-Cutting Symptoms Measure - Adult
    - Patient Reported Outcomes Measurement Information System (PROMIS) Adult Depression Computerized Adaptive Test (CAT)
    - PROMIS Adult Anxiety CAT
    - World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0)
  - DSM-5 Clinician-Rated Dimensions of Psychosis Symptom Severity
  - NINDS CDEs, Neuro-QOL, NIDA Substance Abuse HER Data Elements, NIH Toolbox

Common Data Elements in Neuropsychology

- PI: LUCIA CAVANAGH, PhD: Postdoctoral Fellow in Clinical Neuropsychology at UCLA, General Adult Track, now Chief Fellow
- NAN Clinical Research Grant
- 2018-2019
- Aim 1. Establish Network Infrastructure: By implementing in our clinic, establish technological infrastructure for a shareable, evidence-based clinical protocol and create a model in which other clinics can participate.
- Aim 2. Establish Platform Feasibility and Accessibility Across Settings: Gather feedback on platform content from a nationally-representative group of expert NP consultants

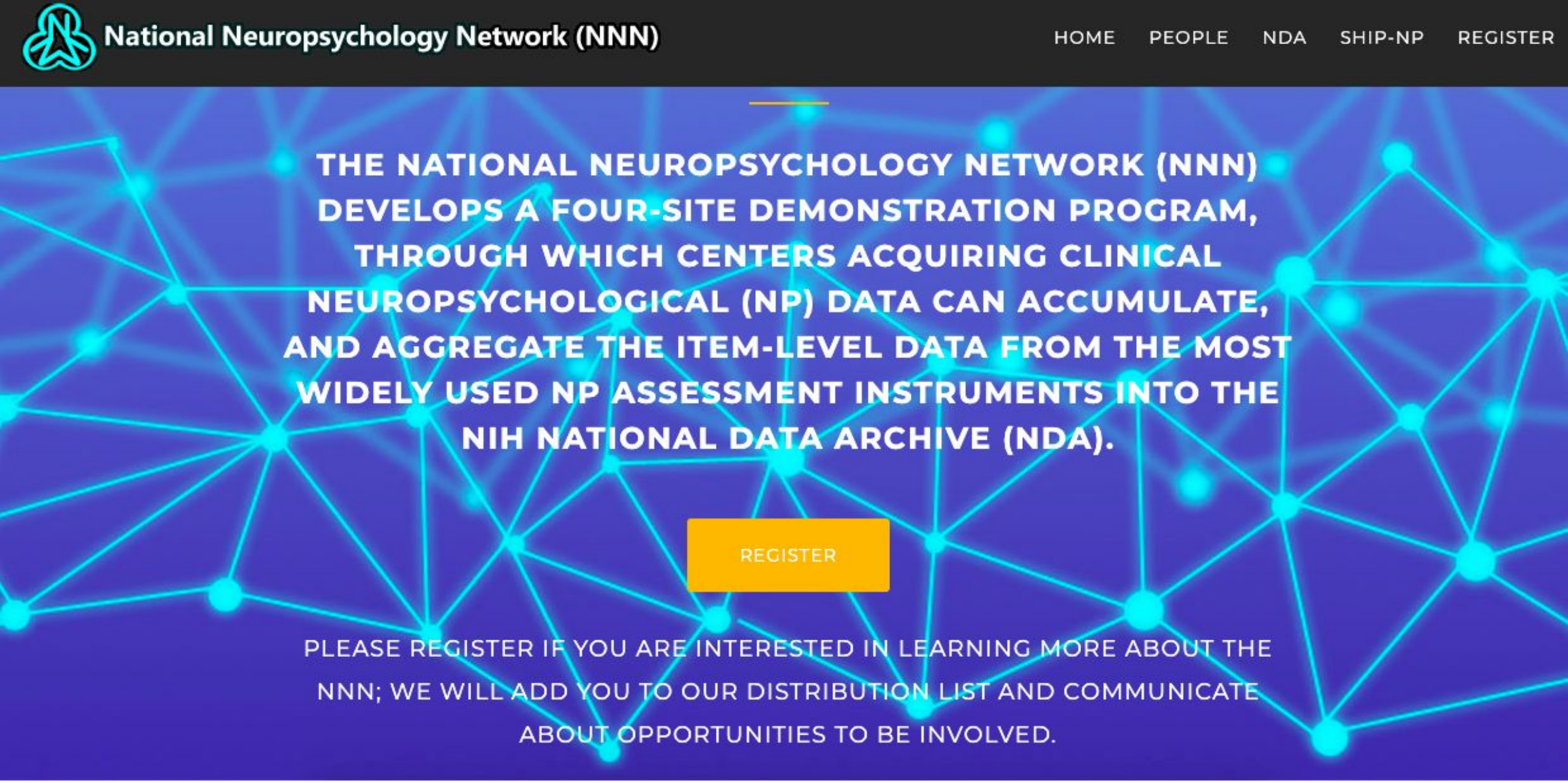
Deliverables: Data

- Collect data on 10,000 cases over 4 years and deposit all item-level data in RDoCdb (enrollment targets are 325 cases per site/year, yielding ~1300 cases/year for the network, or ~5200 cases over the 4-year period of data collection).
- Inclusion/Exclusion criteria:
  - Broad: representative of clinical NP services nationally
  - dementia and degenerative conditions, epilepsies (including psychogenic non-epileptic seizures [PNES]), movement disorders, and other complex neuropsychiatric disorders
  - In all these syndromes, depression, anxiety, or psychotic symptoms are either directly part of the differential diagnosis (e.g., “dementia vs depression”) or the psychiatric symptoms may be critical moderators of cognitive impairment

Table 3. Estimated Clinic Flow for Major Diagnostic Groups						
Condition/Diagnostic Group	**Emory**	**MCW**	**UCLA**	**UF**	**Total Per Year**	**Total x 4 years**
Dementia, MCI, Memory Loss	250	400	280	225	1,155	4,620
Epilepsy	175	75	120	75	445	1,780
Transplant Service, Brief Inpatient Evals	5	10	50	100	165	660
Movement Disorders, Surgical, DBS	150	20	50	250	470	1,880
ADHD/Learning Disability	0	150	50	75	275	1,100
Traumatic Brain Injury	20	750	50	100	920	3,680
Neoplasm, Stroke	50	150	50	50	300	1,200
Primary Psychiatric	55	0	50	25	130	520
TOTAL	**705**	**1,555**	**700**	**900**	**3,860**	**15,440**

Deliverables: Results

- Evidence-based battery selection – this includes adaptive test selection within batteries of tests, to determine which test in the battery provides the highest predictive power for selected differential diagnostic applications, given prior test results
- Computerized adaptive tests – including adaptive item selection within tests, given prior item results, to provide measurement of specific traits with prescribed levels of precision
- Fixed short-forms of tests that increase efficiency of testing even when adaptive testing is not practical
- Analyses will examine test operating characteristics, sensitivity, specificity, positive and negative predictive power of both original and new measures to aid in differential diagnosis of neurocognitive disorders and major psychiatric syndromes
- Establish a testbed for evidence, enabling future measures to be examined directly for equivalence or superiority



https://www.sistat.ucla.edu/NNNWeb/index.html

SHiP-NP
Demo

	A	B	C	D	E	F	G
1	PRE-ASSESSMENT SUMMARY SHEET						
2							
3	Name:				Tested:		
4	MRN#		Examiner(s):	--	Education:		
5	DOB:	--	Age:	--	Supervisor:		
6	Gender:	--	Primary ICD-10:	--	Hand:		
7							
8	ShiP-NP						
9	Form completed by:	*Patient/Other Name*	*Other realtionship*		*## hrs/week spent together*		
10	***Clinical Concerns***						
11	☐	**Bilingual in English and LANGUAGE**					
12	☐	**Endorsed use of DRUG over the past two weeks**					
13	☐	**Currently taking MEDICATION TYPE**					
14	☐	**INFORMANT NAME holds conservatorship for this patient**					
15	☐	**Prior neuropsychological evaluation completed on DATE**					
16							
17	Patient Questions:	*Text*					
18							
19	*Demographics*						
20	Place of Birth:	*MM/DD/YYYY*		Generational Status:	*Born in U.S./Born outside U.S./etc*		
21	Marital Status:	*Married/Single/Divorced*		Language Fluency:	*English*	*L2*	*L3*
22	Sex:	*Male/Female*		Age of Acquisition:	##	##	##
23	Handedness:	*Right/Left*		Education in L2:	--	*Formal edu completed*	
24	Ethnicity:	*(Not) Hispanic/Latino*	*Specifier if Y*	Proficiency:		##	
25	Race:	*Asian/Black/White/etc*	*Specifier if Y*				
26							
27	*Prior Neuropsychological Testing*						
28	Date:	*MM/DD/YYYY*	Provider:	*Name*	Location:	*Institution Name*	
29							
30	*Educational History*						
31	Highest level completed:			College/Major:		Performance:	
32	Grade Failures:	*Y/N*	*Grades failed, if Y*		*Reason*		

ShiP-NP Clinician Output Report

Report prepared for: Jane Doe, Ph.D. **PtID:** 12345 **DOT:** 04/24/2019

CLINICAL NOTES:

- Patient is fluent in English and French.
- Patient endorsed use of Marijuana for several days over the past two weeks.
- Patient reportedly completed prior neuropsychological testing on 09/2010 by Dr. Smith at Cedar's Sinai Hospital.

Name: Jane Doe
MRN: ***
DOB: 01/01/1989
Age: 30
Referral Source: ***
Providers: Jane Doe, Ph.D.

Referral Diagnosis: ***
ICD-10 Code: ***

Dates of Service:

*** *96116, Neurobehavioral status exam with psychologist, first hour*
*** *96138, Neuropsychological testing with technician, first 30 min*
*** *96139, Neuropsychological testing and scoring with technician, additional 30 min*
*** *96132, Professional integration of patient data, first hour*
*** *96133, Professional integration of patient data, additional 30 min*

Reason for Referral:

Ms. Doe is a 30-year-old, right-handed, White, female with a medical history of epilepsy, head injury, diabetes, high blood pressure. She was referred for neuropsychological testing by *** for assessment of her cognitive and emotional functioning and to assist with treatment planning.

Next steps -- on to the Future

- Expand data elements/tests to include both English & Spanish, over time add other languages
- National NP Network in the USA could serve as model for ex-US development
- Modern psychometric specs critical for alignment with test characteristics in other languages and cultures
  - For this – various methods to identify invariance including DIF, “harmonization” and “phenotype alignment” may help
- Ideal – a global bank of methods to be shared freely, used to expand access to high quality NP services and reduce health disparities, and increase knowledge about human health and disease in the broadest sense

UCLA Semel Institute
Tennenbaum Center for the Biology of Creativity

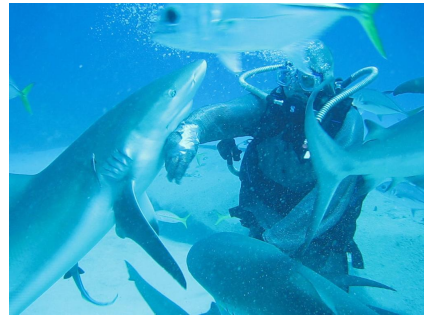
Many thanks!

Consortium for Neuropsychiatric Phenomics (52 investigators); Investigators in current RDoC projects, and Whole Genome Sequencing in Psychiatric Disorders (WGSPD; Freimer et al.). Special thanks to Steve Reise, Catherine Sugar, Gerhard Helleman, Ariana Anderson, and Max Mansolf for measurement issues, and to the NNN PIs: Rus Bauer, Dan Drane, David Loring, Lauren Umfleet, and Dustin Wahlstrom.

Supported by NIH Grants: (CNP) UL1-DE019580, RL1MH083268, RL1MH083269, RL1DA024853, RL1MH083270, RL1LM009833, PL1MH083271, and PL1NS062410; (Cognitive Atlas) RO1NS061771; (Multilevel WM/RDoC) R01MH101478; (Modeling/RDoC) R03MH106922; (WGSPD) U01 MH105578; R01MH (NATIONAL NEUROPSYCHLOGY NETWORK).

NATIONAL INSTITUTES OF HEALTH

John Templeton Foundation



rbilder@mednet.ucla.edu

http://www.semel.ucla.edu/creativity

http://healthy.ucla.edu