Psychological Research with Secondary Aging Data

Jack McArdle, Psychology USC, May 2008

Overview

1. Definition of "Secondary" Data
2. Key Principles of Any Data Analysis
3. Existing Data Benefits and Limitations
4. A Recent Example for the NGCS+HRS
5. Final Suggestions

1. Defining Secondary Analyses

On "Cumulative" Research

- Since about 1970, there has been a growing effort to accumulate information from a large number of empirical research studies.

- There was broad dissatisfaction with the fact that most individual studies of small size, and low accuracy, but there are so many studies that more can be done.

- There was a need for an more formal and organized accumulation of these studies and a recognition of "replicable results" in the "soft-fact" sciences

There are many ways to effectively cumulate evidence

- “Secondary Analysis” of raw data
- “Meta Analysis” of summary data
- “Mega Analysis” of multi-group data
- Some novel methodological tools have become widely available for these kinds of analyses.

Secondary Data Analysis Defined

Definition from *Wikipedia* (2008-05-20):

“In research, Secondary data is (sic) collected and possibly processed by people other than the researcher in question. Common sources of secondary data for social science include censuses, large surveys, and organizational records.”

In sociology, primary data is data you have collected yourself and secondary data is (sic) data you have gathered from primary sources to create new research.”

Used primarily in Sociology, Education, Political Science, and Astronomy.

"Secondary Analysis" Procedures

- Extract previously collected raw data from some other research study or survey.
- Formulate a research question in terms of the focal constructs, and relate these to the available variables and subjects in the available data set
- May need to use unique sampling strategies of subjects within the data set to create a meaningful sample for this sub-study
- Re-analysis of existing data adds to overall body of scientific knowledge.
- *"You cannot analyze a database, but you can analyze a question using a database!"* (McArdle & Horn, '02)

Secondary Analysis Sources

- Over 1,000 studies of the *GSS*, *NES*, *PSID*
- Over 500 studies of the *NELS*
- Over 300 studies of the *NLSY*
- Over 25 studies of the *WAIS* and *WAIS*-*R*
- More Recent Data Sources – *HRS*, *MIDUS*, *NSHAP*

Psychological Barriers?

- The term chosen for the primary analysis of existing data – “secondary data analysis” – looks like it means “after the fact” and this does not instill a great deal of confidence.

- Psychology initially developed as a laboratory science, much like Biology, so it is no surprise that this discipline did not initially embrace other social science paradigms – e.g., (a) causal inference from observational data or
  (b) letting others collect data for you.

- At the same time, psychologists are keen to recognize the need for broad theory statements with \repeatable predictions. So come concerns are reasonable and *self-imposed experimental rigor* is essential.

2. *Principles of Validity In Any Experiment*

Principles of "Any" Data Analysis

1. The most basic principle of experimental design – "We want to (a) *maximize variation* with respect to what we are interested in, and (b) *minimize variation* with respect to what we are not interested in" (Thurstone, 1947).

2. Randomization to assigned treatments balances out other confounds, and makes stat-model assumptions correct in the long run (Fisher, 1929). Note that treatments do not necessarily lead to mean differences only.

3. Be selective -- Focused and limited hypotheses lead to most precision and power (Bock, 1975).

Principles of "Any" Data Analysis

4. *Samples of individuals* (subjects/participants) need to be representative for results to be generalizable. Given representativeness, larger sample sizes lead to more precision. Diversity of sample may be essential.

5. *Measurement makes a difference*, and not all measurement is adequate (some is poor!). The standards of measurement are well-developed and can be used to set standards of evidence.

6. The *situation of measurement* needs to be well defined, clearly understood, and repeatable, or inferences and generalizability will be limited.

Principles of “Any” Experiment

- A successful analysis of any “experiment” starts with the investigator taking “personal responsibility” of the data.

- Analyst must attempt to put themselves in the position of the data collector and carefully explain this to a reviewer.

- Any analysis must be done recognizing data are limited and the limits of the data collection will limit the causal/control inferences possible.

- Any analyst must be *self-critical* and not count on any data collection to have been perfect for their own problem.

Threats in "Any" Experiment

Anyone trained as an Experimental Psychologist will appreciate the precision offered by this methodology.

Anyone who carries out randomized Experiments in a laboratory setting know something always goes wrong.

Anyone can be trained to carry out analyses of large sets of existing data -- especially important in aging research.

We soon find a convergence of the problems of data analysis, including the use of similar stat models and the goal of inferences about "control."

Better term → "Primary Analysis of Existing Data"

3. Benefits and Limitations of Existing Data

Top Ten BENEFITS of Existing Data

1. Low cost expenditure of time and money for the analyst.

2. Avoids potential problems created by poor interview plans and data collectors.

3. Typically data collection was created by a thoughtful team of researchers, so it meets IRB standards.

4. Representative and large samples may be available.

5. Diverse sample may be available as well.

Top Ten BENEFITS of Existing Data

6. Ecological validity may be increased due to a more "real-life" environment of data collection.

7. Analyst can be creative and evaluate hypotheses generated in the laboratory with a great deal of precision

8. Can explore hypotheses to set the stage for future studies in laboratory settings with more controls.

9. Demonstrates the analyst can actually analyze focal ideas using real data.

10. Direct comparisons possible with other analyses.

Top Ten LIMITATIONS of Existing Data

1. Sample of Participants may not be of focal interest.

2. Measurements may not represent the Constructs of focal interest – possible biggest problem to overcome.

3. Conditions of measurement may not be well controlled or completely understood.

4. Representative and large samples may be available but sampling schemes may be complex and difficult to use.

5. Data documentation may not fully describe all problems found in data collection.

Top Ten LIMITATIONS of Existing Data

6. Ecological validity and generalizability may be decreased due to "real-life environments"

7. Complex (i.e., longitudinal) data collections may not be easy to understand (i.e., dropout reasons).

8. Actual conditions of measurement may not be well controlled or completely understood.

9. Representative and large samples may be available but sampling schemes may be complex and difficult to use

10. Other researchers can work on the same topic and not know it – publication priority is never assured.

4. A Recent Example from NGCS+HRS

Cattell (1941) & Horn's (1967) theory of Cognitive Changes

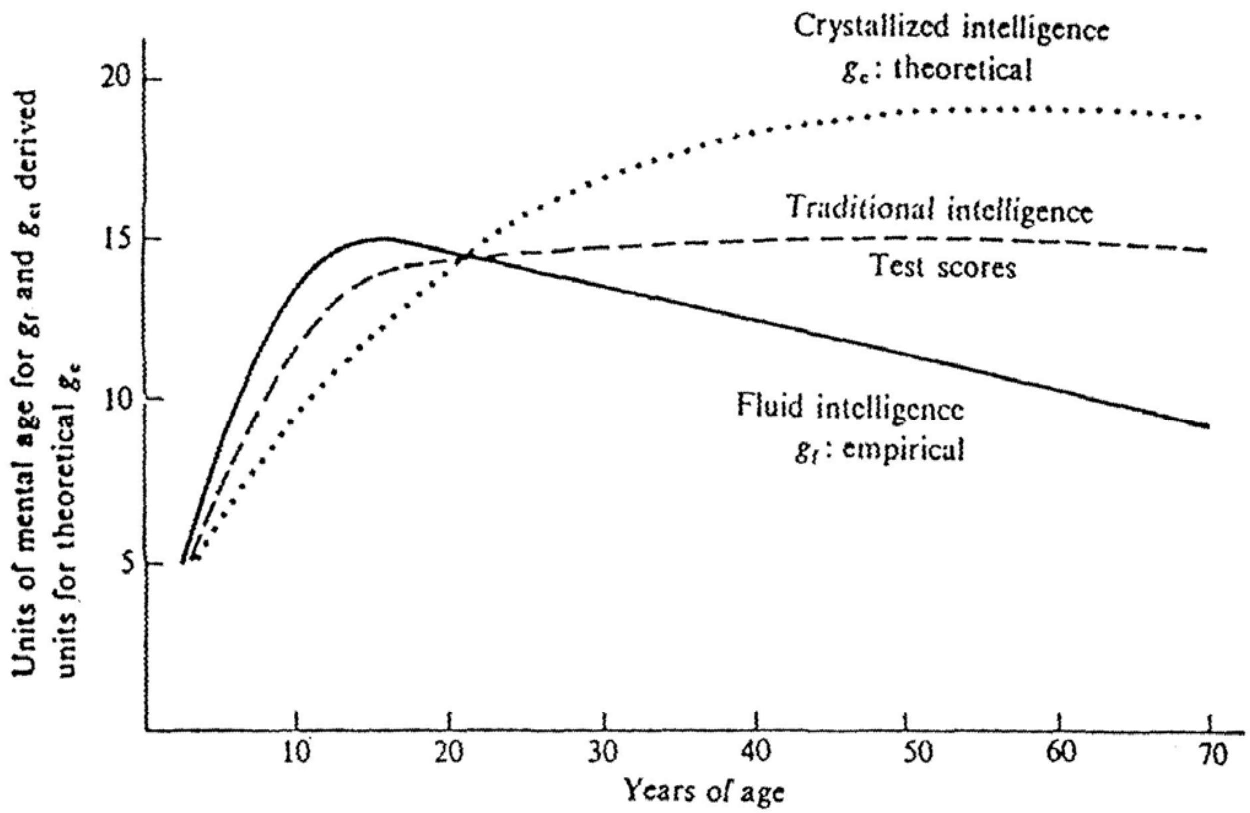


FIG 2.2. Age plots of the investment theory of fluid into crystallized intelligence. Age: 5, 15, 25, 35, 45, 55, 65, 75; g_f values: 9, 15, 14, 13, 12, 11, 10, 9; g_c values: 6.7, 13.0, 16.2, 17.9, 18.8, 19.3, 19.5, 19.6; g_{ct} traditional intelligence test curve.

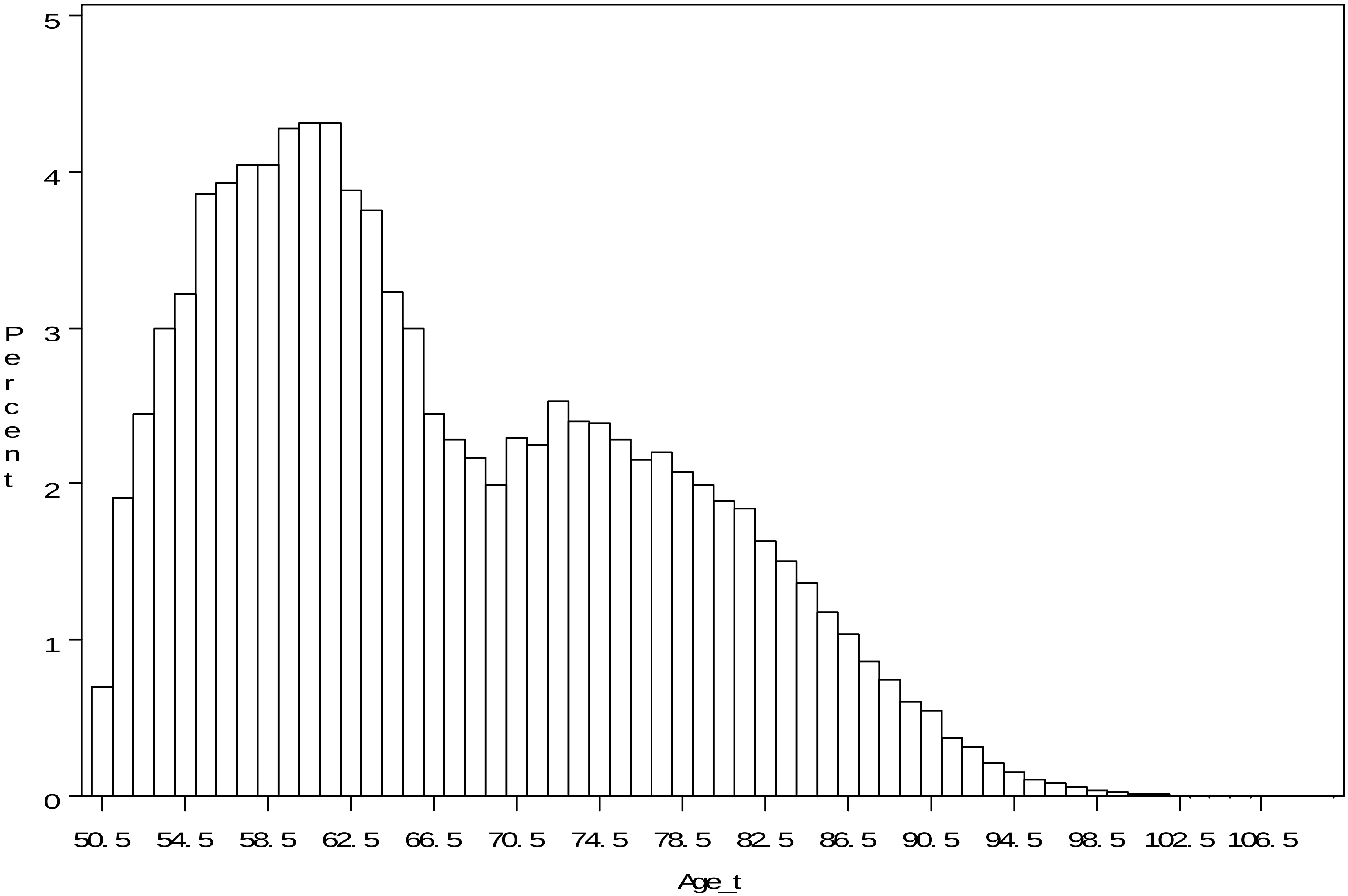
Data Collections part of the US National Growth and Change Studies

1. The *Mega-WAIS Study* (1980) – *N*>5,000 Wechsler Adult Intelligence Scale (WAIS) scores from 12 existing longitudinal archives brought together as one collective.

2. The *Bradway-McArdle Longitudinal Study* (1984) – *N*=111 individuals who were tested at ages 2-7 and seven time until ages 72-77 years old. Retesting now going on in 2008.

3. The *Woodcock-Johnson Retest Study* (1998)

Ages at All Testings in Current HRS (from McArdle, Fisher & Kadlec, 2007)



Current HRS In-Person and Telephone Cognitive Measures

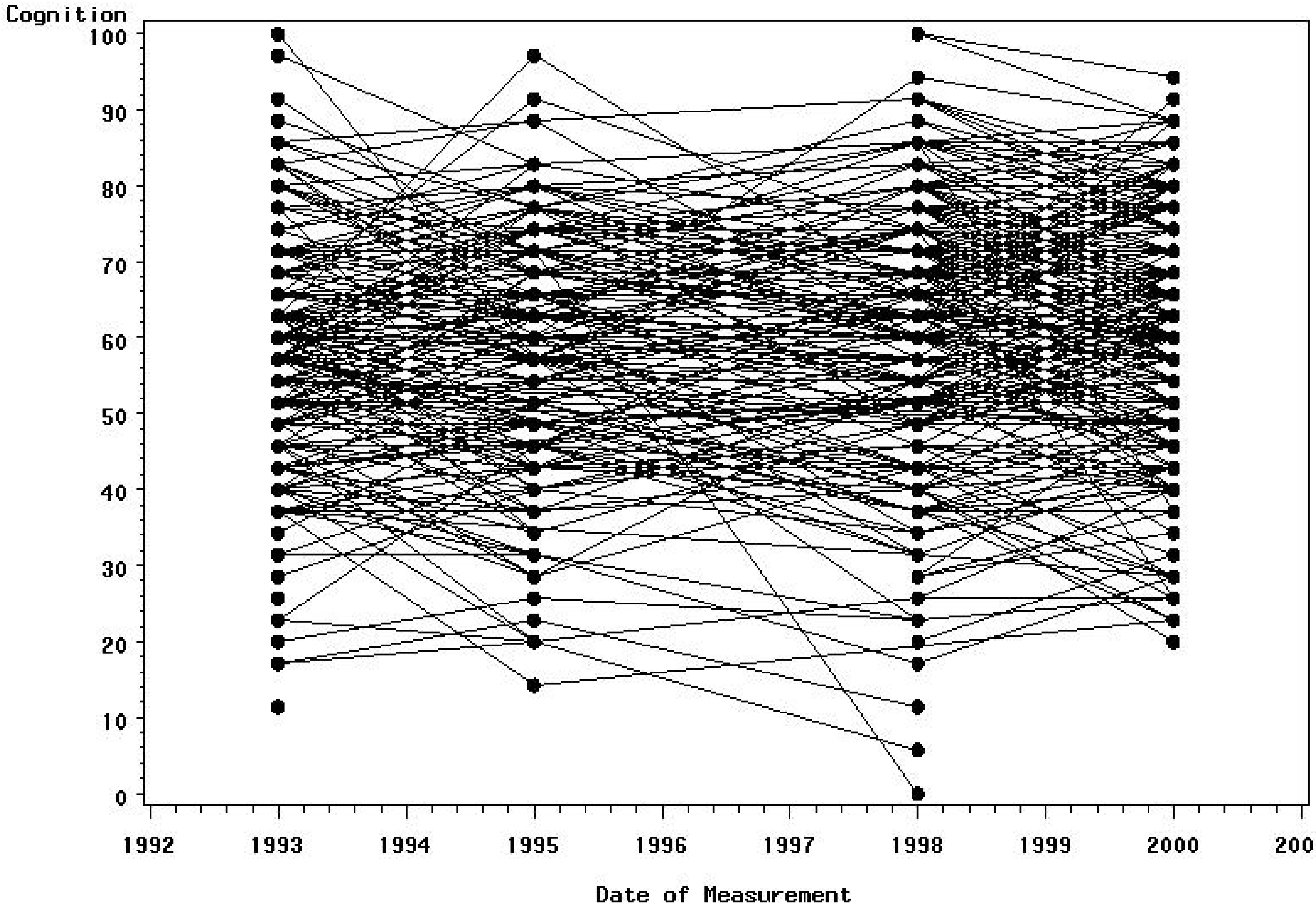
1. Immediate Word Recall (10 items)
2. Delayed Word Recall (10 items)

3. Serial 7s (to assess working memory)
4. Backward Counting (starting with 20 and 86)
5. Dates (Today's date and day of the week)
6. Names (Object naming, President/VP Names
7. Incapacity (to complete one or more of the basic tests)

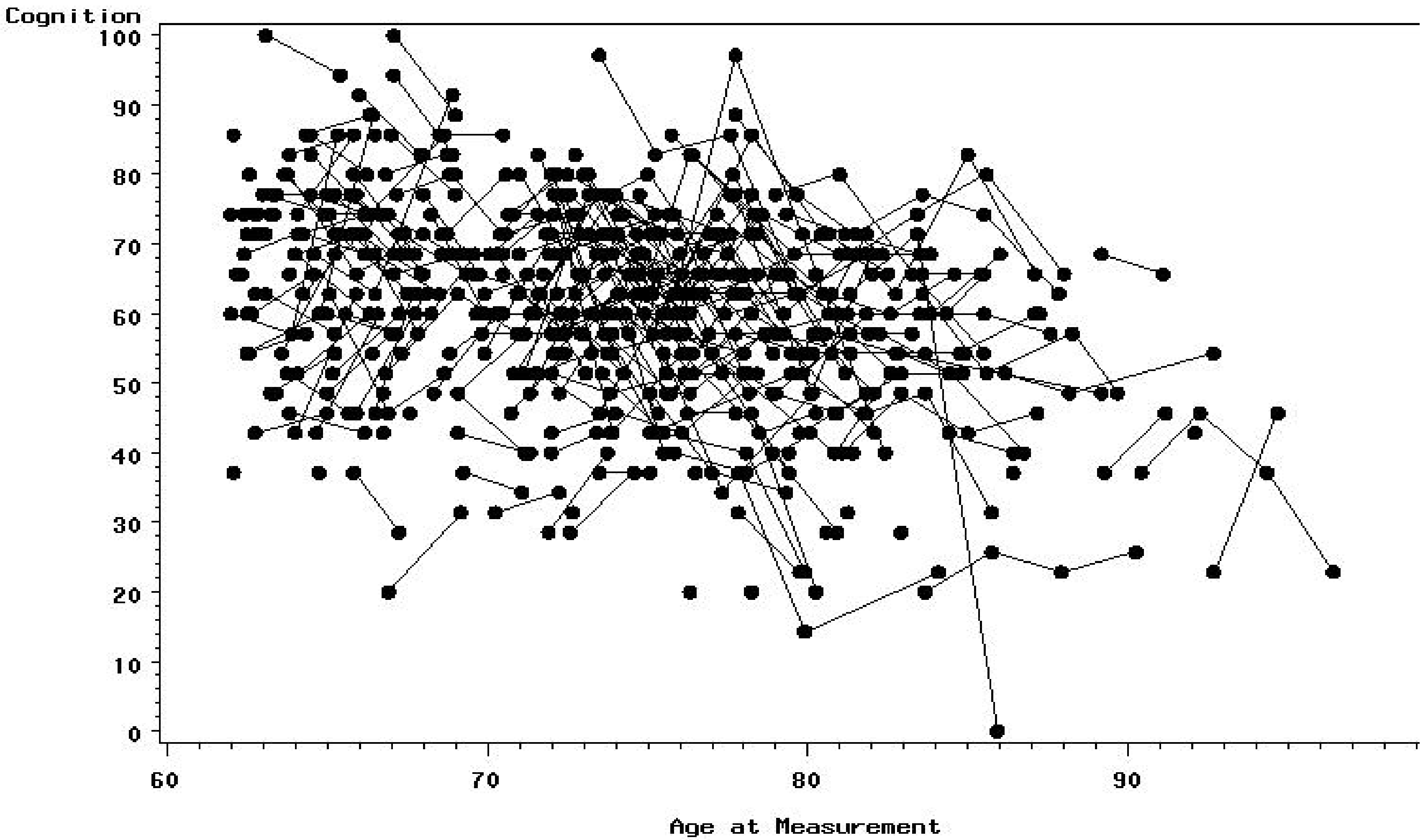
And at some occasions …

8. Vocabulary (adapted from WAIS-R for T>95)
9. Similarities (adapted from WAIS-R for T= 92, 94)
10. Newly created "Adaptive" measures from the WJ-III

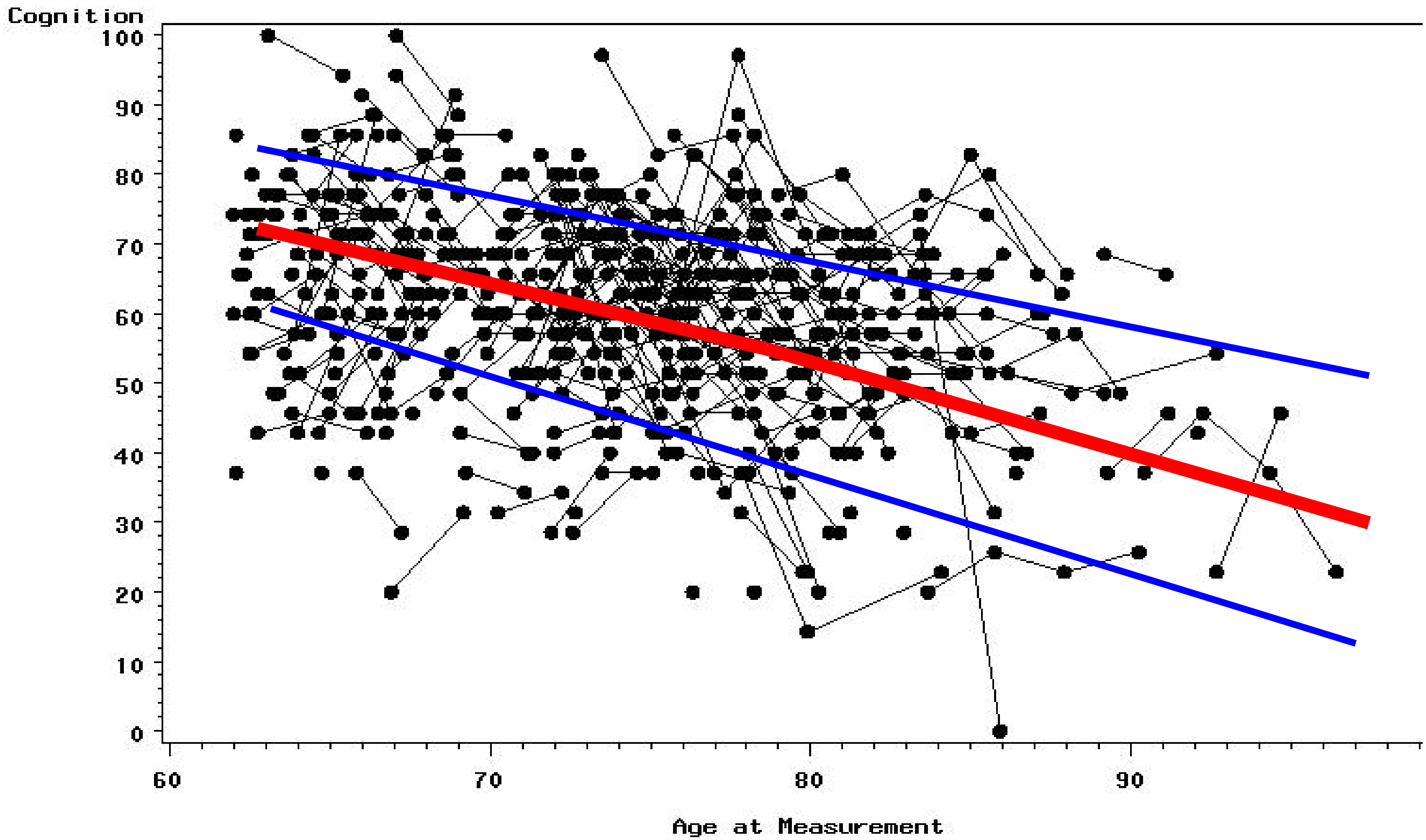
Figure [2b]: HRS Cognitive Data over DATE



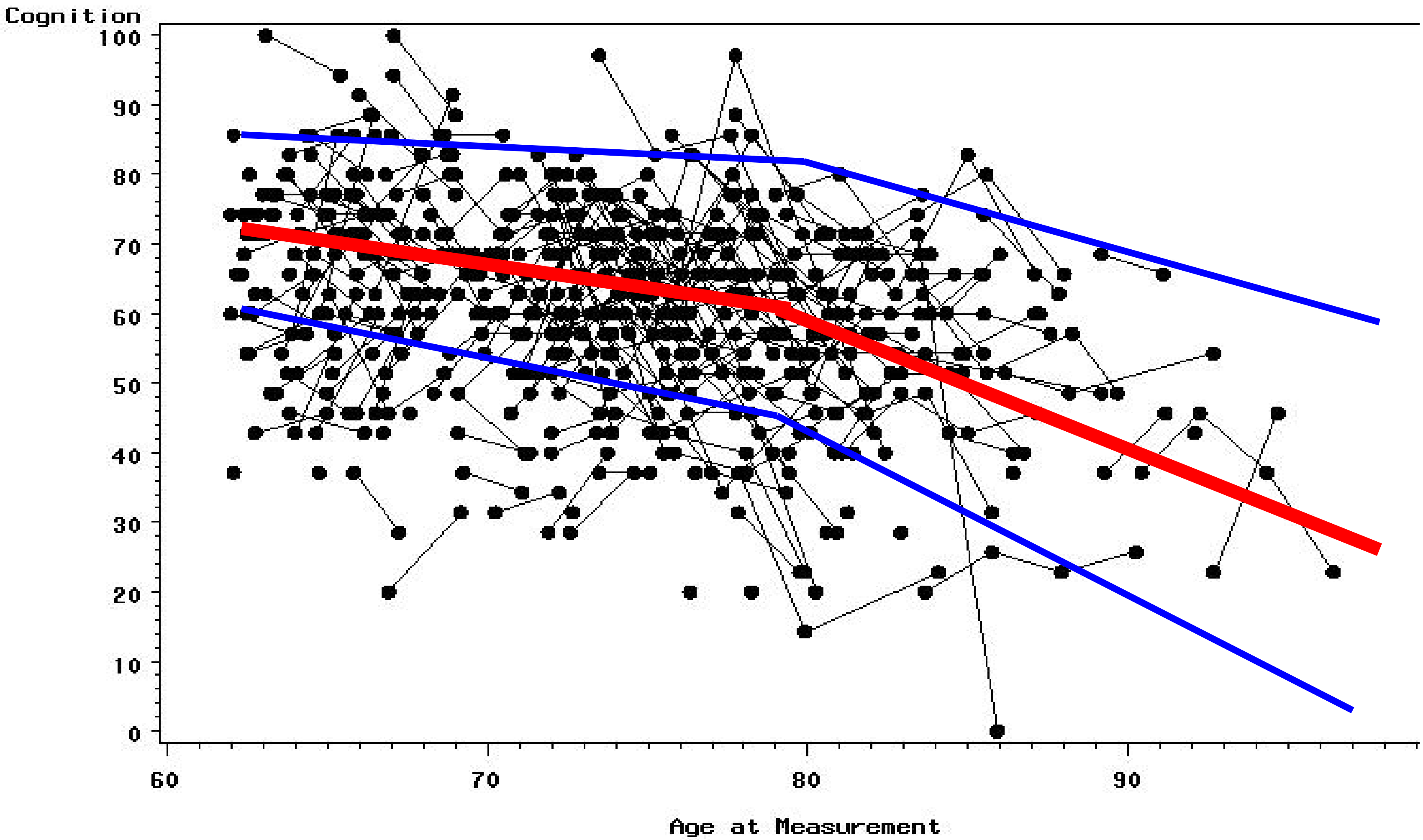
HRS Cognition Scores given AGE of Measurement (N=14,250; D=32,665; T=1-4; Only a sample of data are drawn here, and outliers were excluded)



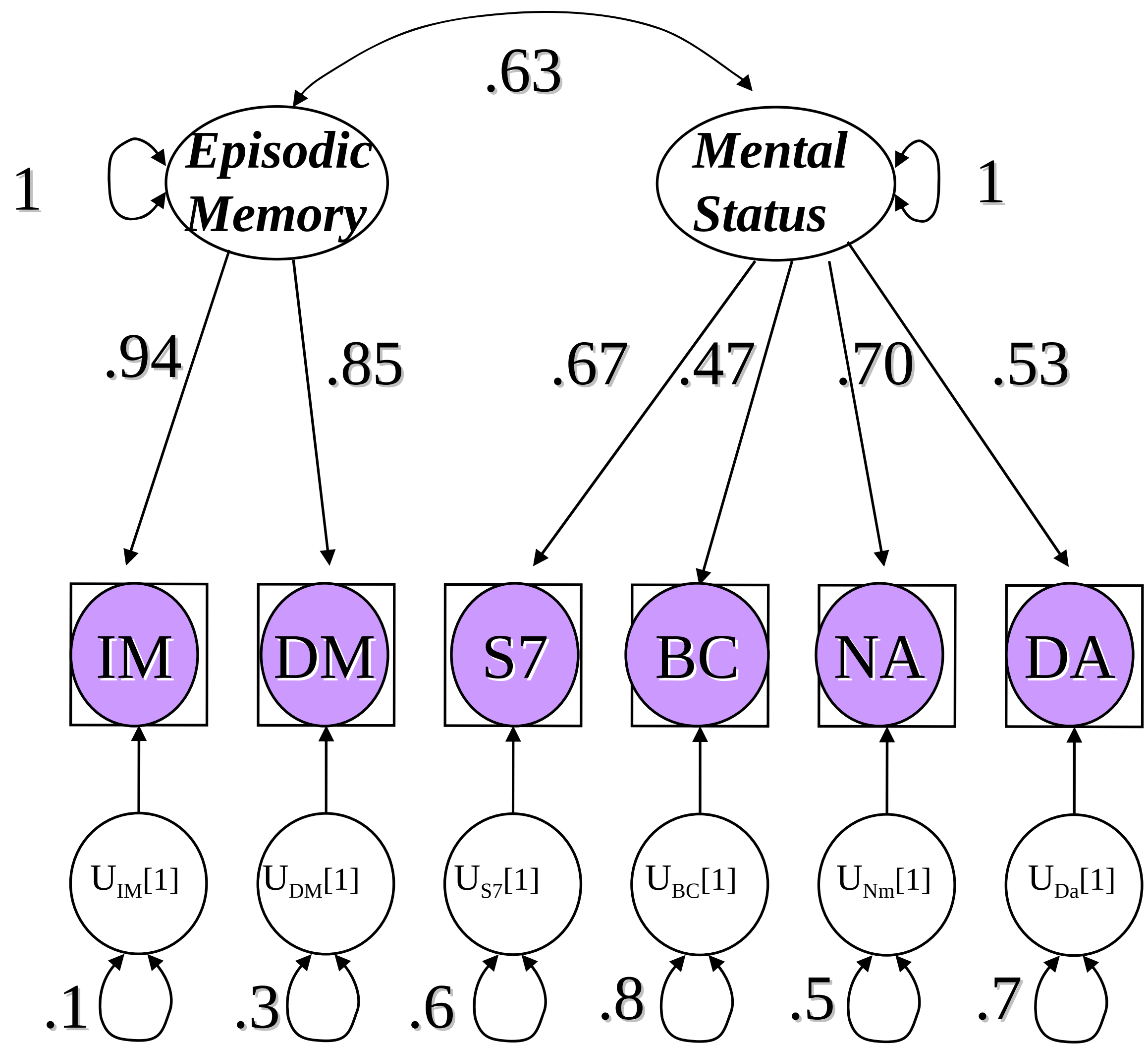
Expected HRS Cognition Scores given AGE of Measurement (N=14,250; D=32,665; T=1-4; Only a sample of data are drawn here, and outliers were excluded)



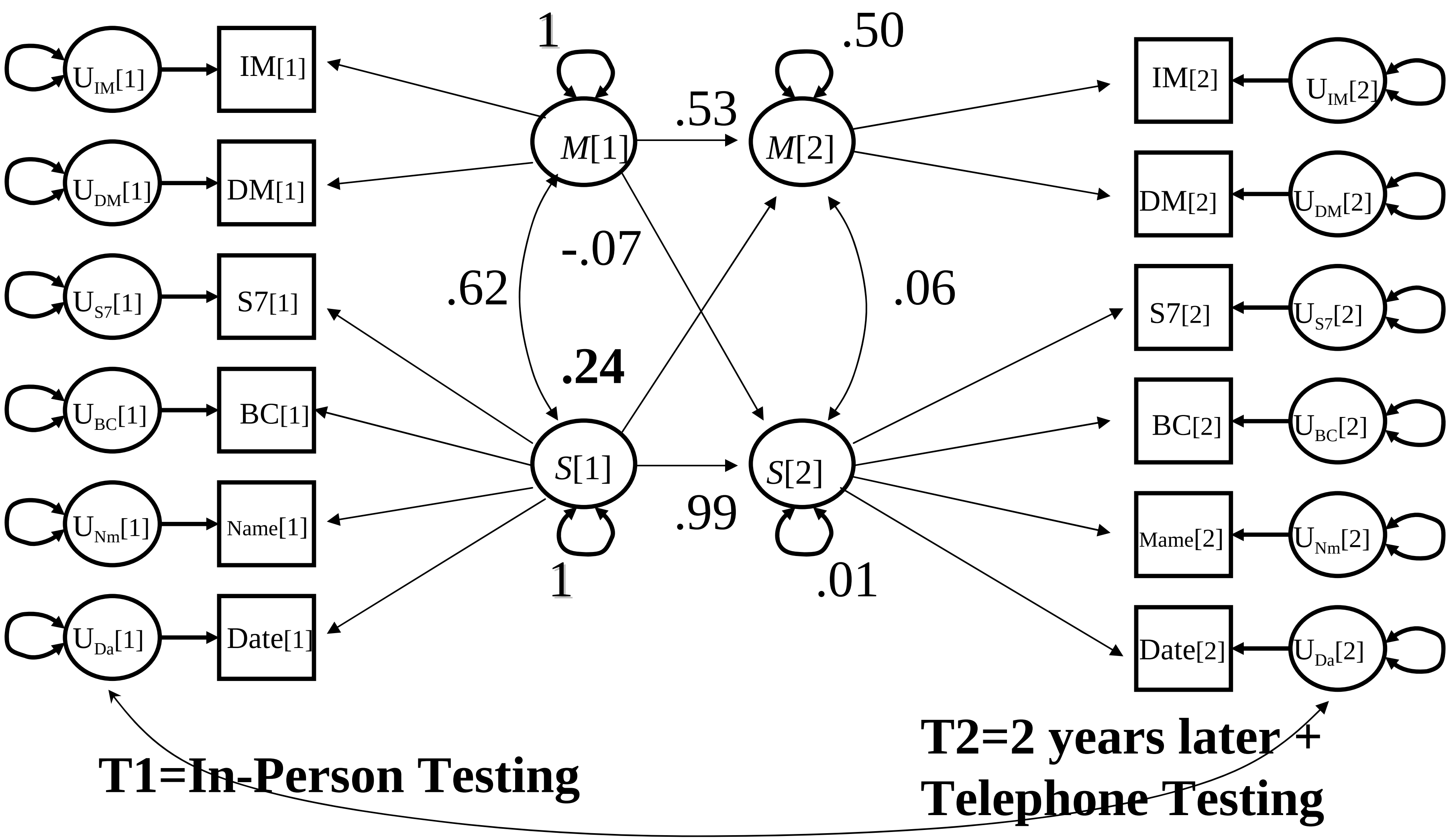
Nonlinear Changes in HRS Cognition Scores given AGE of Measurement (N=14,250; D=32,665; McArdle et al, 2007)



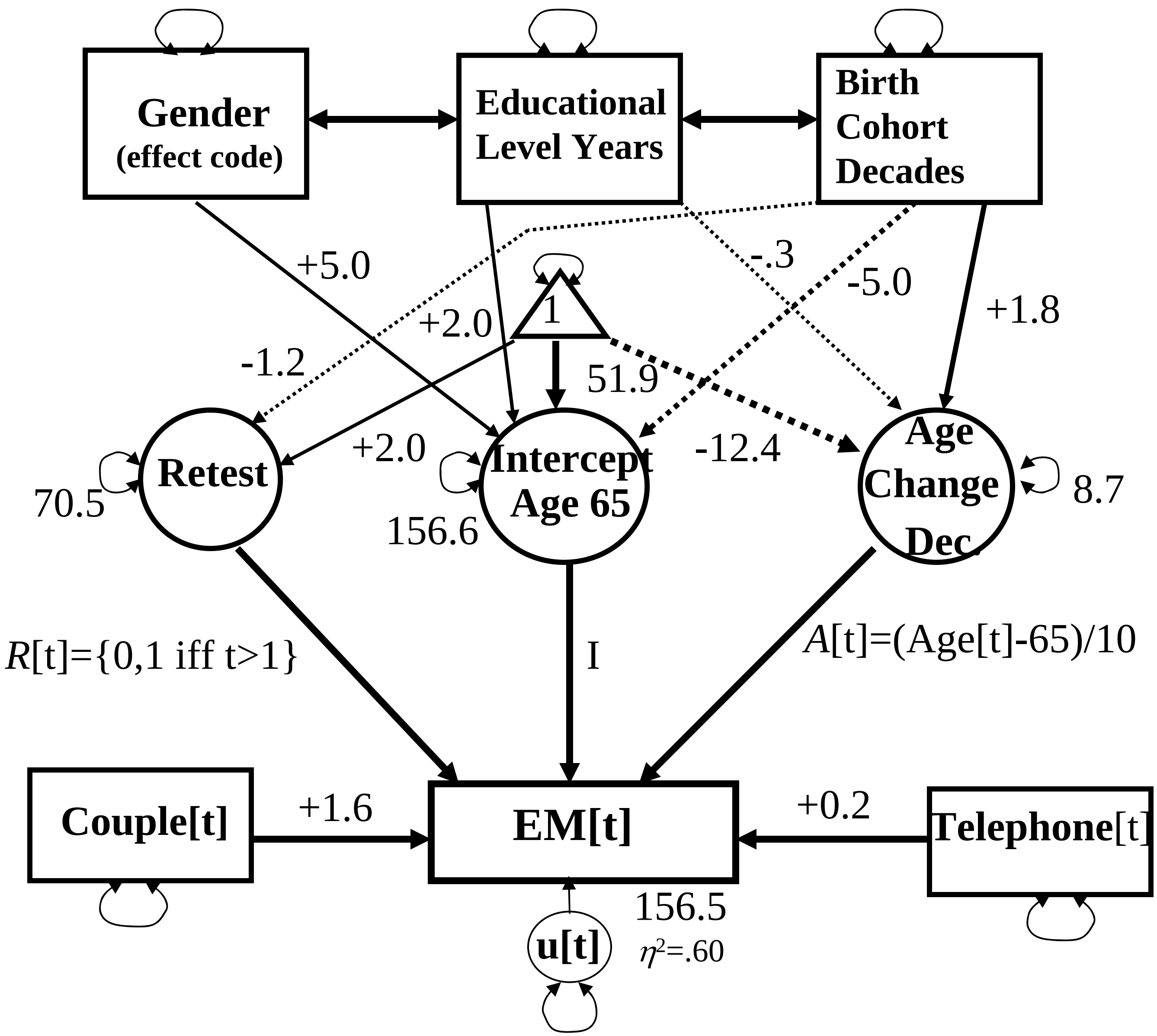
At least Two Useful Cognitive Factors are well measured by the current HRS measures



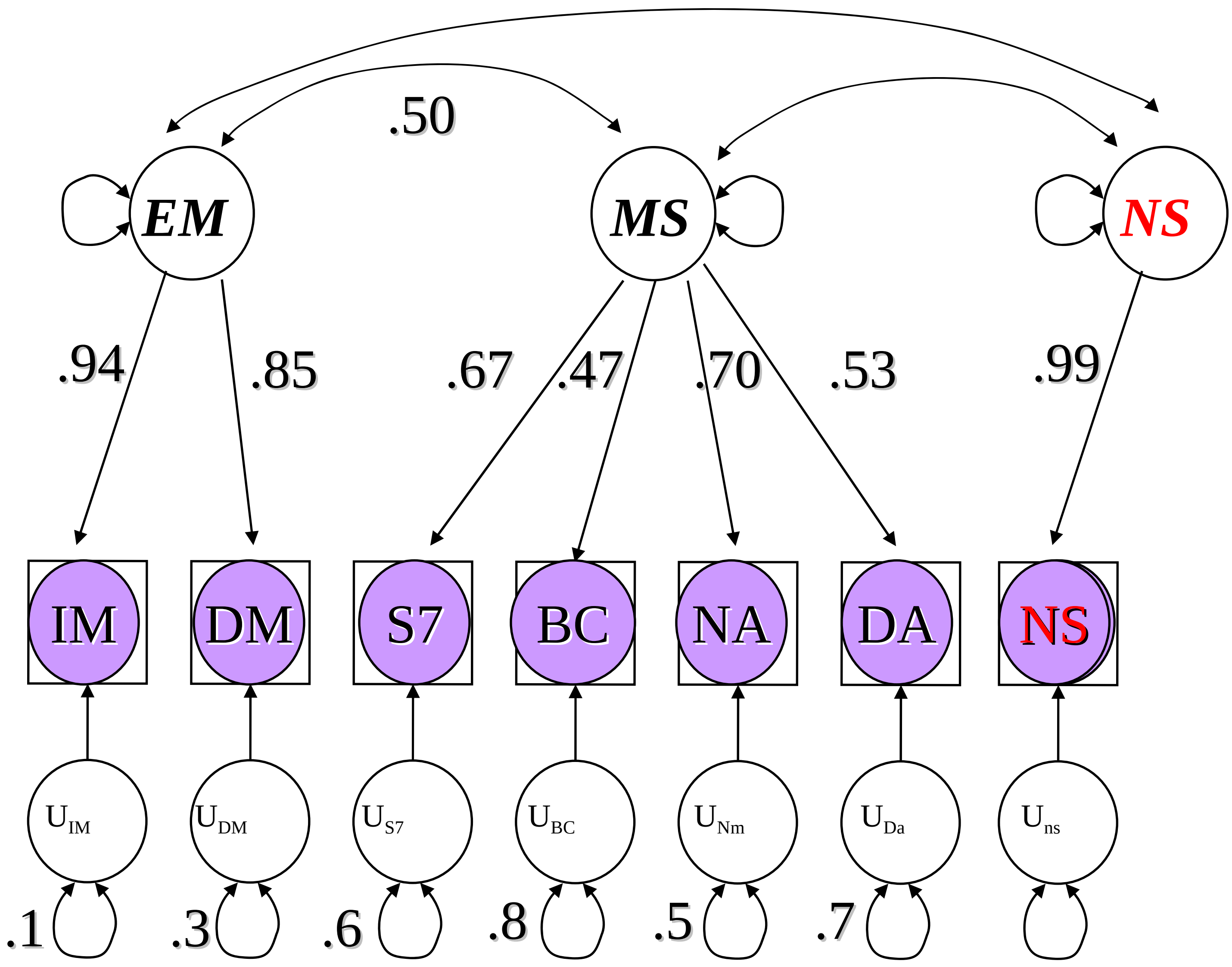
The HRS Cognitive Measures are INVARIANT Over Time and Mode-of-Testing



Longitudinal changes in Episodic Memory (EM[t]) related to demographic indices (McArdle et al, 2007)



Adding to the HRS to Measure Additional Cognitive Abilities



5. Final Suggestions

Review Panels Not Yet Convinced?

- There is a growing recognition for the virtues of secondary/existing data analysis, BUT the NIH/NSF review panels still offer a majority of grants to *new* data collections.

- If we examine how much data that are collected are actually used in published analyses, we find ~ 5% – so more and newer data are not always needed.

- This unfortunate cycle needs to be broken, and more resources need to be set aside for analyses of existing data.

Just Do It!

- There are very few barriers to the analysis of existing data, and these barriers are becoming less every day.

- A big problem we now face is that we have come to realize we do not know how to analyze complex data problems, even if the data are handed to us.

- The analysis of existing data should be a formal requirement before collecting new data on any individual.

- Helpful Hint – think of the question in advance of the data collection -- *"We cannot analyze a database, but we can analyze a question using a database!"*