MEASURING COGNITION IN A MULTI-MODE CONTEXT

Mary Beth Ofstedal, Colleen McClain & Mick P. Couper

With updated results from Jessica Faul and colleagues

Institute for Social Research
University of Michigan
February 2021

HRS HEALTH AND RETIREMENT STUDY

Motivation

- Interviewer-administered longitudinal surveys increasingly incorporating a web option
    - Budget pressures, respondent convenience, changing survey environment, pandemics, …
- Raises concerns for measuring complex constructs, among populations that may have difficulty responding, and over time
- Our focus: **Measurement of cognitive ability in a longitudinal study of older adults with mixed mode data collection**

Surveys With Cognitive Measures

- Interviewer-administered
  - Berlin Aging Study
  - English Longitudinal Study of Ageing
  - Survey of Health, Ageing an Retirement in Europe
  - Household, Income and Labour Dynamics in Australia Survey (HILDA)
- Self-administered
  - *Understanding Society* (CASI and web)
  - Understanding America Study (UAS)
  - UK Biobank
- Both interviewer and self-administered
  - Health and Retirement Study (HRS)
  - Army Study to Assess Risk & Resilience in Servicemembers (STARRS)

Mode Features That May Affect Measurement Of Cognition

- Presence vs. absence of interviewer
  - Social desirability
  - Motivation/focus
  - Understanding
  - Test anxiety
  - Time pressure
  - Use of aids (calculator, Google search, etc.)
  - Interviewer compliance with protocol
- Medium of communication
  - Presentation of material (visual vs. oral/aural)
  - Delivery of response (oral vs. computer/tablet entry)

Other Considerations

- Potential differential effects by:
  - Age
  - Education
  - Cognitive ability
  - Computer literacy
  - Physical and/or sensory impairments
  - Etc.

Existing Mode Comparisons

- Only a few studies of mode effects on cognitive measurement
    - **Runge, Craig, & Jim (2015)** administered word recall tests from HRS on the web to female sample from a pre-existing panel, compared results
    - **Gooch (2015)** assessed differences in "wordsum" vocabulary tests via experimental lab study to web or FTF administration
    - **Al Baghal (2017)** examined differences between two self-administered modes (CASI and web) for several cognitive measures in *Understanding Society* (IP7)
- All find some differences, typically better performance on web
- None are placed in a longitudinal context

Research Questions

- What are the implications of mixing modes for measurement of cognitive performance in a longitudinal setting?
    - Do item missing data rates differ by mode?
    - Do these tests yield equivalent descriptive results across modes?
    - Can we measure change over time?
    - Can we make consistent multivariate inferences about cognitive ability?

DATA AND METHODS

Health and Retirement Study (HRS)

- Panel study of people age 51+ in the U.S.
- Began in 1992
- Study provides information on employment, physical and mental health, access to and use of health services, financial status, family support
- Funded by the National Institute on Aging (NIA U01AG009740) and the Social Security Administration

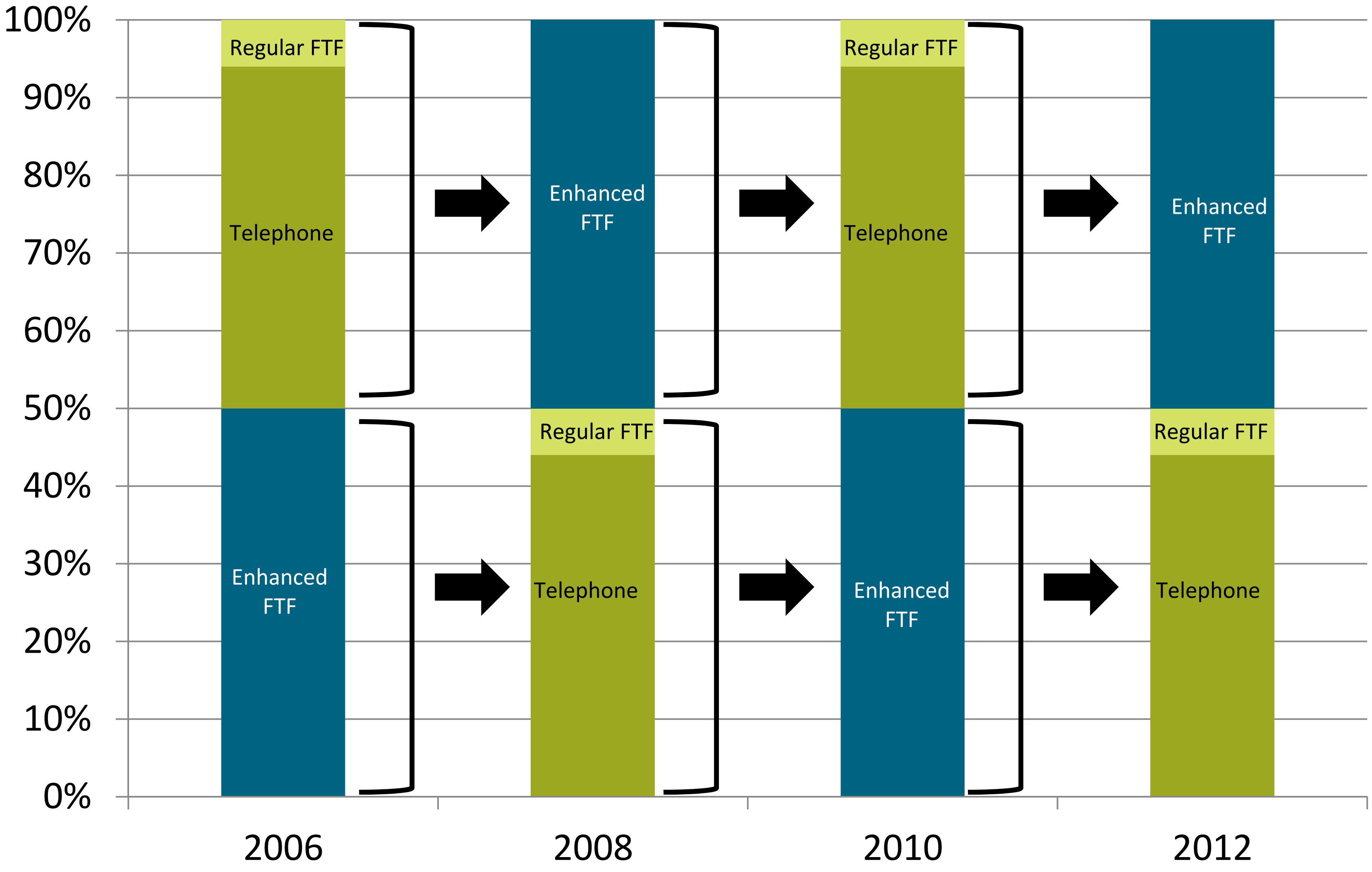
HRS (Cont.)

- Core interviews conducted with ~20,000 participants every 2 years
- Supplemental surveys (via mail, web) in between core interviews
- Sample refreshed every 6 years with cohort age 51-56
- Physical measures, biomarkers and psychosocial self-administered questionnaire added starting in 2006 (enhanced FTF interview)

HRS Multi-Mode Design

- Prior to 2004, telephone was primary mode
  - FTF for baseline only, TEL for follow-up waves
- In 2004, mostly FTF
  - To update Social Security linkage consents
- From 2006 on: half and half
  - Half of sample assigned “enhanced” FTF interview, other half TEL or regular FTF (80+)
  - Assignment flips in next wave; a given R gets enhanced FTF every other wave and TEL/regular FTF in between

Enhanced FTF Sample Rotation

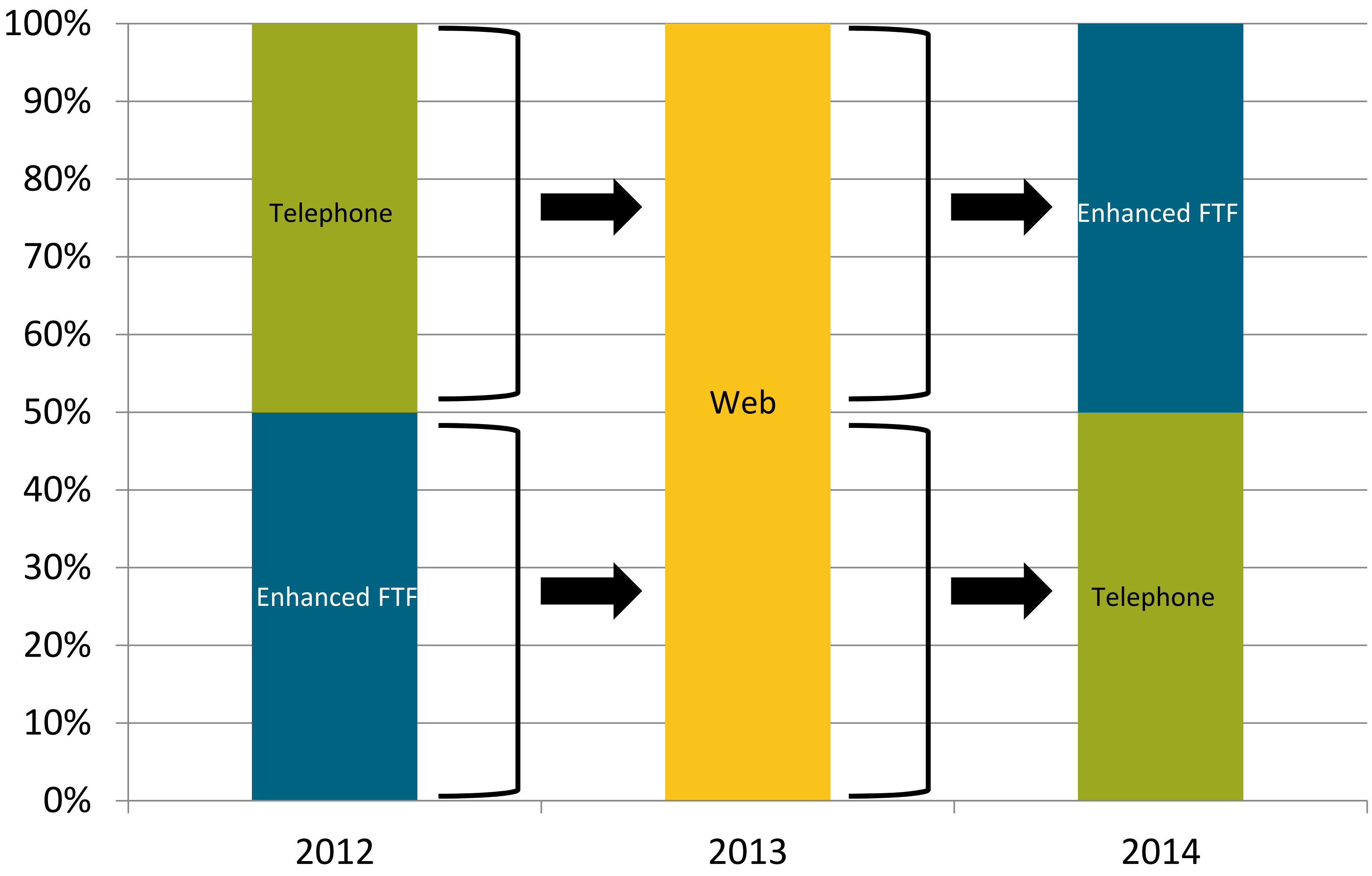


HRS

Analysis Sample

- 2012, 2014: Core interviews
  - Rs <80 who were randomly assigned to TEL or E-FTF in alternate waves
  - Restricted to Rs who self-responded (i.e., not via proxy) in the assigned mode
  - Response rate in both years: 87%
- 2013: Internet survey
  - Administered to a subsample of HRS participants with internet access
  - Response rate: 75%
- 4,223 Rs responded in 2012, 2013 and 2014
  - Control for selection by keeping analytic sample constant

Analysis Sample Rotation



HRS

Cognitive Measures

- Four cognitive tests
  - Serial 7s subtraction
  - Verbal analogies
  - Quantitative number series
  - Numeracy
- Not all tests were administered in all three waves
- Some tests were restricted to random subsamples in one or more waves

Serial 7s Subtraction

- Test of working memory
- Rs asked to subtract seven from 100 five times
  - Given credit for later correct subtractions even if first incorrect
- Key outcome: Count of correct subtractions, 0-5
- Administered in 2012 (IWER), 2013 (WEB) and 2014 (IWER)
- Sample size: 2,113

Verbal Analogies

- Measure of verbal reasoning
- Six-item, block-adaptive test from set of 15 possible items

  "Please finish what I say: Night is to Dark as Day is to ___. "

  - All respondents receive same 3 items in first set
  - Difficulty of second set depends on answers to first set
- Key outcome: Standardized score ranging from 435 to 555
- Administered in 2012 (iwer), 2013 (web), and 2014 (iwer)
- Sample size: 413
  - In 2012, administered to a small, random subsample

Number Series

- Measure of quantitative reasoning/fluency
- Six-item, block-adaptive test from set of 15 possible items

  "For example, if I said the numbers '2 4 6 BLANK,' then what number would go in the blank?"

  - All respondents receive same 3 items in first set
  - Difficulty of second set depends on answers to first set
- Key outcome: Standardized score ranging from 409 to 569
- Administered in 2012 (IWER) and 2013 (WEB) only
- Sample size: 973

Numeracy

- Measure of quantitative ability
- 3 math problems:
    - Chance of getting disease
    - Lottery split
    - Compound interest
- Key outcome: Composite score ranging from 0 to 4 (partial credit for compound interest)
- Administered in 2013 (WEB) and 2014 (IWER) only
- Sample size: 1,069

Analysis Approach

- Primary focus on interviewer versus web administration
  - For interviewer administered, also separate FTF vs. TEL
- Within-subject analysis for IWER vs. WEB
- Between-subject analysis for FTF vs. TEL (within wave)
- Sample restricted to respondents who completed the given test in all waves of administration
  - Sample changes across tests

RESULTS

Do Item Missing Data Rates Differ By Mode?

Percent With Missing Data On Cognitive Tests

	2012		2013	2014	
	TEL	FTF	Web	TEL	FTF
Number series	3.6	2.2	5.3		
Numeracy – item 1			0.6	2.1	1.8
Numeracy – item 2			4.5	7.2	7.4
Serial 7s	1.7	0.8	0.6	1.1	1.4
Verbal analogies	0.0	0.0	2.1	0.0	1.0

Are Descriptive Results Comparable Across Modes?

Means And Standard Deviations For Cognitive Scores

	2012		2013	2014	
Test	TEL	FTF	Web	TEL	FTF
Number series	535.0 (1.1)	532.5 (1.2)	541.4 (0.7)		
Numeracy			2.95 (0.03)	2.56 (0.05)	2.67 (0.05)
Serial 7s	4.12 (0.04)	4.05 (0.04)	4.43 (0.02)	4.20 (0.04)	4.07 (0.04)
Verbal analogies	512.0 (1.7)	515.2 (1.7)	520.5 (1.2)	513.9 (1.9)	519.4 (1.9)

Percent Achieving Maximum Score on Cognitive Tests

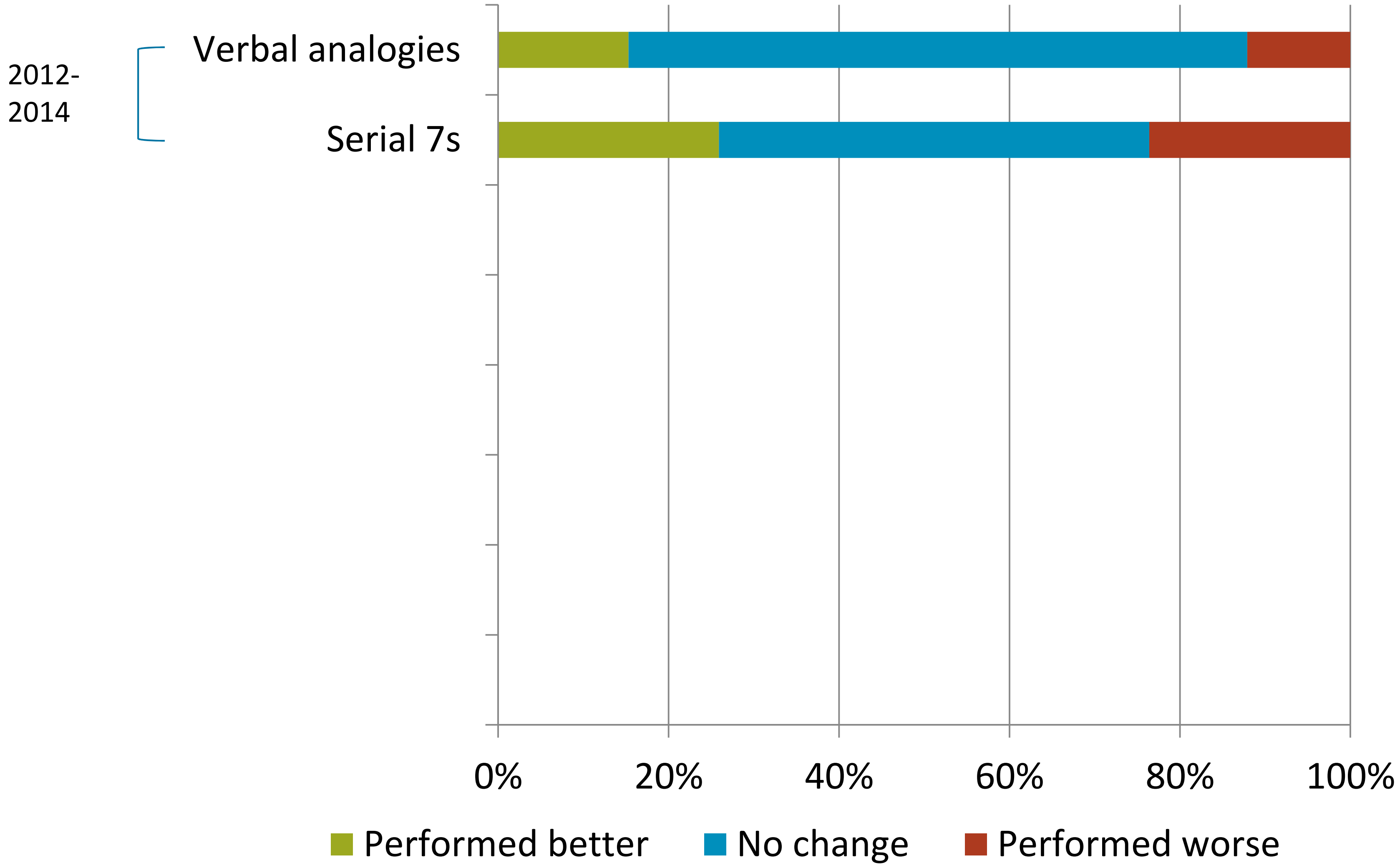
	2012		2013	2014	
Test	TEL	FTF	Web	TEL	FTF
Numeracy	--	--	41.2	24.7	30.0
Serial 7s	57.6	53.8	68.4	58.4	53.9

Within-Test Correlations For Cognitive Tests

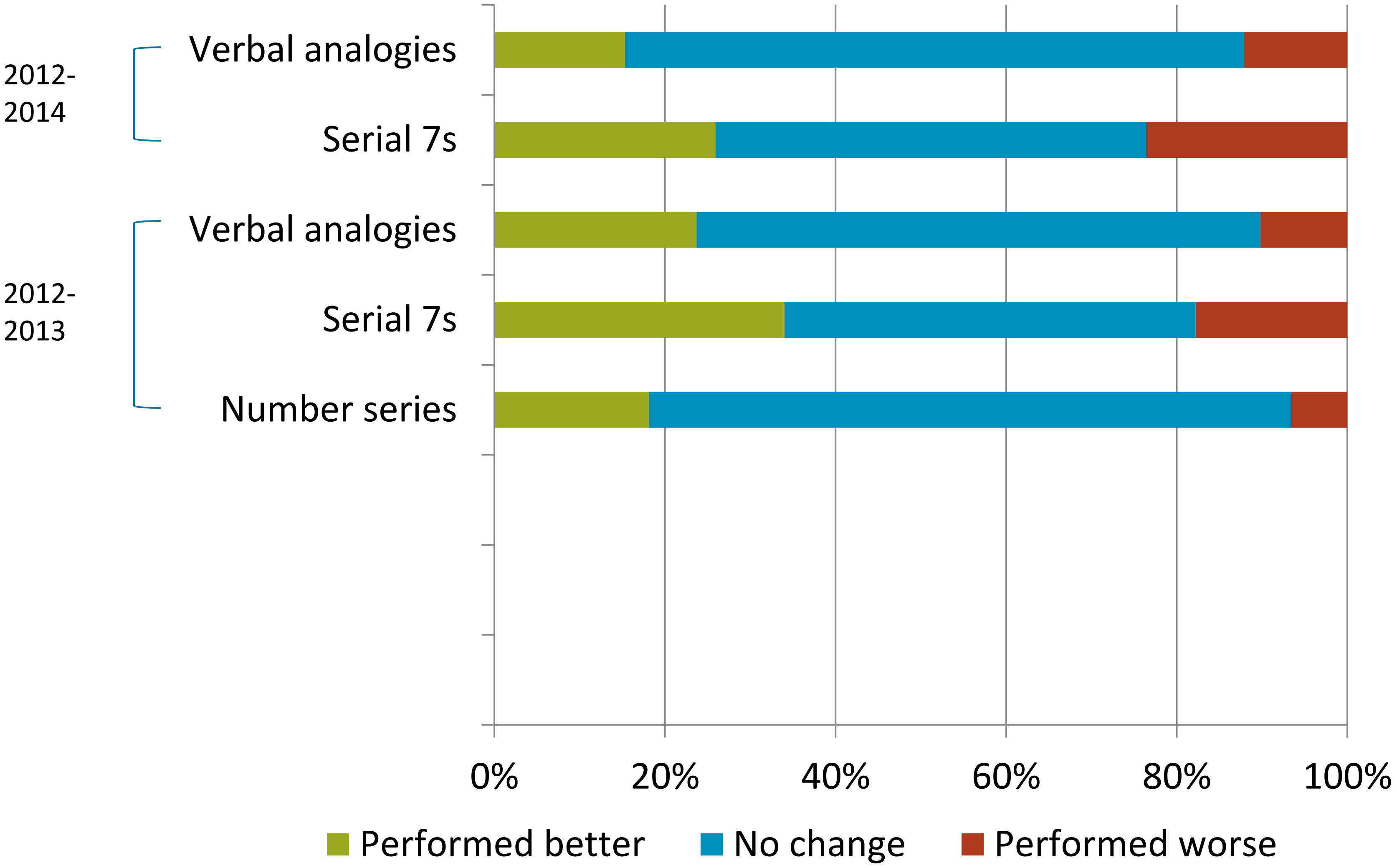
	Pearson Correlations				
Test	2012		2014		2012/2014
	TEL*2013 Web	FTF*2013 Web	TEL*2013 Web	FTF*2013 Web	Iwer*Iwer
Serial 7s	0.22	0.27	0.31	0.22	0.52
Verbal analogies	0.43	0.41	0.44	0.33	0.63

Can We Measure Change in Cognition Over Time?

Percentage Performing Better, the Same or Worse Between Paired Waves

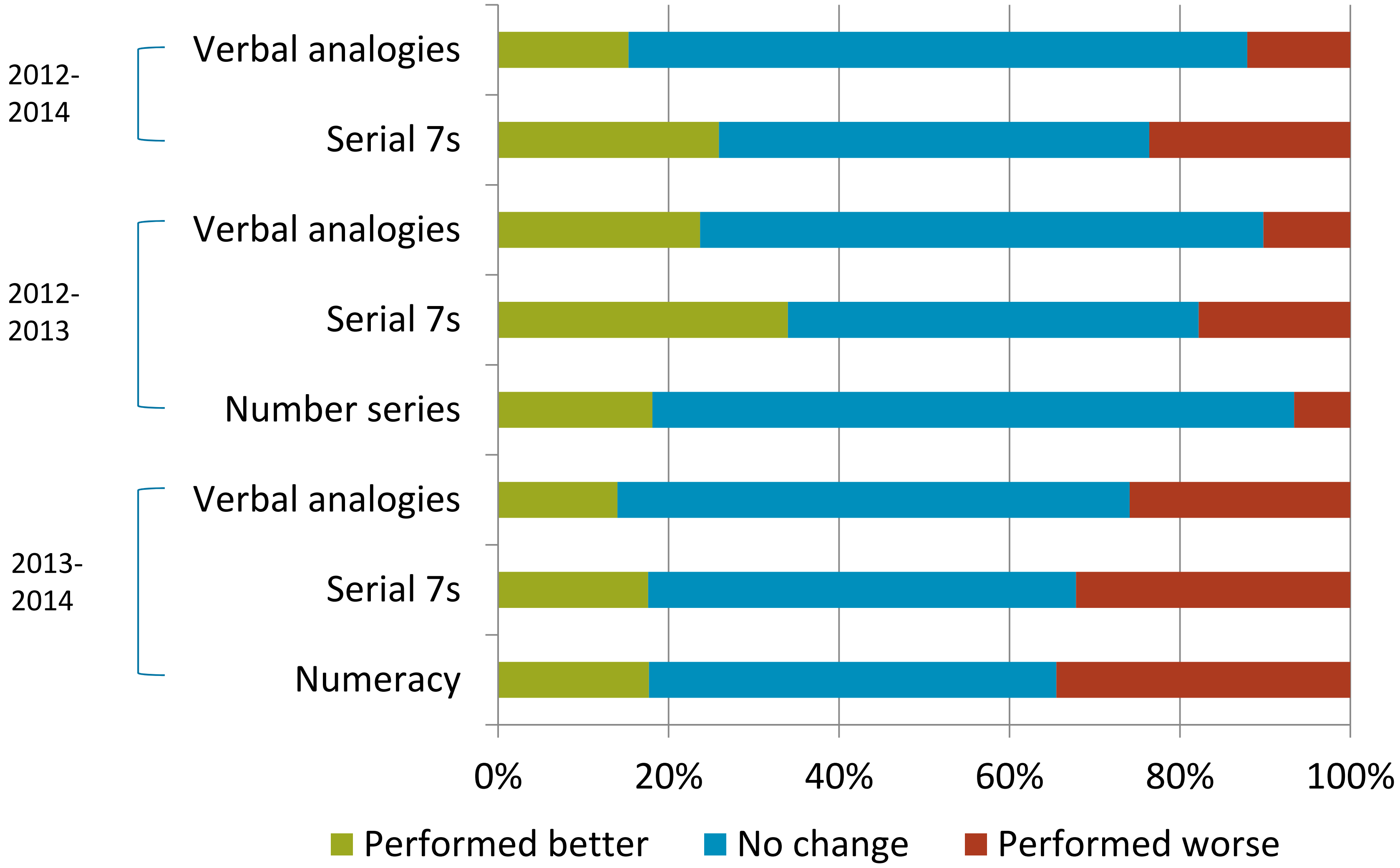


Percentage Performing Better, the Same or Worse Between Paired Waves



HRS

Percentage Performing Better, the Same or Worse Between Paired Waves



HRS

Longitudinal Models

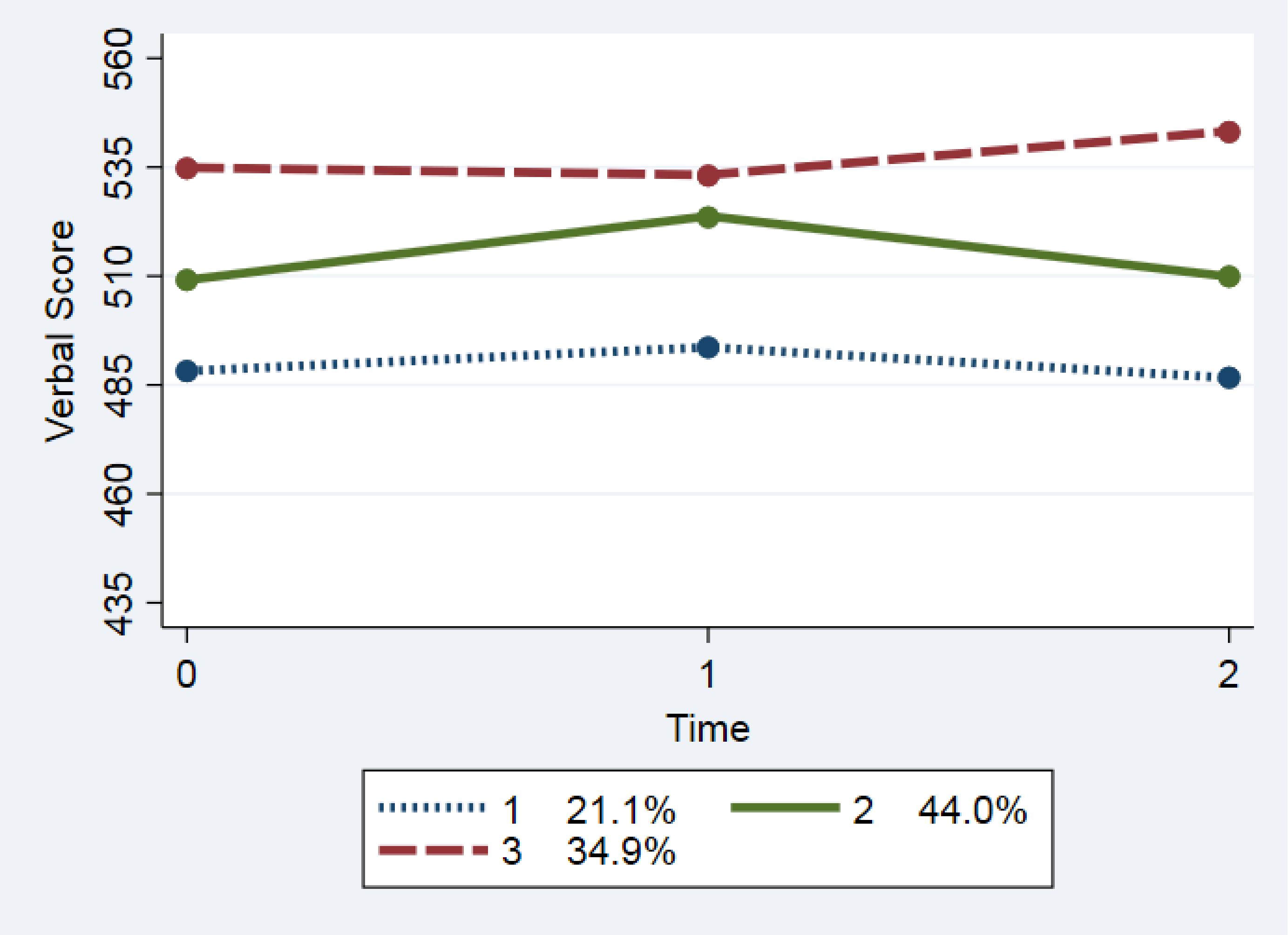
- Approach 1
  - Random intercept repeated measures model
  - Autoregressive error structure
  - Nonlinear fixed effect of time
  - Control for mode (TEL vs. FTF) in 2012
- Approach 2
  - Latent class growth model
  - Quadratic trajectory for each class
  - Tested solutions for one to four classes
- Restricted to cognitive measures with 3 time points

Results From Random Effects Models

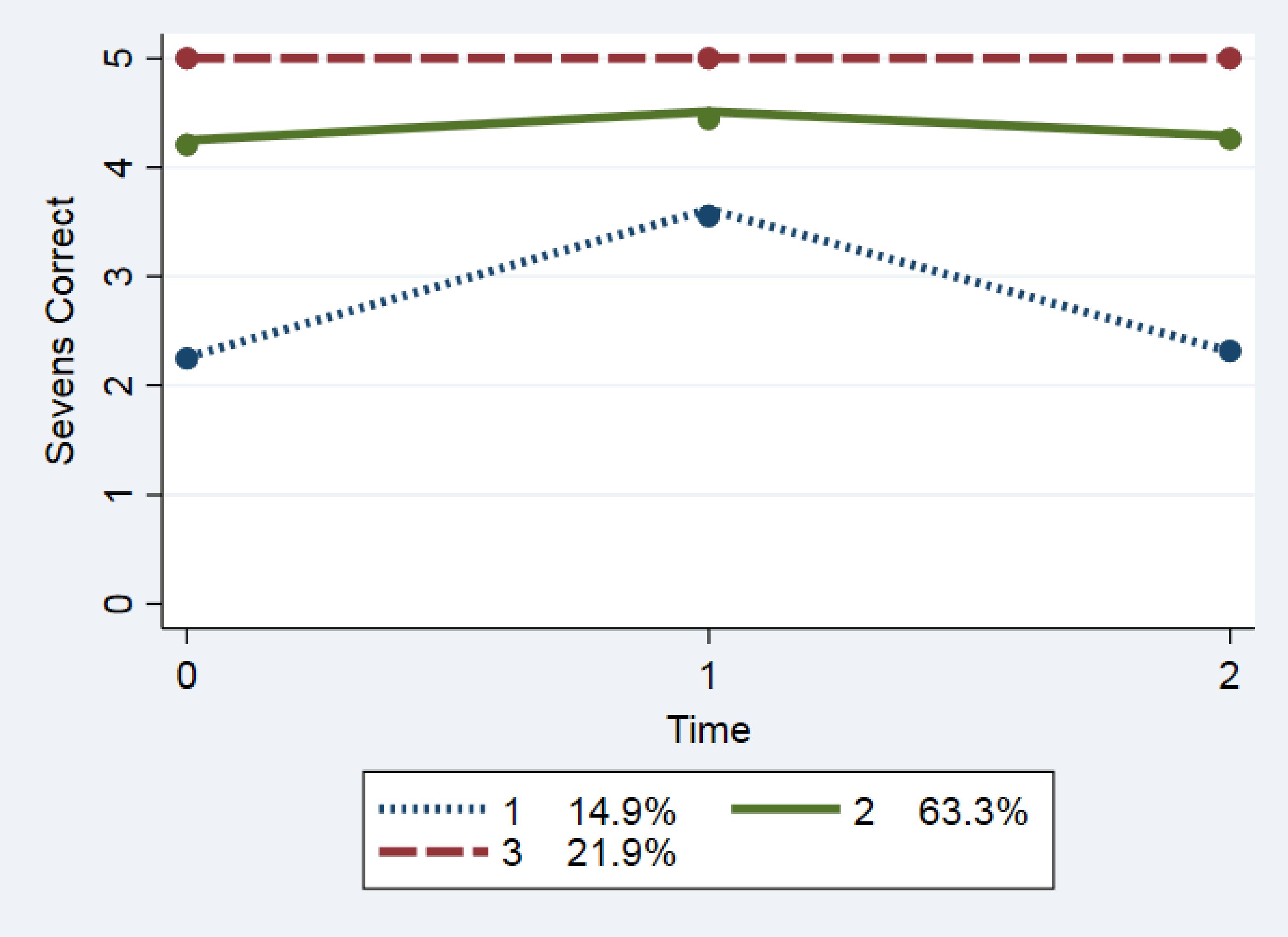
	Verbal Analogies	Serial 7s
Fixed Effects		
Intercept	513.73 (1.624)***	4.081 (0.032)***
Time: 2013 (vs. 2012)	6.889 (1.335)***	0.345 (0.032)***
Time: 2014 (vs. 2012)	2.947 (1.255)*	0.043 (0.028)
2012 FTF (vs. TEL)	-0.182 (2.018)	0.017 (0.038)
Variance Components		
Random intercept	272.64	0.252
Autoregressive errors	0.1302	0.264
Residual variance	374.08	1.137
Model Fit & Sample Size		
BIC	11279.6	19367.4
N	413	2,113

***p<.001, **p<.01, *p<.05

Plot of Three-class Solution from Latent Class Growth Model: Verbal Analogies



Plot of Three-class Solution from Latent Class Growth Model: Serial Sevens



HRS

Can We Make Consistent Multivariate Inferences About Cognitive Ability?

HRS

Results From Multivariate Models Predicting Cognitive Score: Summary

- Models yield inconsistent conclusions by mode, explain less variance on the web
- For example:
  - Education is positively associated with Serial 7s via IWER, not WEB
  - Hispanics score lower than Whites on Serial 7s via IWER, not WEB
  - Rs with higher income score higher on Serial 7s via IWER, not WEB
  - Women had higher verbal scores than men in WEB, not IWER
- For 18 out of 54 regression coefficients, results were substantively different for 2013 WEB vs. 2012 or 2014 IWER

SUMMARY, UPDATE AND DISCUSSION

HRS

Summary of Initial Results

- There are strong selection effects into mode (not discussed here)
    - Web respondents tend to be younger, better educated, more computer literate, and with higher cognitive functioning
- Even controlling for selection (restricting the sample to those who used both interviewer-administered and web modes), we find measurement differences

Summary of Initial Results (Cont.)

- Survey mode influences estimates of cognitive ability
  - Small differences between TEL and FTF
  - Larger differences between WEB and IWER (WEB > IWER)
  - Some tests more problematic than others
- Lower construct validity (correlations) on the web
- Descriptive change estimates suggest a skew toward **improvement** over time when moving from IWER to WEB; **decline** over time when moving from WEB to IWER
- Multivariate relationships using cognition as an outcome are somewhat inconsistent by mode
- Mixed findings regarding item missing data, but overall rates low

Limitations

- Non-experimental design
    - Not a true mixed-mode design
- Mode is confounded with time
    - Only one web data point
    - Insufficient data points and mode transitions to measure stable trajectories
- Small sample sizes for some comparisons
- Analysis sample is not representative of full HRS sample
    - Results are likely attenuated given the selection effects

2018 Web Administration

- In 2018, HRS added web as part of sequential mixed-mode (Web → phone) design in core biennial interview
  - Web offered for regular TEL/FTF respondents only; respondents assigned to enhanced FTF not eligible for web
  - 3,700 eligible for web – criteria included prior report of Internet access, English speaking, self-respondent, non nursing-home resident in prior wave
  - 60% of eligible cases randomly selected for web-first sample; remainder got usual mode of TEL/FTF (controls)
- Web-assigned cases used sequential mixed-mode
  - Web non-respondents followed up by TEL: 81% RR (62% WEB, 19% TEL)
  - Control sample (TEL/FTF): 80% response rate

2018 Mode Comparison – Preliminary 1

- Preliminary analyses being done by Gabor Kezdi, Ben Domingue, Ben Stenhaug & Jessica Faul
- Comparison of web-first versus control group: intent-to-treat comparison
    - Immediate and delayed word recall – similar means/medians by mode assignment, but larger standard deviation for web-assigned group
    - Racial differences by mode assignment vary by measure: disadvantage of African Americans in word recall is stronger in web mode, but weaker for serial 7s, and no racial difference by mode for numeracy items

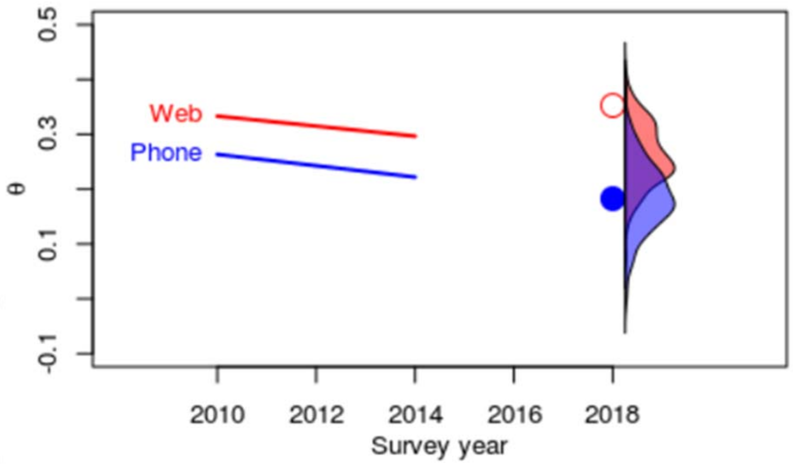
2018 Mode Comparison – Preliminary 2

- Analysis of performance on Telephone Interview for Cognitive Status (TICS) tests using item response theory (IRT) and differential item functioning (DIF)
- First, estimate the difference in cognitive functioning by mode of completion (not assignment), based on prior longitudinal cognition data (both groups TEL in 2010 and 2014)
    - See next slide
- Second, estimate the overall effect of taking the survey via the web as compared to the phone
- Third, explore item-level variation in the magnitude of the mode effect (in progress)

2018 Mode Comparison – Preliminary 3

- Both groups show decline between 2010 and 2014
- Some evidence of selection bias – in both 2010 and 2014, respondents who do the 2018 survey online did somewhat better on average

Figure 3: Difference in group performance.



- The blue dot is the observed cognitive ability for 2018 phone responses – it is consistent with the trend suggested by the blue line
- The hollow red dot is the observed ability for 2018 web respondents – it looks high, suggesting that the web-based test is easier than the phone-based test

Discussion Points

- Is calibration across modes feasible?
    - Our results suggests a simple (constant) adjustment for mode may not work, given variation in cognitive performance across subgroups by mode and differences across cognitive measures
- How do we deal with the issue that different modes are including people at different points on the cognitive performance continuum?
    - Particular challenge for those at the low end of cognitive performance
- How do we interpret or use results from survey-based cognitive tests?
    - Broader question about reliability of these measures for classifying individuals

THANK YOU

HRS