NIH Toolbox Equivalence Study

Jerry Slotkin, University of Delaware

September 27, 2017

Presentation Overview

- Background – how we got here
- Equivalence Study Design
- Equivalence Study Results
- Implications and Actions

Background

- NIH Toolbox publicly released in September 2012 on Web
  - National norms (enhanced in 2015), well validated
- Large, NIH-sponsored research projects desired more portable solution
- iPad solution initially released for NIH use in 2014, public use through App Store in 2015
- Some studies continued to use Web version, some initiated with iPad; others transitioned and have both types of data

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

NIH Toolbox Web vs. iPad

Web Version	iPad Version
Stand-alone laptop computer with separate 19” monitor for test presentation	Uses single screen for both test presentation and participant responses – 9.7”
Participant responses recorded via keyboard (reaction time tests) and/or mouse	Uses touchscreen interface for all participant responses
Internet connection required, data loaded from web	No internet connection required for testing; only for uploading data

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

NIH Toolbox Web vs. iPad Effects?

- After review, we concluded that no mode effects would be expected for Motor, Sensation or Emotion measures
  - Vision test distance shortened to 3 M, optotypes mathematically adjusted for distance, screen size and resolution
  - DVA Test not yet converted to iPad
  - Others do not require interaction with iPad or involve response selection only (e.g., PROs)
- Performance differences on some Cognition tests deemed possible (hypothesized not likely for language measures)

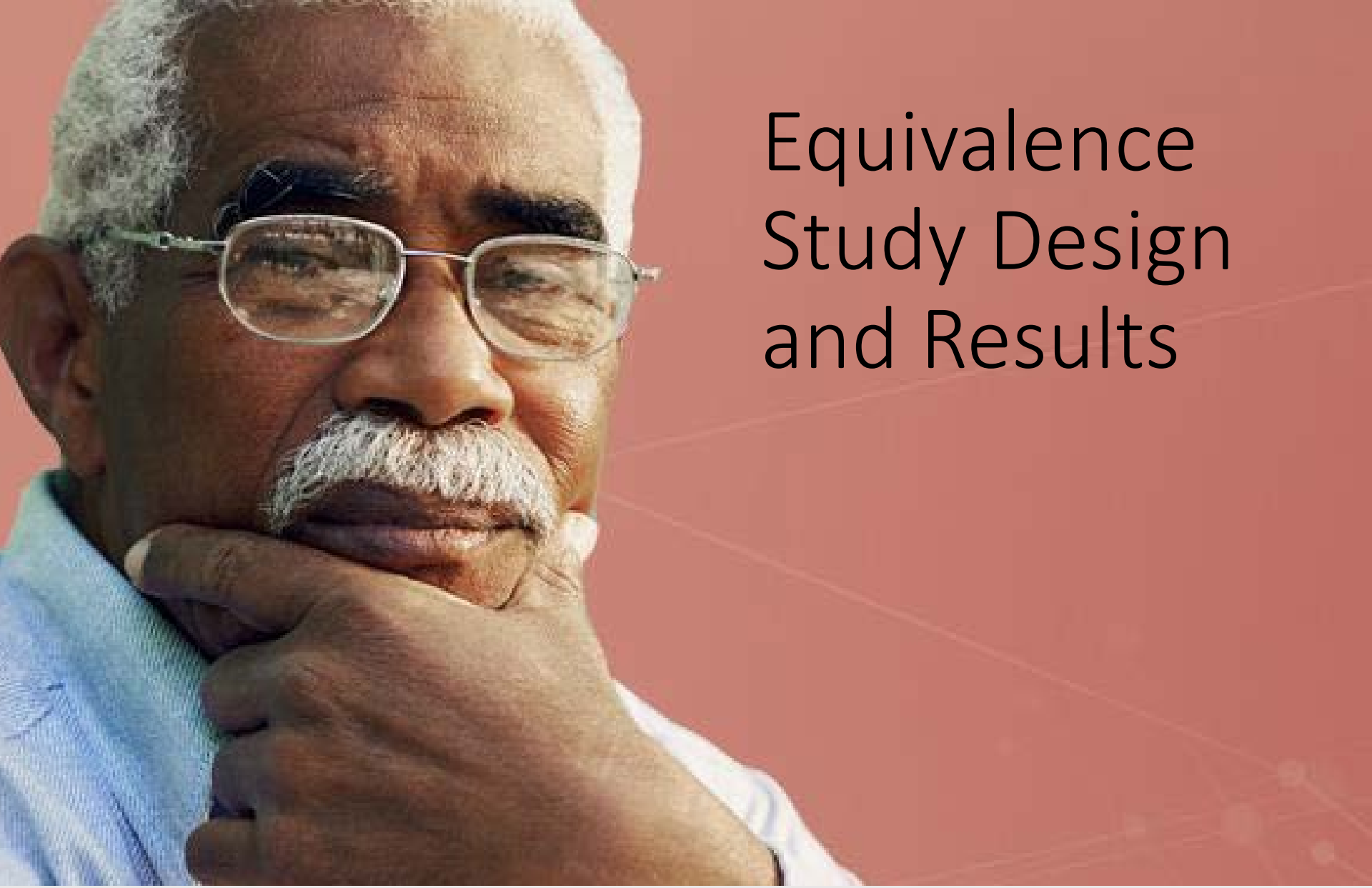
HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Cognition Battery Equivalence Study

- When a test is available in multiple formats, it is essential to ensure all scores are comparable across platforms
- **Study Objective:** To evaluate comparability of the iPad and web-based versions of the Cognition measures to ensure that web-based norms can be applied to scores on the iPad
- Goal is assurance of equivalent interpretation, regardless of format
- Study conducted in 2016-17

NIH TOOLBOX Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED



Equivalence Study Design and Results

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Equivalence Study Design

- Random equivalent-groups design (industry standard)
- In-person assessment; participants randomly assigned either the web- or iPad-based NIH Toolbox Cognition Battery
- Participant random assignment stratified by age, assuring adequate representation of racial/ethnic groups and educational levels
- At least 20 participants per each age bin were assigned to each of the 2 conditions

Equivalence Study Design (cont.)

- Age stratified into 15 levels: 3, 4, 5, 6, 7-9, 10-12, 13-15, 16-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+
- Each participant completed instruments only once
- A research panel company was used to identify and test participants in two major metropolitan areas
- 648 participants (331 children aged 3-17; 317 adults aged 18+) were enrolled in the study
- 321 were randomized to web- and 327 to app-based administration

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Analysis Plan

- Test for significant differences by mode of administration for each test
  - Effect size differences greater than 0.20 were considered non-equivalent (industry standard)
  - Analyzed both most basic scores (e.g., raw score, theta, or computed scores) and normative scores (uncorrected, age-corrected, and fully corrected scores)
- For non-equivalent tests:
  - Evaluate all four moments (mean, sd/variance, skew, kurtosis) of the tests by group
  - Determine optimal equating approach (Mean, Linear, or some other format)

Results

	AC/Web Mean (SD)	iPad Mean (SD):	Effect Size [95% CI]	p-value
Vocabulary	1.90 (4.09) N=313	1.91 (4.18) N=327	< 0.01 [-0.15, 0.16]	0.97
Reading	3.69 (3.54) N=238	3.81 (4.19) N=252	0.03 [-0.15, 0.21]	0.72
List Sorting	17.0 (4.8) N=225	16.7 (4.4) N=249	-0.08 [-0.26, 0.10]	0.39
Picture Sequence Memory	-0.55 (1.09) N=281	-0.87 (1.00) N=309	-0.30 [-0.47, -0.14]	<0.001
DCCS	6.7 (2.6) N=297	7.4 (2.8) N=316	0.26 [0.10, 0.41]	<0.01
Flanker	7.0 (2.2) N=302	7.7 (2.2) N=318	0.31 [0.15, 0.47]	<0.001
Pattern Comparison	52.9 (15.5) N=237	42.0 (9.6) N=248	-0.84 [-1.03, -0.65]	<0.001

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Results Summary

- Mode of administration did not influence test scores on three NIHTB tests:
  - Picture Vocabulary
  - Reading
  - List Sorting
- Four tests showed differences between Web and iPad, requiring adjustments:
  - Picture Sequence Memory
  - DCCS
  - Flanker
  - Pattern Comparison

PROMIS Neuro QoL ASCQ-Me Adult Sickle Cell Quality of Life Measurement Information System NIH TOOLBOX Motor Emotion Sensation Cognition HealthMeasures TRANSFORMING HOW HEALTH IS MEASURED

PROMIS

NeuroQoL

ASCQ-Me℠
Adult Sickle Cell Quality of Life Measurement Information System

NIH TOOLBOX Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Picture Sequence Memory Test

- Score variance was approximately equal
- Mean differences varied by age
- Mean Equating was conducted within age bands

DCCS

- DCCS scores are composed of Accuracy and Reaction Time sub-scores
- Accuracy scores were equivalent
- Reaction time was faster on the iPad (touchscreen) than on the web (keyboard)
- Adjustments were made to the reaction time scores only, using Mean Equating

Flanker

- Flanker scores are composed of Accuracy and Reaction Time sub-scores
- Accuracy scores were equivalent
- Reaction time was faster on the iPad (touchscreen) than on the web (keyboard)
- Adjustments were made to the reaction time scores only, using Mean Equating

Pattern Comparison

- Scores on the iPad were lower and had less variability than scores on the web
- Investigation of iPad programming revealed that Apple (iOS) was automatically adding 200ms to each item – individuals were getting significantly fewer items on the iPad than on the web
- Linear Equating allowed adjustment of both scores and score variance

Study Implications

- Four cognition tests showed evidence of score differences between the Web and iPad versions, *primarily* due to differences in speed of responding on iPad touchscreen vs. keyboard
- Mean and linear equating adjustments successfully addressed these score differences across the lifespan assessed by NIH Toolbox (ages 3-85)
- The norms for the NIH Toolbox, originally developed for the Web, can now be confidently applied to the iPad-administered tests

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Completed and Ongoing Efforts

- iPad app rescoring process implemented
  - Existing, pre-equivalence study data rescored as part of app update, early 2017
  - All new administrations of the app will automatically apply the conversions, so that rescoring will not be necessary in the future
- Analysis Guide published – a user-oriented brief document describing what users can or should do based on mode(s) of administration used
- Equivalence Study manuscript – in process

Questions?

NIH TOOLBOX® Motor Emotion Sensation Cognition

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED