Reliability and Error in Measurement Instruments developed with Classical Test Theory and Item Response Theory

Allan J. Kozlowski, Ph.D., B.Sc. (PT)
Assistant Professor of Rehabilitation Medicine
Icahn School of Medicine at Mount Sinai

Mount Sinai

Disclosures

None.

Objectives

1. Review two frameworks for validity in rehabilitation measurement.
2. Review, compare, and contrast reliability from the Classical Test Theory (CTT) and Item-Response Theory (IRT) perspectives.
3. Discuss the implications for interpreting scores and conducting analyses using scores from CTT- and IRT-based measurement instruments.

Outline

1. Review two rehabilitation measurement validity frameworks
    1. Conventional (COSMIN)
    2. Contemporary (Messick)
2. Describe reliability
    1. From the Classical Test Theory (CTT) perspective
    2. From the Item-Response Theory perspective
    3. In regard to common sources of measurement error
3. Describe the implications for
    1. Interpreting scores at a single time point
    2. Interpreting change over time
    3. Modeling longitudinal data
4. Questions, Answers, and Discussion

Every measurement has error.

COSMIN: A Conventional Validity Framework

Conventional Validity Framework: COSMIN

- **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments

- http://www.cosmin.nl/

- Resources
  - Taxonomy
  - Checklist
  - Systematic reviews of Measurement Properties

- de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide.* Cambridge, UK: Cambridge University Press; 2011.

Conventional Validity Framework: COSMIN Definitions

Validity:

The degree to which an instrument measures the construct(s) it purports to measure.

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html

Conventional Validity Framework: COSMIN Definitions

- **Face:** …instrument *appears* to adequately reflect the construct
- **Content:** … instrument *content* adequately reflects the construct
- **Construct:** … scores of an instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct
  - Structural
  - Hypothesis Testing
  - Cross-Cultural
- **Criterion:** … degree to which the scores of an instrument adequately reflect a 'gold standard'
- http://www.cosmin.nl/cosmin-taxonomy_3_0.html

Conventional Validity Framework: COSMIN Definitions

Validity:

The degree to which *an instrument* measures the construct(s) it purports to measure.

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html

Messick: A Contemporary Validity Framework

Contemporary Validity Framework: Messick

Construct Validity: … is not simply a property of a measure but is a reflection of and resides in the conditions of its use.

- Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores

- Messick S. *Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. Am Psychol 1995;50(9):741-9.

Contemporary Validity Framework: Messick

Six Aspects of Construct Validity:

- **Content:** relevance, representativeness, technical quality
- **Substantive:** empirical evidence fits theoretical basis
- **Structural:** scores ≡ instrument ≡ construct
- **Generalizability:** samples → population groups, settings, tasks
- **External:** convergent and discriminant evidence
- **Consequential:** implications of interpretation and action

Conventional 'types' are methods of establishing evidence.

Reliability from the Classical Test Theory Perspective

Classical Test Theory: Reliability definition

- **Reliability:** The degree to which the measurement is free from measurement error

$$X = T + E$$

  - X = Observed score
  - T = True score
  - E = Error

  - Sample: assumes errors are normally distributed about a mean of zero

- http://www.cosmin.nl/cosmin-taxonomy_3_0.html

Classical Test Theory: Reliability types

- **Internal consistency:** The degree of the interrelatedness among the items
- **Rater reliability:**
  - Inter-rater: consistency in scores across two or more raters
  - Intra-rater: consistency for a single rater
- **Test-retest reliability:** consistency over time (stable or unchanging sample)
- **Group average of summary scores**
- http://www.cosmin.nl/cosmin-taxonomy_3_0.html

Reliability from the Item-Response Theory Perspective

Item Response Theory: Reliability

- **Item response theory (IRT):** attempts to explain the response of a person to an item based on the idea that the probability of a correct/keyed response is a function of
  - Person ability
  - Item difficulty

- IRT extends the CTT concept of reliability:
  - Measurement precision varies across ranges of item difficulty and person ability (i.e., observed scores).
  - Scores nearer the floor and ceiling tend to have larger errors than scores nearer the mid-range.

- Rasch model can be considered a 1-parameter IRT model

- http://en.wikipedia.org/wiki/Item_response_theory

Item Response Theory: Reliability

- **Item response theory (IRT):** attempts to explain the response of a person to an item based on the idea that the probability of a correct/keyed response is a function of
  - Person ability
  - Item difficulty

- IRT extends the CTT concept of reliability:
  - Measurement precision varies across ranges of item difficulty and person ability (i.e., observed scores).
  - Scores nearer the floor and ceiling tend to have larger errors than scores nearer the mid-range.

- Rasch model can be considered a 1-parameter IRT model

- http://en.wikipedia.org/wiki/Item_response_theory

Item Response Theory: Reliability

- **Misfit:**
  - Overfit: not enough variation in responses; may indicate redundant items
  - Underfit: unexpected patterns of responses; may indicate 'noisy' items
  - Infit: information-weighted fit statistic
  - Outfit: outlier-sensitive fit statistic
- **Differential item functioning (DIF):**
  - The extent to which item scores are influenced by confounding.
  - E.g., Male/female differences.
- **Category Disorder:**
  - Higher category is more likely at a lower point than a lower category
- Test effect of collapsing categories and removing items
- http://en.wikipedia.org/wiki/Item_response_theory

Rasch Measures and Standard Error

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
1	.00E	15.31	24	31.06	2.26	47	43.32	2.48
2	9.05	7.68	25	31.64	2.23	48	44.05	2.57
3	13.39	5.01	26	32.20	2.19	49	44.84	2.68
4	15.65	3.95	27	32.75	2.16	50	45.70	2.79
5	17.17	3.38	28	33.28	2.13	51	46.64	2.92
6	18.35	3.04	29	33.80	2.11	52	47.66	3.04
7	19.33	2.82	30	34.30	2.09	53	48.76	3.15
8	20.19	2.67	31	34.80	2.07	54	49.94	3.24
9	20.97	2.57	32	35.29	2.06	55	51.17	3.29
10	21.71	2.50	33	35.77	2.05	56	52.42	3.31
11	22.42	2.46	34	36.26	2.05	57	53.68	3.32
12	23.11	2.44	35	36.74	2.05	58	54.96	3.37
13	23.79	2.44	36	37.22	2.05	59	56.30	3.48
14	24.47	2.44	37	37.70	2.06	60	57.78	3.70
15	25.16	2.44	38	38.19	2.07	61	59.50	4.07
16	25.84	2.45	39	38.69	2.09	62	61.67	4.65
17	26.53	2.44	40	39.20	2.11	63	64.60	5.48
18	27.21	2.44	41	39.71	2.14	64	68.77	6.63
19	27.89	2.42	42	40.25	2.18	65	75.11	8.31
20	28.56	2.40	43	40.80	2.22	66	85.65	11.06
21	29.21	2.37	44	41.38	2.27	67	100.00E	17.31
22	29.85	2.33	45	41.99	2.33			
23	30.46	2.30	46	42.63	2.40			

Item Response Theory: Other Benefits

- **Computer-Adaptive Test (CAT):** Subsequent questions depend on response to previous questions
  - Only include relevant items
  - Reduce burden to patient/subject: fewer items with similar reliability
  - Reduce burden to clinician or researcher: computer administration

- **Short Forms**:
  - Fixed
  - Tailored

Common Sources of Measurement Error

Common Sources of Error: CTT

- **Patient or Subject:**
  - Normal variability of construct
  - Comprehension
- **Instrument:**
  - Irrelevant items included
  - Important items excluded
- **Rater:**
  - Variability in skill
  - Variability in observation
  - Potential for bias?
- **Environment:**
  - Variations in equipment, space, etc.
  - Distractions to patient/subject
  - Distractions to rater

- Potential for bias?

Common Sources of Error: IRT

- **Patient or Subject:**
  - Normal variability of construct
  - Comprehension
- **Instrument:**
  - Irrelevant items included
  - Important items excluded
- **Rater:**
  - Variability in skill
  - Variability in observation
  - Potential for bias?
- **Environment:**
  - Variations in equipment, space, etc.
  - Distractions to patient/subject
  - Distractions to rater

- Potential for bias?

- **Dimensionality**
- **Item Misfit**
- **DIF**

Implications for Interpreting Scores at a Single Time Point

Interpretation of Single Point Scores: CTT

- **Point Estimate:**
  - Observed score
  - Often ordinal
- **Margin of error:**
  - Standard Error of Measurement (SEM)
  - ±1 SEM ≡ 67% Confidence interval (CI)
  - ±1.96 SEM ≡ 95% CI
- **Conditional SEM**
  - Margin of error varies across scale range

Interpretation of Single Point Scores: IRT

- **Point Estimate:**
  - Rasch measure (logit or transformed)
  - Interval level
  - Summary or item score
- **Margin of error:**
  - Standard Error (SE)
  - ±1 SE ≡ 67% Confidence interval (CI)
  - ±1.96 SE ≡ 95% CI
  - Varies across scale range

Implications for Interpreting Change Over Time

Interpretation of Change over Time: CTT

- **Minimal Detectable Change (MDC):**
  - Margin of error for change score
  - Derived from stable (unchanged) sample
  - Two-point change: which time points?
  - Conditional MDC: varies for range of baseline score

- **Minimal Clinical Important Difference (MCID):**
  - Index for important change
  - Derived from sample who have changed an important amount
  - Importance anchored to
    - Patient
    - Clinician
    - Both
    - Other

Interpretation of Change over Time: IRT

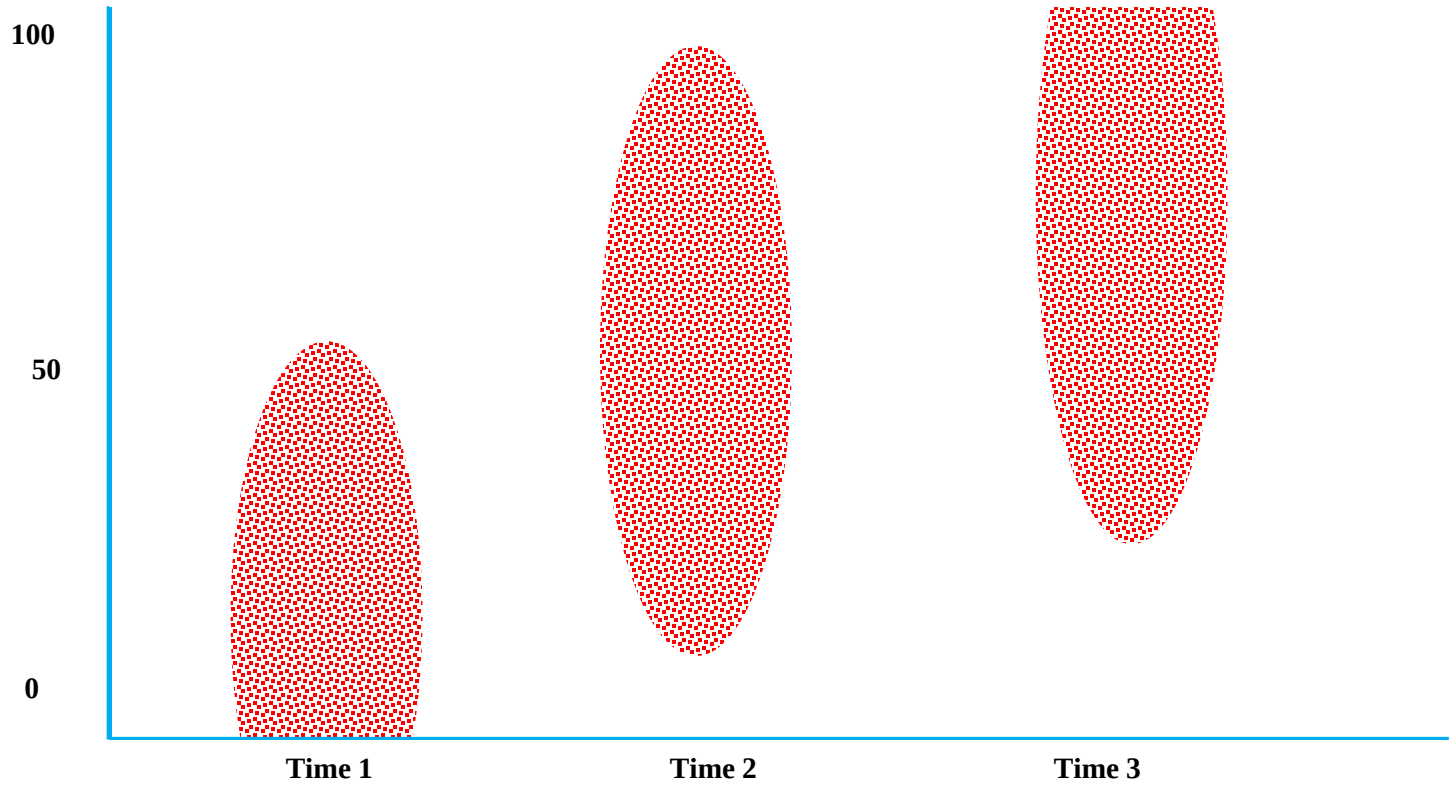
- **No MDC or MCID**
- **Change on summary score**
- **Keyform Maps**
  - Pattern of change (Bode 2014)
  - Item category thresholds (Velozo 2011)

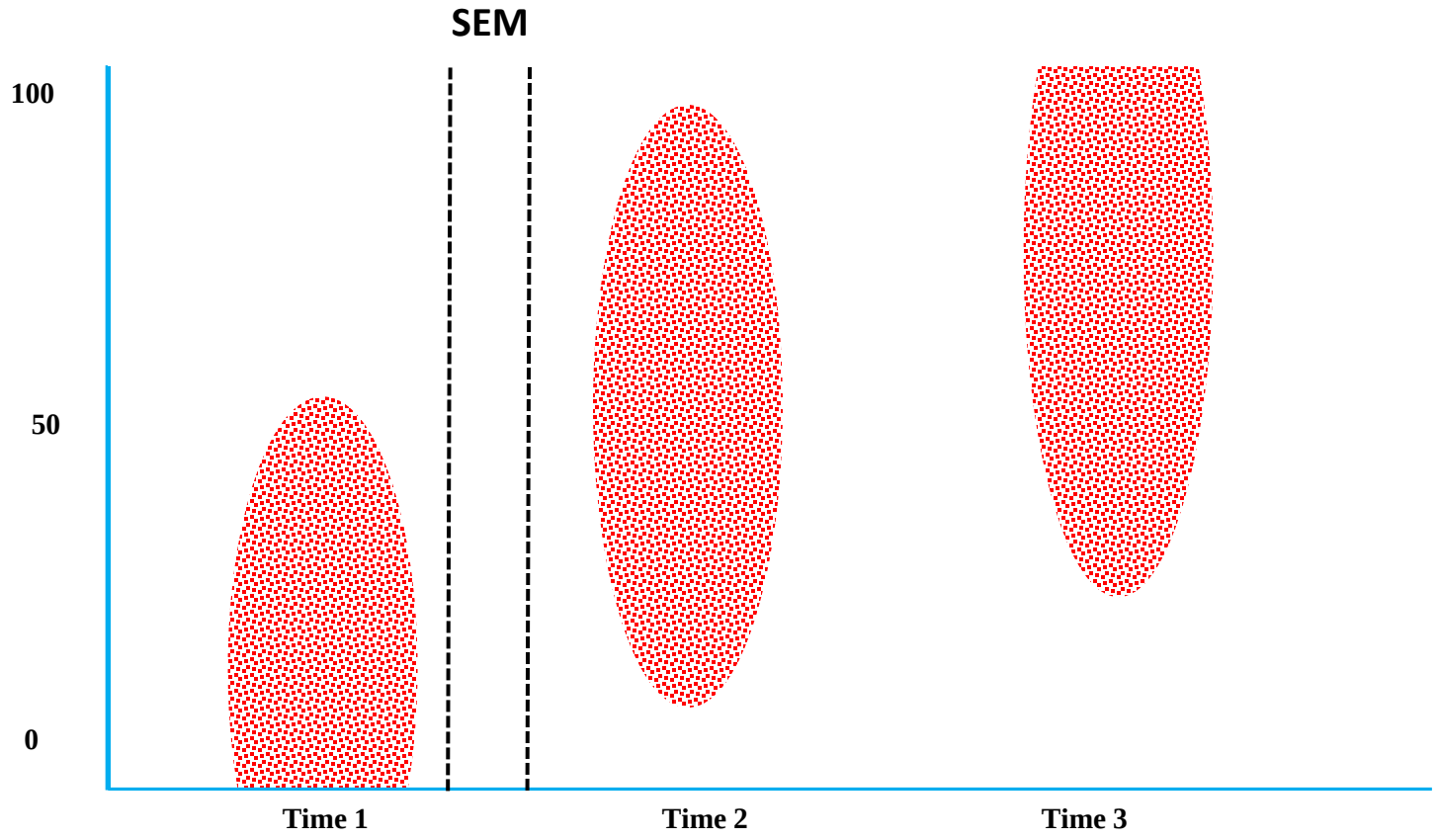
- Bode RK, Heinemann AW, Kozlowski AJ, Pretz CR. Self-Scoring Templates for Motor and Cognitive Subscales of the FIM Instrument for Persons With Spinal Cord Injury. Archives of physical medicine and rehabilitation 2014;95(4):676-9 e5.
- Velozo CA, Woodbury ML. Translating measurement findings into rehabilitation practice: an example using Fugl-Meyer Assessment-Upper Extremity with patients following stroke. J Rehabil Res Dev 2011;48(10):1211-22.

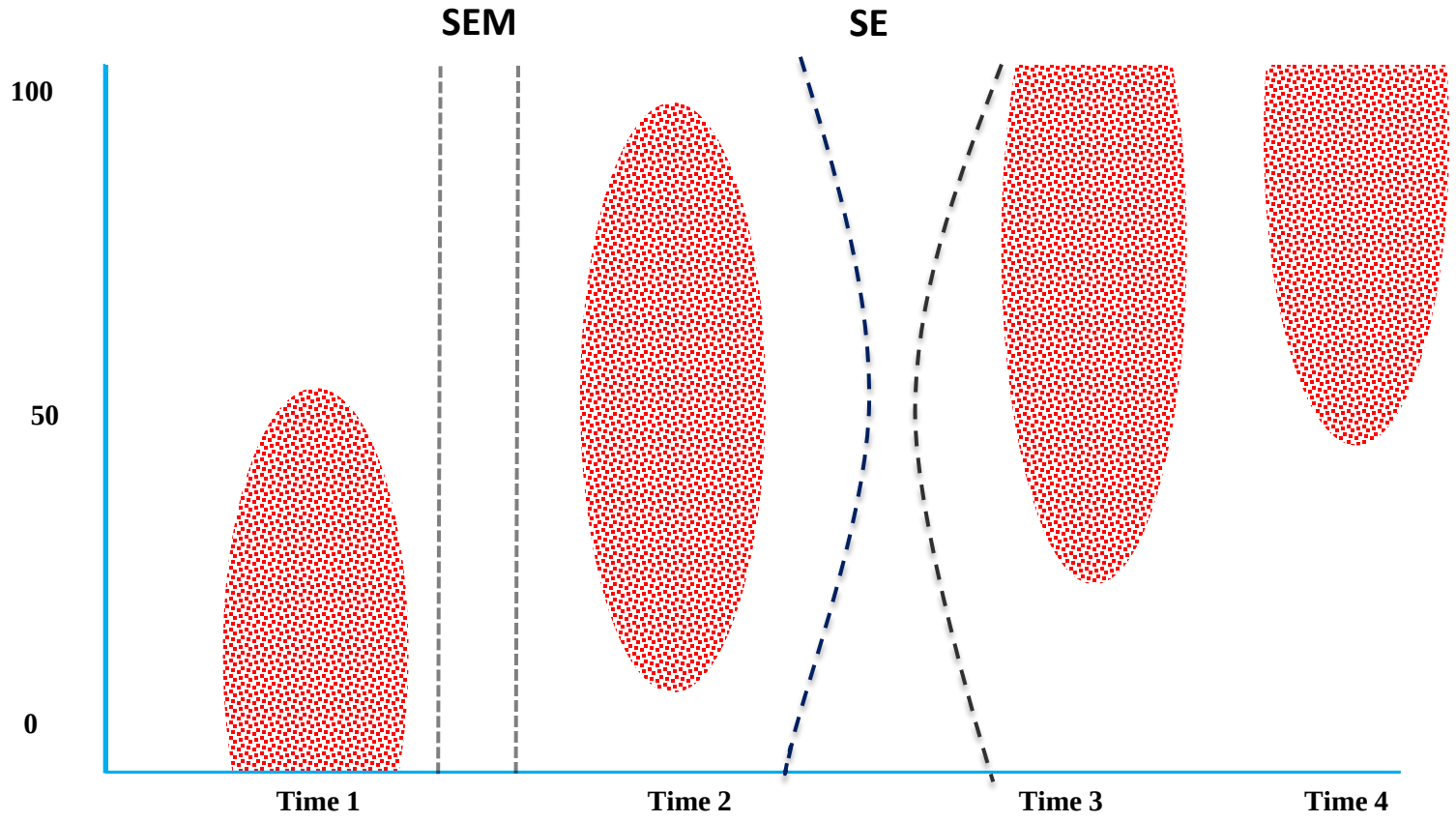
Implications for Modeling Longitudinal Data

Implications for Modeling Longitudinal Data

- **Do errors vary over time?**
- **Does timing of data collection matter?**
- **Ordinal or Interval?**

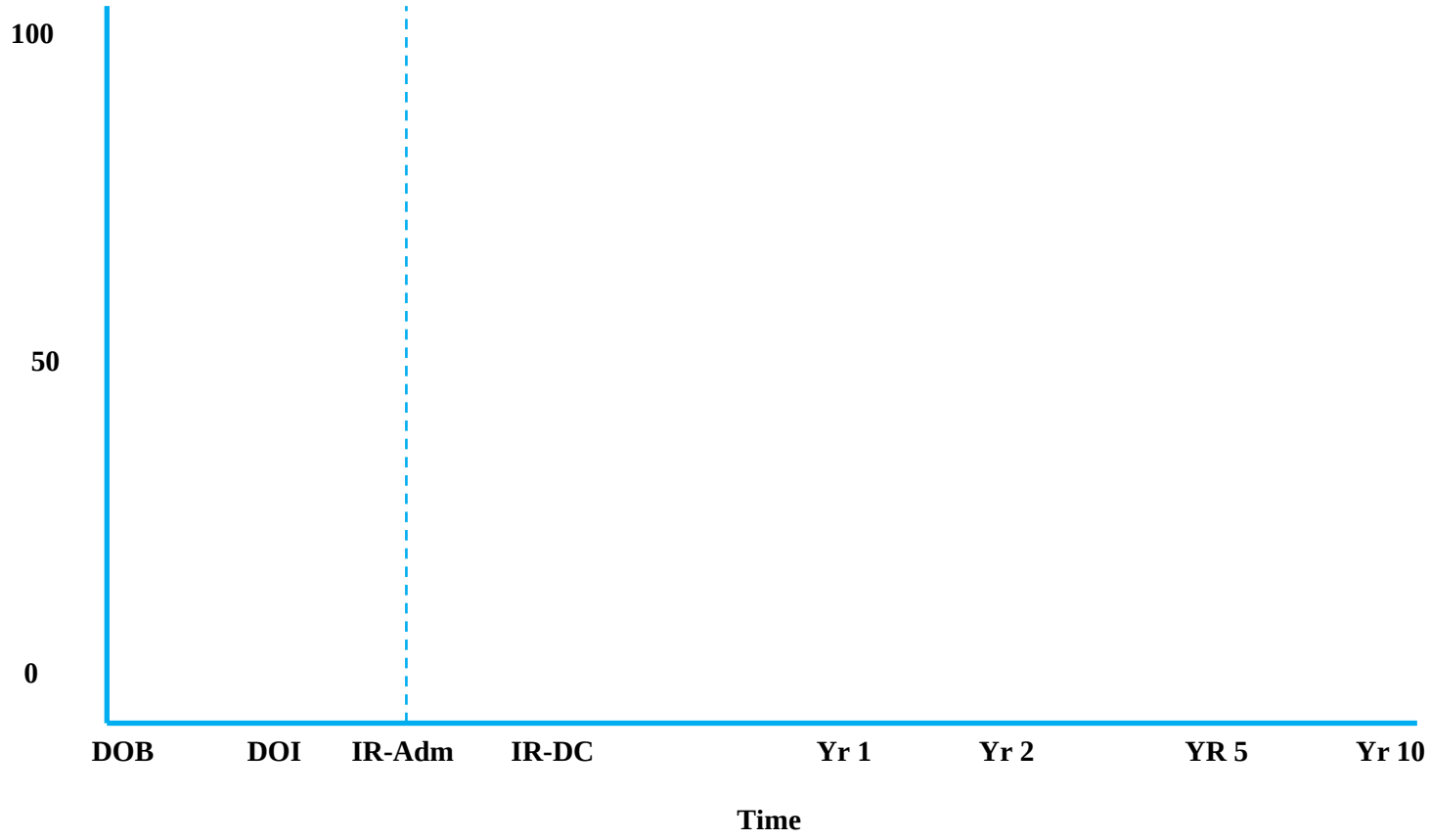

100
50
0
Time 1
Time 2
Time 3

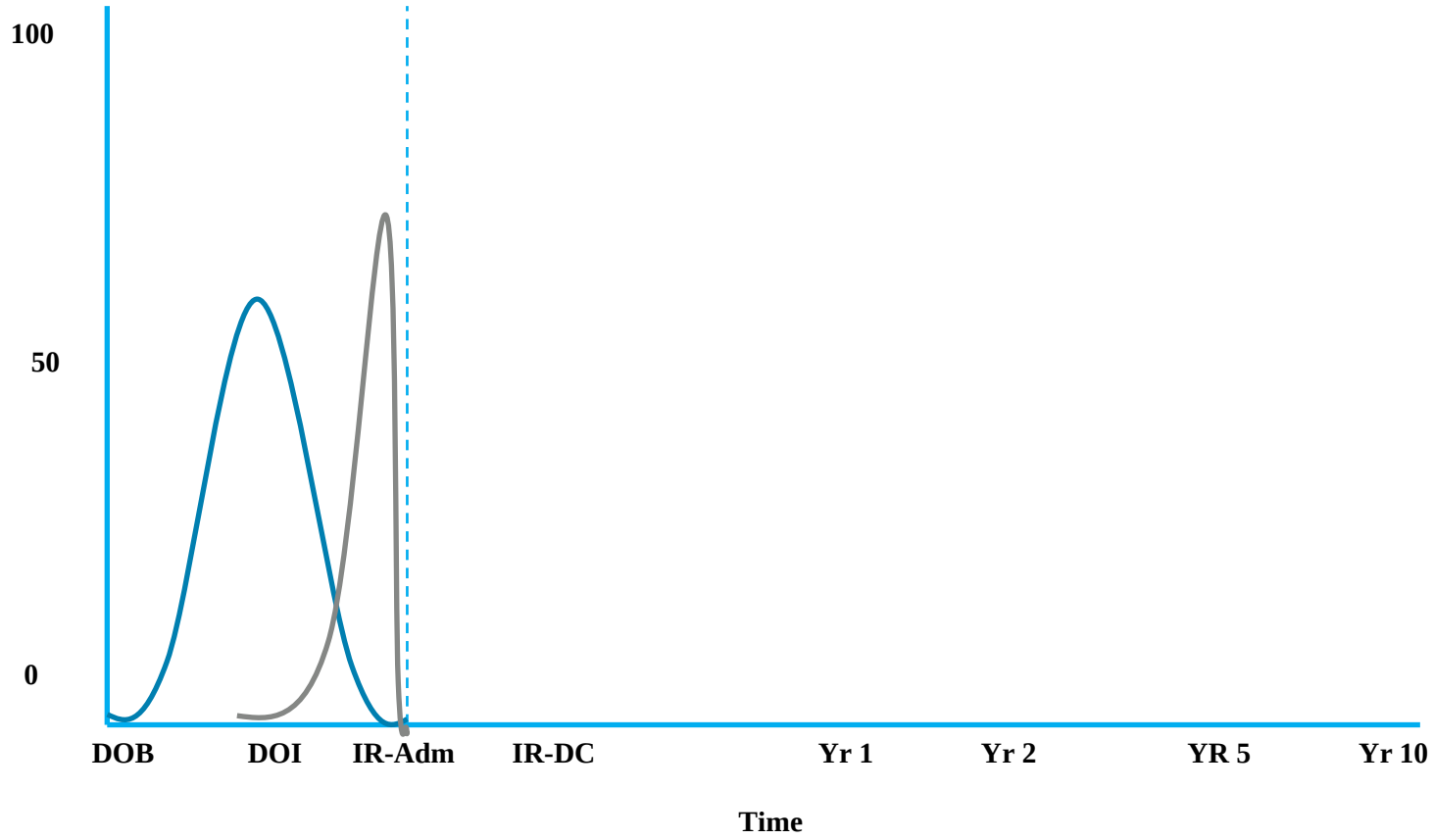

SEM
100
50
0
Time 1
Time 2
Time 3

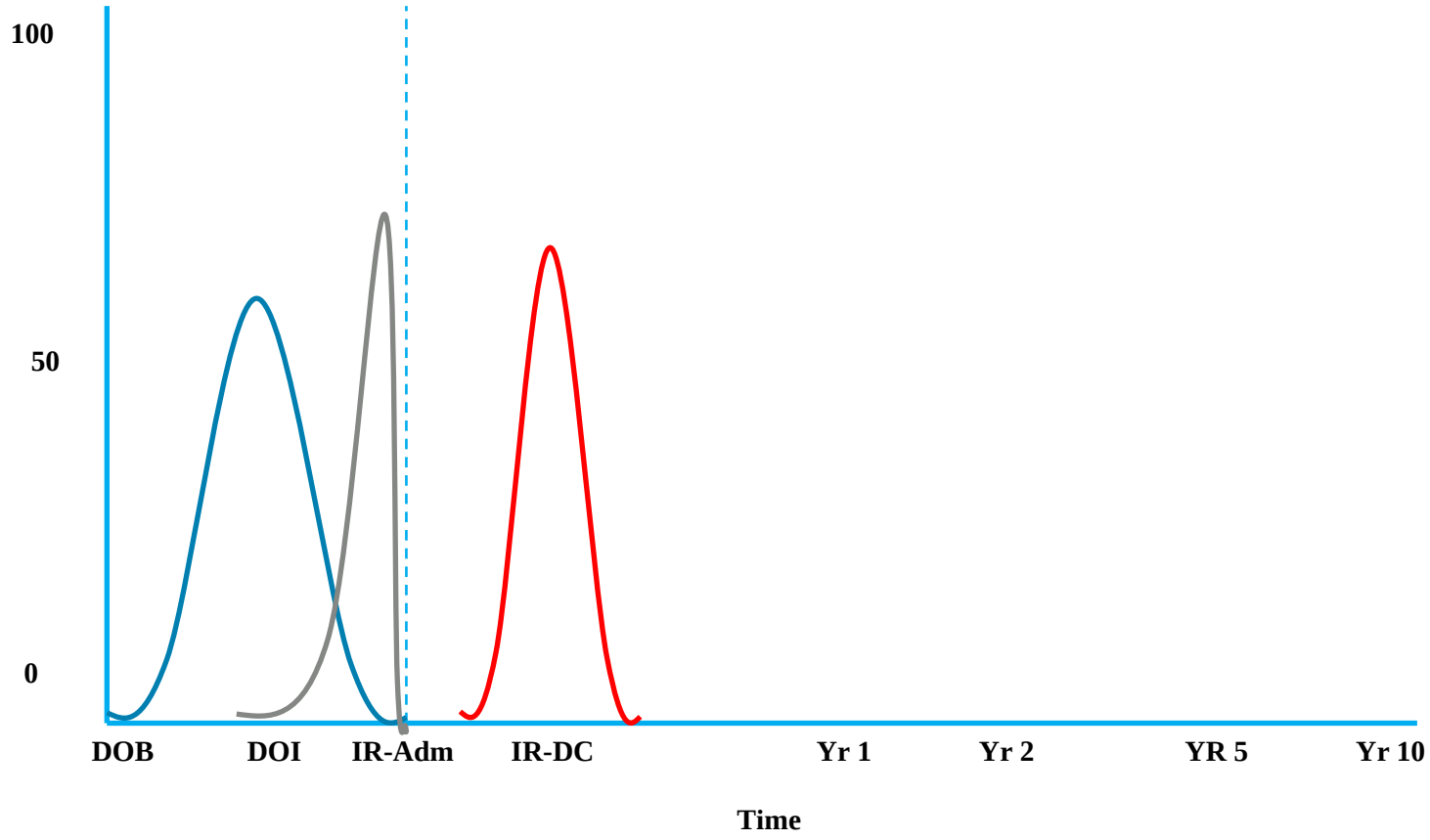

SEM
SE
100
50
0
Time 1
Time 2
Time 3
Time 4

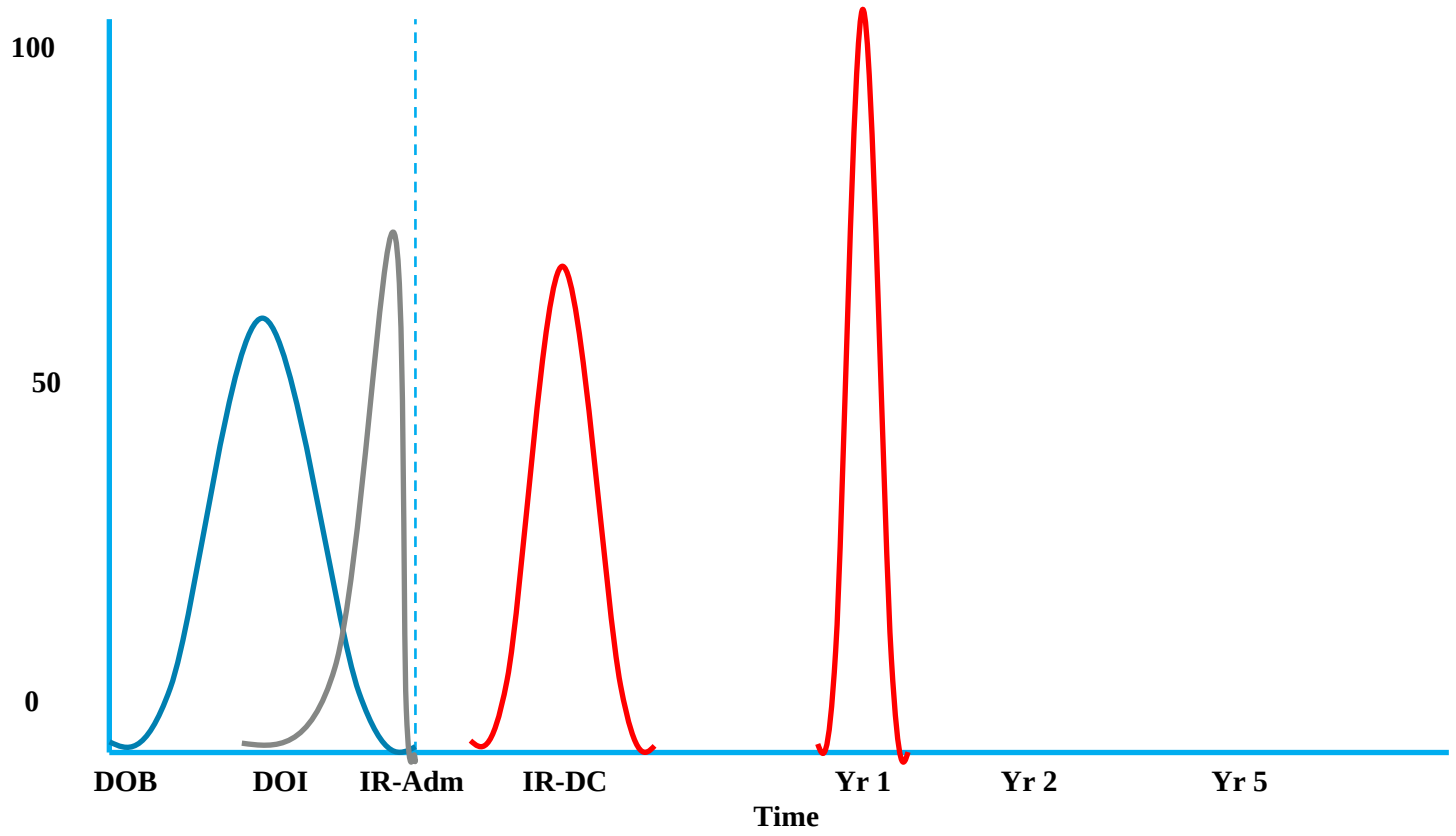
Implications for Modeling Longitudinal Data

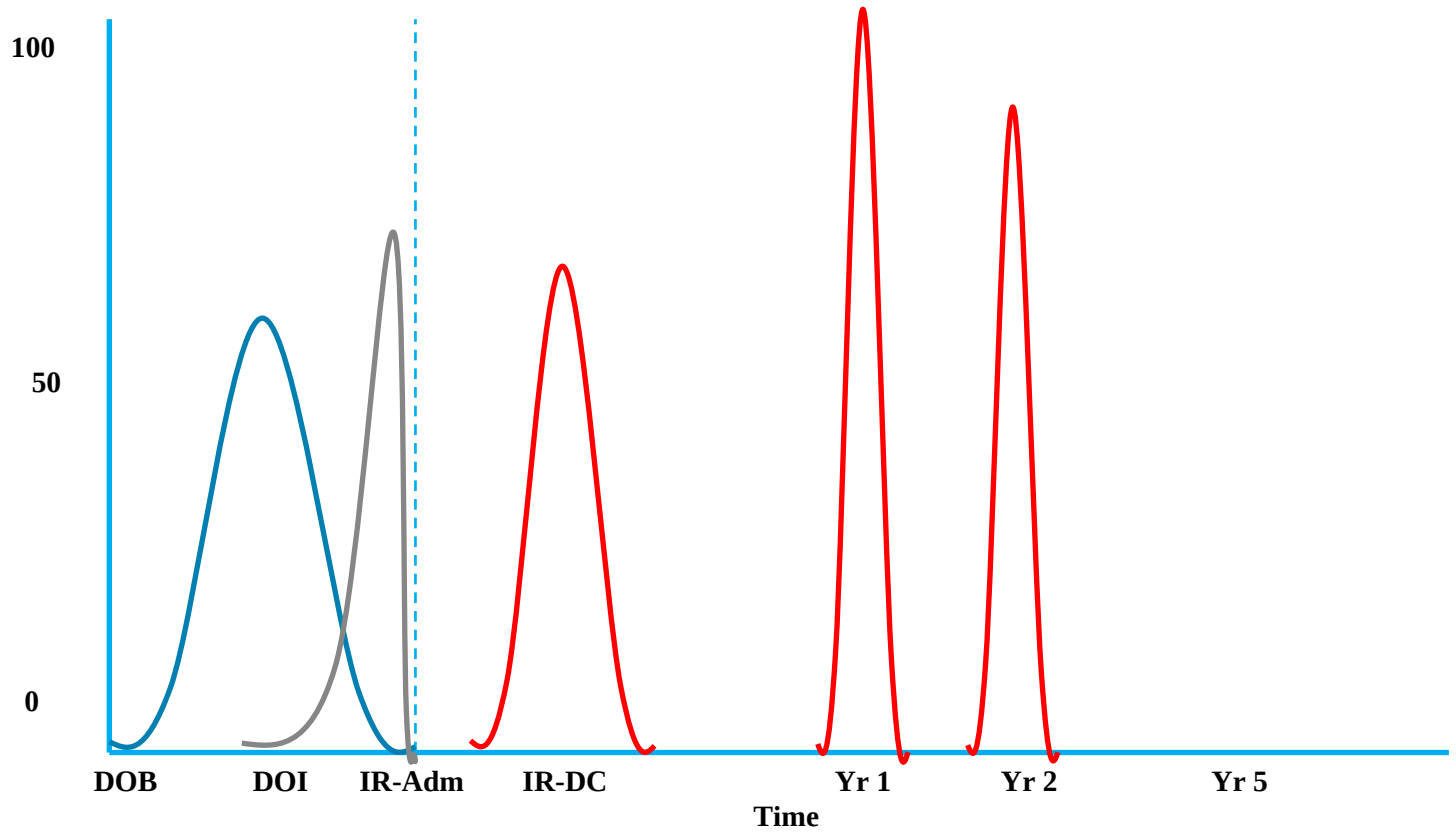
- **Do errors vary over time?**
  - Change in score distributions: random sampling (Bode 2014)
  - Response shift
- **Does timing of data collection matter?**
- **Ordinal or Interval?**

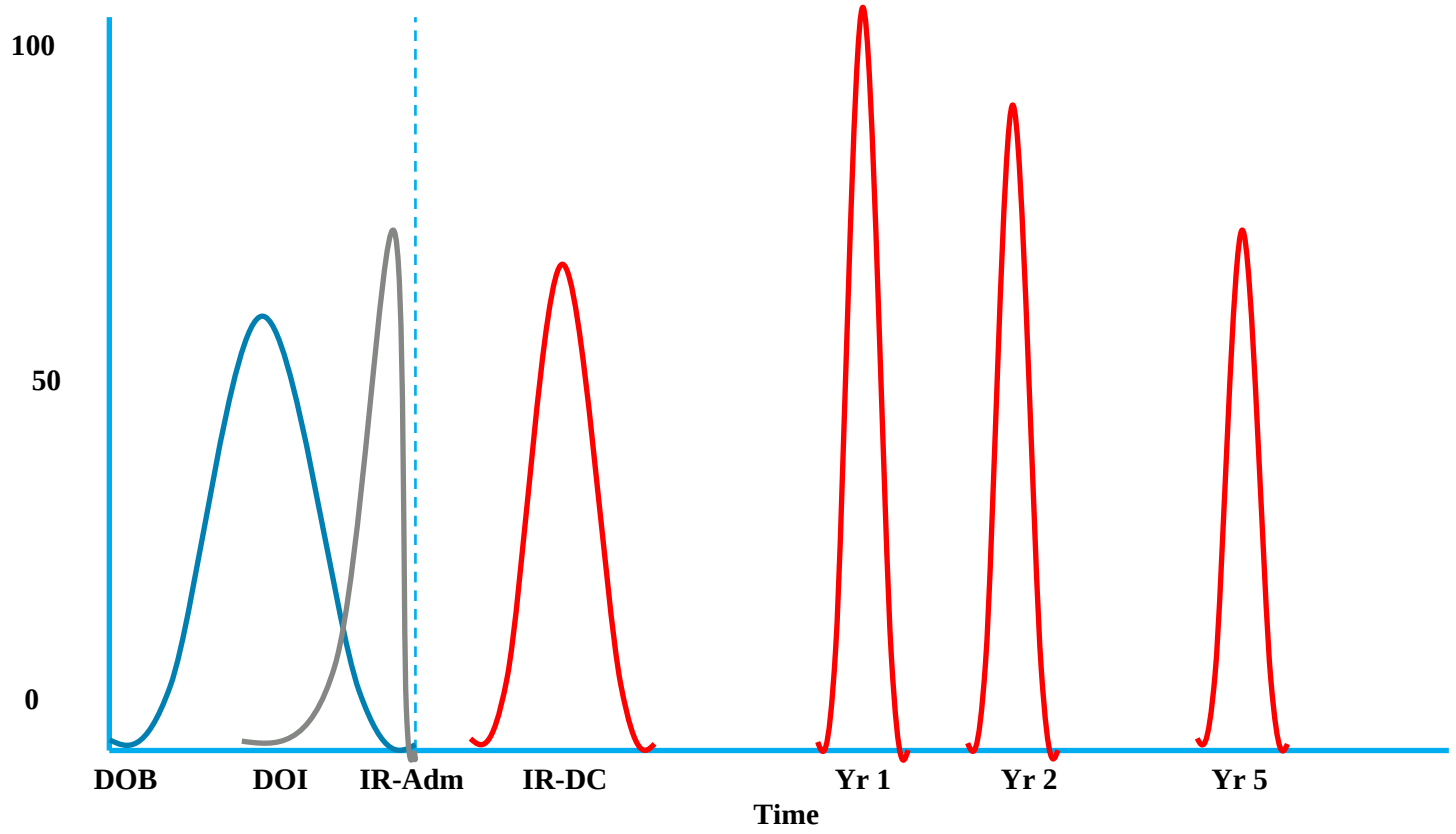

100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
YR 5
Yr 10
Time


100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
YR 5
Yr 10
Time


100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
YR 5
Yr 10
Time

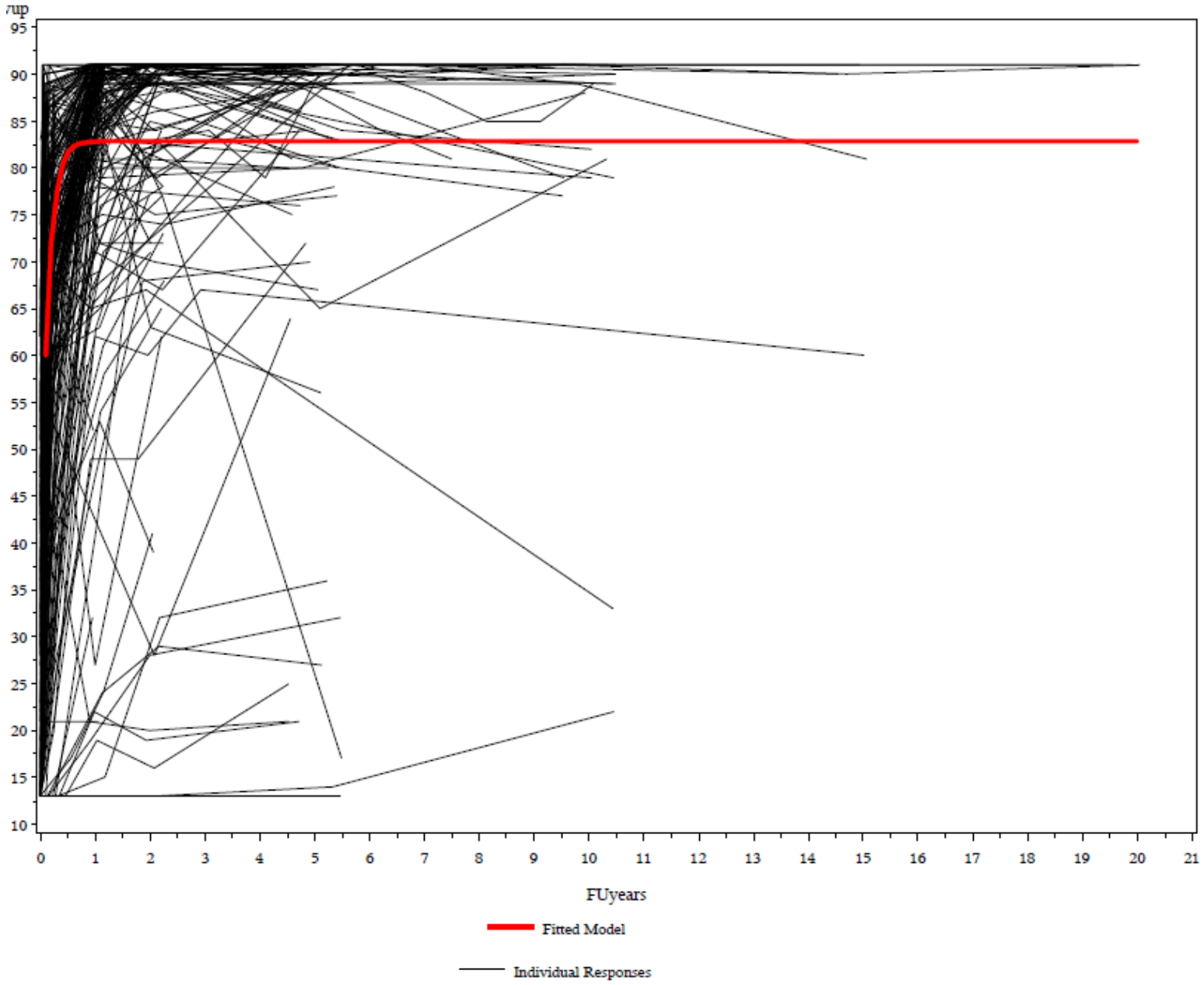

100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
Yr 5
Time


100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
Yr 5
Time


100
50
0
DOB
DOI
IR-Adm
IR-DC
Yr 1
Yr 2
Yr 5
Time

Implications for Modeling Longitudinal Data

- **Individual Growth Curve Models**
  - Hierarchical linear model
  - Explicitly model time
  - Account for within-person correlations of scores over time
  - Simultaneously model individual and group trajectories
  - Accommodate missing at random outcome scores and variable time point intervals
  - Linear, curvilinear, non-linear models
  - Covariate associations explain variance
  - More suited to large data sets


7up
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
FUyears
Fitted Model
Individual Responses

- Kozlowski AJ, Pretz CR, Dams-O'Connor K, Kreider S, Whiteneck G. An introduction to applying individual growth curve models to evaluate change in rehabilitation: a National Institute on Disability and Rehabilitation Research Traumatic Brain Injury Model Systems report. Archives of physical medicine and rehabilitation 2013;94(3):589-96.

- Pretz CR, Kozlowski AJ, Dams-O'Connor K, Kreider S, Cuthbert JP, Corrigan JD et al. Descriptive modeling of longitudinal outcome measures in traumatic brain injury: a National Institute on Disability and Rehabilitation Research Traumatic Brain Injury Model Systems study. Archives of physical medicine and rehabilitation 2013;94(3):579-88.

- Pretz CR, Dams-O'Connor K. Longitudinal description of the glasgow outcome scale-extended for individuals in the traumatic brain injury model systems national database: a National Institute on Disability and Rehabilitation Research traumatic brain injury model systems study. Archives of physical medicine and rehabilitation 2013;94(12):2486-93.

Summary

- Validity can be thought of as an attribute of scores, and the social consequences of interpreting and using scores for decision-making.
  - No such thing as a ‘valid and reliable instrument’
- CTT considers reliability and validity at of scores at the level of the instrument (scale or subscale summary scores only).
- IRT considers score reliability to be a probability function of person-ability and item-difficulty, and provides
  - additional item-level diagnostic tests.
  - Item- or score-level reliability estimates (SEs)
  - Additional tools for interpretation (e.g., key-form maps)
  - Alternate modes of administration (CAT, tailored short forms)

Summary

- IRT offers advantageous for modeling change over time:
  - May reduce error at single time points
  - Strategies to account for variability in errors over time
  - Does not preclude developing CTT-based indices to aid interpretation
  - Enhances interpretation of change for longitudinal models

Questions, Answers, and Discussion