Time as a Variable

Marty McCall

Northwest Evaluation Association

Brian D. Bontempo

Mountain Measurement, Inc.

On the matter of time…

I know well enough what it is, provided nobody asks me; but if I am asked what it is and try to explain, I am baffled.

St. Augustine

Confessions, Book 11, Chapter 14

Characteristics of Time

- Type of Variable
  - Manifest
  - Latent
- Type of Variable
  - Continuous
  - Discrete

Collecting Time Data

- Physical Science Tools
  - Stopwatch
- Universal Metric
  - Second
- Computerized Testing
  - Item Response Times

Aggregating Item Response Time

- Student's Test Taking Speed
  - Cognitive Processing Speed
  - Reading Speed
- Item's Duration
  - Cognitive Complexity
  - Readability

Using Item Response Times

- Non-Task ("Not on Task") Behavior
  - Engagement
  - Rushing
  - Motivation
  - Distraction
- Human Performance Factors
  - Fatigue
  - Warm-Up Effects
- Cheating

Models for On-Task Behavior

- Dichotomous
  - Engaged
  - Not Engaged
- Continuous
  - Threshold of Engagement

Detecting Non-Task Behavior

- Consistent Flag – Working consistently too fast or too slow to be engaged
- Aberrant (Unexpected) Flag – Items that deviate from expected “Cherry Picking”

Operationalization

Item Response Times

Test Taking Speed

Unexpected Item Response Times

Test Taking Speed

Item Duration

Engagement

Two Models of Test Taking Speed

Continuous (Van der Linden)

- Test taking speed is a Manifest variable
- Continuous Data
- Continuous Output

Rating Scale (Bontempo)

- Test taking speed is a Latent variable
- Discrete Data
- Continuous Output

Assumptions

- Test taking speed is stable and consistent
- Within test effects
- Warm-up effects (1st item was removed)
- Fatigue
- Administration engine was not perfect (Items longer than 10 minutes were not useful)

Data

- NWEA MAP Mathematics Assessments
- Spring 2003
- ~11,000 8th grade students
- ~500,000 responses
- ~3,000 items
- Slight differences in the data
  - More data to gain sample on low N items
  - Items with N < 30 were removed

Continuous Model

$$RT_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij}$$

- RT_{ij} is the ln of the observed response time for person j taking item i

  μ is the grand mean of all response times

  μ_i is the mean RT for item i; σ_i is the RT standard deviation for item i

  μ_j is the mean RT for person j; σ_j is the RT standard deviation for examinee j

  ε_{ij} is a residual term, ~N(0, σ_ε^2)

Non-Engagement Flags

- Unexpected Response Flag: $\varepsilon_{ij} < -2\sigma_{\varepsilon}$
- Consistent Flag: $RT_{ij} < \mu_j - 2\sigma_i$

Continuous Model Results

Flags per student	Unexpected Flag		Consistent Flag	
	N	%	N	%
0	4,207	38%	6,119	55%
1	3,076	28%	2,233	20%
2	1,699	15%	955	9%
3	853	8%	505	5%
4	481	4%	308	3%
5	270	2%	210	2%
6	178	2%	149	1%
7	113	1%	121	1%
8	77	1%	113	1%
9	50	0%	69	1%
>=10	134	1%	356	3%
	11,138	100%	11,138	100%

Rating Scale Model

Measurement Scale

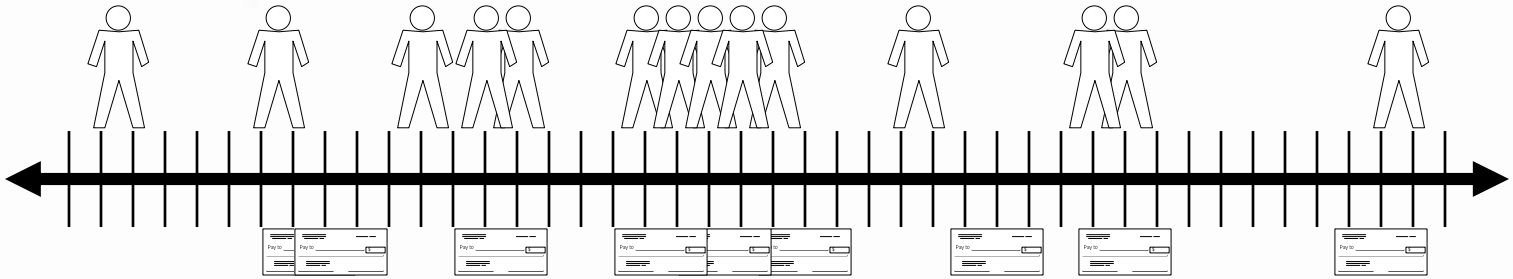
Test Takers : Examinee Speed

Slow Examinees

Fast Examinees

Θ = -3

Θ = +3



b = -3

b = +3

Short Items

Long Items

Test Items : Item Duration

Why IRT & Item Response Times

- IRT handles missing data quite easily
  - Enables analyses of CAT data
- IRT has a variety of fit indices that help asses the fit of an individual response
  - Enables detection of non-task behavior such as random answering, spacing, or taking a break
  - Enables detection of time related test taking phenomena such as fatigue, warm-up, and rushing
- IRT makes linking and equating easy
  - Allows for repeated measures designs that are void of preview effects

IRT & Item Response Times

- Each item response is converted to a rating scale value
  - ln time is calculated
  - ln time distribution is converted to a discrete distribution by creating equidistant boundary points for each category

Rating Scale

10 point rating scale (RS10)			
minimum item response time (seconds)	ln item response time lower bound	ln item response time upper bound	Rating Scale Value
0	-∞	1.50	9
4.48	1.50	2.00	8
7.39	2.00	2.50	7
12.18	2.50	3.00	6
20.09	3.00	3.50	5
33.12	3.50	4.00	4
54.60	4.00	4.50	3
90.02	4.50	5.00	2
148.41	5.00	5.50	1
244.69	5.50	∞	0

Non-Engagement Flags

- Unexpected: Z Score Residual < -2
- Consistent: Predicted Item Duration < (Item Duration - 2*SEM for Item)

Rating Scale Model Results

	Unexpected Flag		Consistent Flag	
Flags per student	**N**	**%**	**N**	**%**
0	4,466	42%		
1	3,012	28%		
2	1,404	13%		
3	681	6%		
4	363	3%		
5	239	2%		
6	149	1%		
7	101	1%		
8	76	1%		
9	53	0%		
>=10	74	1%		
	10,618	100%	10,618	100%

Conclusion

- Both models successfully modeled Test Taking Speed
- Both models were successful at detecting unexpected responses
- Easier to use continuous model for detecting consistent non-task behavior